Series 5

Genomics and bioinformatics - Week 5

October 18, 2011

1 Markov model

Plasmodium Falciparum (protozoa responsible of malaria in humans) has a GC content of about 20%, and a genome length of 23Mb. Suppose now that a protein has a strong affinity with GC rich isochores (long regions of DNA with a relatively homogeneous GC content, which tend to be more flexible and contain more genes). We are interested in finding isochores to discover potential binding sites. In the case of Falciparum, only two 7Kb DNA isochores have a 50% GC content.

1. Draw a Hidden Markov Model that reflects the situation: identify hidden states and observed variables.
2. What are the emission probabilities from each state?
3. What is the probability, taking a random position in the genome, to be in an isochore? Call this probability $x = P(I)$, the complementary $y = 1 - x = P(N)$.
4. Call p, q the transition probabilities between your two states N and I. p represents the probability $P(I|N)$, and q is $P(N|I)$. What are then $P(I|I), P(N|N), P(I)$ and $P(N)$, as functions of p and q ?
5. Find p and q solving the equations obtained above.
6. Call $M_{\sigma,s}$ the matrix of transition probabilities, and $E_{T_n,s}$ the emission probability of character T_n from state s. Write the fist steps $F_{n,s}$ of the Forward Algorithm for the following sequence: AAGGCTT.

2 Reading frame

In this exercise you are given a nucleotide sequence which contains a coding region somewhere. You have to deduce what is the reading frame of this coding region.

The general procedure to find the right frame for reading a nucleotide sequence is to convert the nucleotide sequence into the corresponding possible amnio acid sequences and see which one makes the most sense. As you know, the base pairs are read three by three and translated into amino acids. One can hence read a sequence in three different ways: A, B and C.

So, to convert a base pair sequence (e.g. `CAGATTCTC`...) to a amino acid sequence (e.g. `GWLPHLQRI`...) you cut the base pair sequence in pieces of 3 nucleotides (e.g. `ČAG;̧` `ÄTT;̧`...), and use a conversion table that links any possible 3-mer to one of the 21 amnio acids. For instance, `CAG` codes for glutamine.

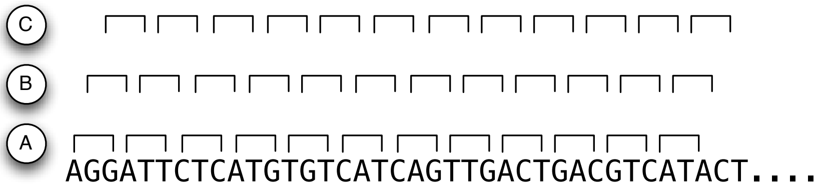

Figure 1: Three possible reading frames.

2.1 All 3-mers

To build the conversion table that links 3-mers to amino acids. We first need to build an exhaustive list of 3-mers. Write the code that takes as input the list of the four base pairs and generates as output all the possible permutations of size three.

The output should start like this and have 64 elements:

```
bases = ["t", "c", "a", "g"]
codons = [’ttt’, ’ttc’, ’tta’, ’ttg’, ’tct’, ’tcc’, ’tca’, ’tcg’, ’tat’, ...
```

2.2 3-mer to amino acid

We can now build a dictionary that links every 3mer to an amino acid. If you built the list correctly in the previous step, the corresponding amino acids are the following.

```
aminos = "FFLLSSSSYY**CC*WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG"
codon_to_amino = {’aaa’: ’K’, ’aac’: ’N’, ’aag’: ’K’, ’aat’: ’N’, ’aca’: ’T’, ...
```

2.3 Sequence to protein

You can now write the function that takes a nucleotide sequence as entry and outputs a protein sequence.

```
def seq_to_prot(seq): ...........
```

You should be able to use it like this:

```
seq_to_prot(’cagattctc’)
>>> QIL
```

2.4 Testing the three reading frames

You can now load the file "sequence.fa" and call the function you wrote in the last step with the three different possible frames and decide which one is right one.

3 Blast

Lorem ipsum.