

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 isochores? 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 $P(N)$, in function of p and q ? Write equations with all these probabilities in order to find p and q .

2 Reading frame

In this exercise you are given an 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 amino 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.

AGGATTCTCATGTGTCATCAGTTGACTGACGTCATACTTACTATACTCGCG...

Figure 1: Three possible reading frames.

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

2.1 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 function that takes four base pairs this as input and generates all the possible permutations of size three.

```
def all_threemers(bases=['a','g','t','c']):  
    ...
```

The output should start like this and have 213 elements:
asdasddsa

2.2 BP to AA

We can now build the function the takes a three-mer as entry and outputs an amino acid. If you build the list correctly in the previous step the corresponding amino acids are the following:

FLLSSSSYY**CC*WLLLLPPPHHHQRRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG

2.3 Sequence to protein

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

2.4 Sequence to protein

You can now 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.