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 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.

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. C̈AG; 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.

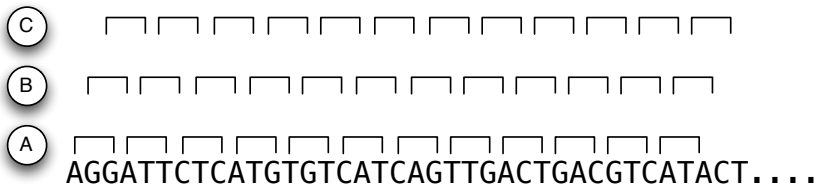

Figure 1: Three possible reading frames.

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

A common use of the BLAST tool is to identify the function of an unknown sequence. You have been provided with the sequence of a DNA fragment, `"fragment_007.fasta"` from an unknown micro-organism. Your aim is to use the NCBI BLAST programs to determine what kind of protein is `fragment_007` is likely to encode.

`http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome`

3.1 Nucleotide BLAST

Perform a nucleotide BLAST using `fragment_007.fasta`

1. Do you get any matches to `fragment_007`? Which parameters did you use? Record the alignment statistics for the top hits.
2. Extract the sequence of the hit with the highest query coverage (this may not necessarily be the top hit) and perform another nucleotide BLAST, using the same parameters. Record the alignment statistics for the top hits.
3. What differences do you observe between the two BLAST results?
4. To which parameter could you attribute the changes in the E-values?
5. What is the default threshold for the E-value on NCBI BLAST?
6. Do you have any significant hits suggesting a possible function for `fragment_007`?

3.2 Protein BLAST

Using the python function from the previous exercise, obtain the amino acid sequences for `fragment_007.fasta` in three reading frames. Choose the appropriate reading frame and save the corresponding protein sequence in `aa_007.fasta` format.

Perform a protein BLAST with `aa_007.fasta`

1. Are any well-known protein domains found?
2. Do you get any significant hits? Record the alignment statistics for the top hits.
3. What is the possible function of the protein encoded by `fragment_007`?
4. Which species is most predominant in your BLAST output?
5. Could you have obtained the same results using another BLAST program, without having to translate the nucleotide sequence of `fragment_007`?
6. How do results from Protein BLAST compare with the results from Nucleotide BLAST?

4 Finding orthologs

You have been provided with the sequence of the Pho2p protein from the famous yeast, *Saccharomyces cerevisiae*. Run BLAST to find out whether any putative orthologs of Pho2p are present in another, less famous yeast, *Candida glabrata*.

In a paper comparing the phosphate signal transduction in *S. cerevisiae* and *C. gabrata* (Genetics 182:471-9,2009), the authors identified the ortholog of Pho2p in *C. gabrata*.

Are the findings of the paper consistent with your observations?

5 Some BLAST tips

1. These can be used as a guide but should be considered with common sense.

   - E-value $<$ 10e-100 Identical sequences. You will get long alignments across the entire query and hit sequence.
   - 10e-50 $<$ E-value $<$ 10e-100 Almost identical sequences. A long stretch of the query protein is matched to the database.
   - 10e-10 $<$ E-value $<$ 10e-50 Closely related sequences, could be a domain match or similar.
   - 1 $<$ E-value $<$ 10e-6 Could be a true homologue but it is a grey area.
   - E-value $>$ 1 Proteins are most likely not related
   - E-value $>$ 10 Hits are most likely junk unless the query sequence is very short.

2. *Reciprocal Best Hit*

   This is a simple and commonly used test for predicting orthologous sequences.

   Genes A (from species X) and B (from species Y) will be considered as (putative) orthologs if (1) a search of similar sequences of A in species Y yields as the best hit B, AND a search of similar sequences of B in species X yields as the best hit A.