Biologia Computazionale 2011- II sem. BIA

Universitá degli Studi di Milano

Docente: Giorgio Valentini Istruttore: Matteo Ré

Identificazione di regioni genomiche codificanti proteine (I)

Abstract

In questa esercitazione proveremo ad identificare regioni genomiche codificanti proteine. Inizieremo ad affrontare il problema considerando alcune caratteristiche delle regioni codificanti del genoma.
Tutti i nostri test verranno realizzati tenendo conto del fatto che i codoni sinonimi non vengono utilizzati con la stessa frequenza. Questo correla con la diversa disponibilitá dei t-RNA. Tale fenomeno é conosciuto con il nome di *codon bias*. Posta questa base biologica, ed osservato che delle regioni codificanti devono, per forza di cose, codificare il loro contenuto di informazione sottoforma di codoni, costruiremo un test che ci permetta di calcolare, data una sequenza osservata, la probabilitá che essa sia piú aderente ad un modello che non prevede un utilizzo differenziale dei codoni oppure ad un modello in cui i codoni vengono utilizzati con frequenze diverse.
Questo pone un problema: di fatto non cercheremo di calcolare la probabilitá di una data sequenza ma, piuttosto, la probabilitá che una data sequenza sia stata originata da un certo modello. Cercheremo quindi di calcolare uno score in grado di aiutarci a capire "quale modello (tra quelli considerati) é piú *verosimile*, data un'osservazione". Questo score é noto con il nome di *Likelyhood*. Come sempre i nostri test verranno realizzati utilizzando il linguaggio Perl.

1 Il fenomeno del codon bias

Codon bias é la probabilitá che un dato codone venga utilizzato per codificare un amminoacido rispetto ad altri codoni che codificano lo stasso amminoacido (es. GGT vs GGG per Gly).
Geni che vengono sempre espressi ad alto livello mostrano, nella maggioranza dei casi, delle forti preferenze per alcuni codoni rispetto ad altri codoni sinonimi. Tale fenomeno non si verifica nel caso di geni che vengono espressi a livello minore. Una seconda osservazione interessante é che geni che passano da bassi ad

alti livelli di espressione in risposta a determinate variazioni nelle condizioni intra e/o extracellulari mostrano un bias nell'utilizzo dei codoni simile a quello dei geni espressi costantemente ad alto livello. Potremmo quindi ipotizzare che geni il cui prodotto deve essere disponibile in tempi rapidi in risposta a determinati eventi non sono "liberi" di utilizzare i codoni sinonimi in modo casuale. L'origine di questo fenomeno vanno ricercate nelle abbondanze relative dei t-RNA: durante il processo di **traduzione** la presenza di codoni corrispondenti a t-RNA "rari" introduce pause traduzionali rallentando il processo. Se l'organismo ha necessitá di rendere disponibile in prodotto genico in tempi rapidi questo potrebbe costituire uno svantaggio. Ne deriva che nel corso dell'evoluzione, gli organismi si sono adattati ad utilizzare preferenzialmente, nei geni espressi ad alti livelli, codoni corrispondenti a t-RNA maggiormente disponibili. Questa teoria riguardo all'origine del fenomeno del codon bias é detta teoria **selezionista**. Una teoria alternativa é l'ipotesi **mutazionalista** (detta, a volte, *neutralista*). Essa dice che il fenomeno del codon bias é la risposta degli organismi al fatto che i pattern di mutazione nei genomi *non sono casuali* ma sono, al contrario, specifici degli organismi. Questa teoria, quindi, afferma che alcuni codoni sono soggetti a mutazione piú frequentemente di altri e il codon bias sarebbe quindi una semplice preferenza nell'utilizzo di codoni "piú stabili". La differenza rispetto alla teoria selezionista, quindi, é che nel corso dell'evoluzione non esiste una pressione selettiva per garantire il funzionamento efficiente dei geni altoespressi in maniera costitutiva, ma piuttosto una scelta piú generale a favore di codoni piú stabili dal punto di vista mutazionale.
Il problema dell'origine del codon bias é ancora aperto ma, a favore dell'ipotesi mutazionalista, sono stati pubblicati alcuni risultati che dimostrano come il parametro piú rilevante per la spiegazione delle differenze tra i particolari codon bias osservati in diversi organismi é il contenuto in GC. É ragionevole pensare che tale caratteristica (il contenuto in GC) risponda a processi generici che agiscono su tutto il genoma e non a processi che agiscono solo sulle regioni codificanti. Partendo da questo punto di vista é possibile pianificare esperimenti che tentino di dimostrare che é possibile predire il codon bias utilizzando informazioni derivanti dalle regioni genomiche non codificanti. Un esperimento di questo tipo ha dimostrato che é possibile predire le differenze di codon bias negli eucarioti basandosi unicamente su statistiche effettuate a partire dalle regioni intergeniche [2]. Per chi fosse interessato alle origini del fenomeno codon bias, segnalo una review interessante sull'argomento [1].
Indipendentemente dalla sua origine, dobbiamo fare alcune considerazioni sul codon bias come base di uno strumento di biologia computazionale avente lo scopo di identificare regioni codificanti. Il test del codon bias dipende dal contenuto in codoni. Ne deriva che, per ogni sequenza di DNA, il test va effettuato in *tutte* le possibili frame di lettura. Un altro fattore di cui tener conto é che il codon bias varia non solo tra genomi ma anche tra geni appartenenti al medesimo genoma. É quindi irrealistico aspettarsi che un test basato su codon bias sia in grado di rilevare tutti i geni codificanti proteine di un organismo.

Likelihood

Data una tabella di utilizzo dei codoni ed una sequenza di DNA S, maggiore é il numero dei codoni presenti nella sequenza riportati nella tabella come codoni ad "alta freauenza di utilizzo" e maggiore é il *potenziale codificante* della sequenza S (assimilabile alla probabilitá che S codifichi una proteina o parte di essa). Il potenziale codificante puó essere calcolato come segue:

$$\frac{p(C_1)p(C_2)p(C_3)...p(C_m)}{\frac{1}{64}\frac{1}{64}\frac{1}{64}...\frac{1}{64}}$$

Dove $C_1, C_2, C_3...C_m$ sono codoni (ad es. C_1 =ACC). Questa espressione é nota come *likelihood ratio* ed é normalmente calcolata in termini logaritmici, tanto che di solito ci si riferishe ad essa come *log-likelihood ratio*. Notiamo che, in questa formula, testiamo il rapporto tra le frequenze dei codoni sotto due ipotesi differenti: al numeratore assumiamo che S sia una sequenza codificante (e quindi utilizziamo le frequenze disomogenee dei codoni dovute all'esistenza del codon bias) mentre al demoninatore, assumiamo un utilizzo dei codoni puramente casuale, come sarebbe lecito aspettarsi se S fosse una sequenza non codificante.

DOMANDA : Perché la likelihood ratio viene calcolata in termini logaritmici?

Esistono diversi modi di utilizzare la log-likelihood ratio per stimare il potenziale codificante di una sequenza attraverso una valutazione della forza del codon bias. In prima istanza calcoleremo la log-likelyhood ratio **complessiva** per tutta la sequenza ma nulla vieta di utilizzare delle finestre scorrevoli lungo la sequenza di DNA. Il secondo approccio risulta particolarmente indicato per sequenze genomiche molto estese per le quali non é ragionevole aspettarsi che tutta la sequenza considerata codifichi proteine (pensiamo, ad esempio, a sequenze di interi cromosomi...).

Realizzazione del test in Perl :

Dal sito del corso (`http://homes.dsi.unimi.it/~re/corsobc11.html`), scaricate il file della tabella di utilizzo dei codoni umana (cdutable.txt). La tabella ha il seguente formato:

```
GGG 0.01708
GGA 0.01931
GGT 0.01366
GGC 0.02494
GAG 0.03882
GAA 0.02751
```

```
GAT 0.02145
GAC 0.02706
...
```

STEP I: Il primo passo sará quello di leggere i valori presenti nella tabella e caricarli in un array associativo (o *hash*) in cui ogni valore é associato ad una chiave testuale (che, nel nostro caso, sará costituita dal particolare codone a cui si riferisce il valore riportato nella seconda colonna). É sempre buona norma scrivere, prima del codice sorgente dello script Perl, una lista delle operazioni che vogliamo svolgere sottoforma di pseudocodice (una descrizione ad alto livello delle operazioni che verranno svolte dal programma in fase di esecuzione). Prevediamo fin d'ora che al programma passeremo due argomenti: il nome del file della tabella di utilizzo dei codoni ed un file contenente una sequenza nucleotidica in formato FASTA. L'output desiderato é il valore di log-likelihood ratio calcolato secondo la formula presentata nella sezione precedente.

```
acquisisci nome_file_tabella

apri nome_file_tabella

FINCHE' riga_corrente = leggi_riga(nome_tabella)
p(codone_prima_colonna) <- valore_seconda_colonna
fine FINCHE'

chiudi nome_file}
```

Come si vede é previsto l'utilizzo di un ciclo (FINCHE'). Questa parte del programma aprirá un file e ripeterá le stesse operazioni su ogni riga del file fino alla fine del file (momento in cui non ci saranno piú righe disponibili). L'operazione svolta sará, nel nostro caso, quella di leggere separatamente i valori delle colonne (quindi dovremo spezzare la riga in "parole") e salvare i valori in una varibile hash. Il valore nella prima colonna (il codone) verrá utilizzato come chiave mentre il valore nella seconda colonna (frequenza di utilizzo del codone), verrá salvato come valore associato alla chiave. Il tutto, in Perl, diventa:

```
$filectable = shift;

my %hashcodonusage=();

open in, $filectable;

while(<in>){
chop $_;
@vals = split(/\s+/,$_);
$hashcodonusage{$vals[0]}=$vals[1];
}

close in;
```

A questo punto esiste, nel programma, una variabile di tipo hash (array associativo) contenente tutte le frequenze di utilizzo dei codoni. Potremo accedere ad ogni valore unicamente mediante la chiave utilizzata durante la lettura del file della tabella di utilizzo dei codoni. L'effetto di queste istruzioni

```
$a = "GGG";
$codusage = $hashcodonusage{$a};
print "$codusage\n";
```

é la stampa a video della frequenza di utilizzo del codone GGG (il cui valore, letto dal file, era 0.01708).

STEP II: Il secondo passo sará la lettura dell'unica sequenza contenuta in un file FASTA. Il file puó essere scaricato dal sito del corso (link *exon sequence*, file exn.fa). Il suo contenuto é il seguente:

```
>HUMHBB.2ex
GCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACT
CCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCT
TTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCT
GCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGG
```

L'obiettivo di questa fase é quello di inserire la sequenza nucleotidica in un'unica variabile. I fatti rilevanti a proposito del file sono i seguenti:

- La prima riga del file é l'intestazione della sequenza fasta
- Tutte le altre righe (non sappiamo quante) costituiscono parti della sequenza

Il file verrá letto, come nel caso della tabella delle frequenze di utilizzo dei codoni, riga per riga. Lo pseudocodice per riuscire a salvare la sequenza in un'unica variable potrebbe essere il seguente:

```
acquisisci nomefile
crea SEQUENZAVUOTA

Apri file
finche' e' possibile leggi_riga

SE numero_riga_corrente < 2
salva_intestazione
ALTRIMENTI
ACCODA riga_corrente a SEQUENZAVUOTA
FINE SE
fine finche'
Chiudi file
```

In Perl il tutto diventa:

```
$fileseq=shift;
my $header;
my $sequence;
open in,$fileseq;
while(<in>){
chop $_;
if($.<2){
$header=$_;
}else{
$sequence = $sequence.$_;
}
}
close in;
```

La variabile $sequence ora contiene l'intera sequenza nucleotidica contenuta nel file FASTA.

STEP III: calcolo della log-likelihood ratio. Per realizzare quest'ultimo passaggio é necessario svolgere le seguenti operazioni:

- spezzare la sequenza nucleotidica in modo da ricavare tutti i codoni
- salvare i codoni in una variabile "lista di codoni"
- calcolare la log-likelihood ratio sotto forma di somma di termini (uno per ciascun codone osservato) corrispondenti al logaritmo del rapporto: $P(codone|modello_{codonusage})/P(codone|modello_{random})$

Per quanto riguarda la lista di codoni essa puó essere prodotta mediante un'unica istruzione in Perl:

```
@codons = ($sequence =~ m/.../g);
```

Il calcolo della log-likelihood ratio, quindi, puó essere realizzato mediante un ciclo attraverso tutti gli elementi dell'array codons e l'utilizzo dei valori contenuti nell'hash (array associativo) generato nello STEP I. Il codice Perl che realizza il tutto é il seguente:

```
my $result = 0.0;
foreach $curcodon (@codons){
$result = $result + log($hashcodonusage{$curcodon}) - log(1/64);
}
```

Alla fine di questo ciclo, la variabile $result contiene la log-likelihood ratio che ci permetterá di valutare se la sequenza esaminata é piú aderente al modello basato su un'utilizzo disomogeneo dei codoni (la cui osservazione é attesa solo in regioni codificanti) oppure al modello basato sull'utilizzo uniforme dei codoni. Per stampare il risultato utilizzeremo il comando **printf** (una forma specializzata di print):

```
printf "%.2f\n",$result;
```

Il comando printf permette di controllare il formato con cui vengono stampati i numeri. Nella forma che abbiamo appena utilizzato printf visualizza il contenuto della variabile $result stampando 2 cifre dopo la virgola.

Esercizi

Questi esercizi dovrebbero essere svolti esattamente nell'ordine in cui sono presentati. Tenete presente che **tutte** le informazioni necessarie a svolgere questi esercizi sono presenti in questo file PDF (esercizi 1-3) o nelle slide della lezione dedicata alle basi del linguaggio Perl che potete trovare sul sito del corso (esercizi 4-5).

1. Scrivere lo script Perl descritto in questa esercitazione. (3 punti)
   Eseguire lo script sulla sequenza FASTA esonica ed annotare lo score ottenuto.

2. Utilizzare lo script Perl per calcolare il potenziale codificante di una sequenza intronica (2 punti).
   Il file puó essere scaricato dal sito web del corso (link "intron sequence", file int.fa) ed annotare lo score ottenuto.

3. Gli score ottenuti vi permettono di discriminare tra la sequenza codificante e la sequenza non codificante? Giustificate la risposta (4 punti)

4. Se al punto 3 avete identificato un problema nella logica del programma realizzate uno script che lo risolva. Per verificare la vostra ipotesi di soluzione utilizzate lo script appena realizzato per verificare il potenziale codificante delle due sequenze coinvolte in questa esercitazione. (5 punti)

5. Spiegate il motivo **biologico** per cui la prima versione dello script (esercizi 1 e 2) *non riusciva* a discriminare tra la sequenza esonica ed intronica. (4 punti)

References

[1] R. Hershberg and D.A. Petrov. Selection on Codon Bias *Annu. Rev. Genet.*, 42:287–299, 2008.

[2] S.L. Chen, W. Lee, A.K. Hottes, L. Shapiro and H.H. McAdams. Codon usage between genomes is constrained by genome-wide mutational processes. Proc. Natl. Acad. Sci., USA, 101:3480–3485, 2004.