Using HMMs for eukaryotic promoter elements prediction

Shengting Li

1. *Introduction*

Transcription initiation is the result of specific interactions between proteins and DNA regions called promoters. In order for transcription to take place, the RNA polymerase must attach to the DNA near a gene. Promoters contain specific DNA sequences and response elements, which provide a binding site for RNA polymerase and for proteins called transcription factors that recruit RNA polymerase. There are many types of transcription factor binding site (TFBS) in the promoters. It's very difficult to confirm the TFBSs for a special gene experimentally[1]. And it's also difficult to predict a gene's promoter elements, especially for eukaryotic genes. Currently, the major method to predict promoter elements is using weight matrix[2, 3].

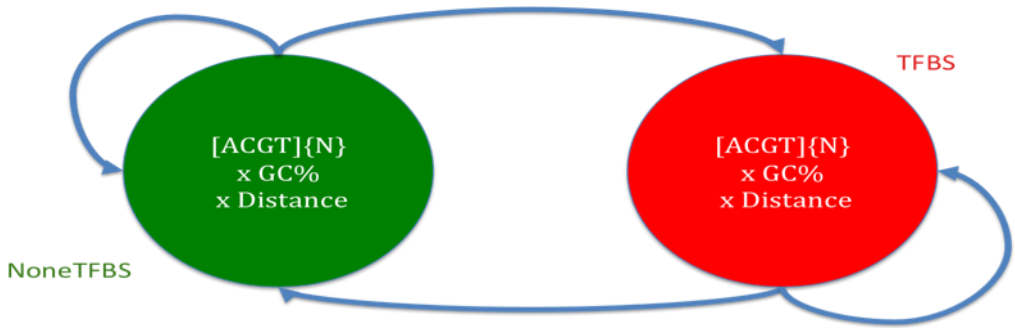
In this project, I'm trying to use HMM method for eukaryotic TFBS prediction. For convenience, I only focus on TATA-box and transcription start site (TSS) at first. These two are the most important elements in gene transcription procedure.

2. *Methods and Data*

The general idea of using HMM method in TFBS finding comes from using HMM for Gene finding[4].

Candidate Hidden Parameters:

- N bp TFBS box (4^N states)
- GC Content (0%, 5% ,10%, …, 100%=20 states)
- Distance to TSS (10bp, 20bp, 30bp, …,)



Used hidden parameters in TSS predict:

- N bp box (4^N states)
- GC Content (0%, 5% ,10%, …, 100%=20 states)

Used hidden parameters in TATA-box predict:

- N bp box (4^N states)
- Distance to TSS (10bp, 20bp, 30bp, …,)

Training Data:

- Training data for TSS predict:

1795 human promoters downloaded from EPD. Each promoter sequence contains -499 to 100 bp (total 600bp) around the TSS. Use 897 of them as training data and another 898 promoters as test data.

- Training data for TATA-box predict:

I've looked through several databases (EPD[3], TRANSFAC[1], JASPAR[2]…) to find annotated promoters as training data. But I failed. There are several annotated genes on TRANSFAC but far away from being sufficient for HMM training data. EPD database has enough promoters but only provides the TSS information. All databases provide the same weight matrix to predict TATA-box. This matrix comes from paper *Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoters. J. Mol. Biol. 212, 563-578.* So I have to create the training data by using weight matrix method first. Then use the artificial data as my training data.

Predict the TATA-box in the same 1795 promoters with matrix[6]:

	A	C	G	T	
01	61	145	152	31	S
02	16	46	18	309	T
03	352	0	2	35	A
04	3	10	2	374	T
05	354	0	5	30	A
06	268	0	0	121	A
07	360	3	20	6	A
08	222	2	44	121	W
09	155	44	157	33	R
10	56	135	150	48	N
11	83	147	128	31	N
12	82	127	128	52	N
13	82	118	128	61	N
14	68	107	139	75	N
15	77	101	140	71	N

Use 897 of them as training data and another 898 promoters as test data.

3. Results and Discussion

Unfortunately, the results are disappointed.

For TSS predict:

The result is totally useless.

GC_win_size: 25	3bp box	5bp box	7bp box
GC step: 10%	3/898 correct	4/898 correct	1/898 correct
GC step: 5%	2/898 correct	4/898 correct	0/898 correct

For TATA-box predict:

With 6bp box and 10bp distance step, I got

False Positive predicted length: 1222 bp (total promoters length 449000bp)
False Negative predicted length: 1514 bp (total promoters length 449000bp)
Correctly predicted length: 976bp (total TATA-box length 2490bp)

Which is also a bad prediction.

So, here I come to the conclusion that the HMM is not a better method for eukaryotic TFBS prediction for now.

But there are maybe some possible improvements that I can do in the future:

a. Get more real data rather than the artificial data for training.

b. Add more detail states in the HMM (only 2 now: None-TFBS, TFBS)

4. Reference:

1. TRANSFAC. http://www.gene-regulation.com/
2. JASPAR. http://jaspar.genereg.net/
3. EPD. http://www.epd.isb-sib.ch/
4. Hidden Markov models.
   http://www.cs.au.dk/~cstorm/courses/ML_f09/slides/hidden-markov-models-1.pdf
   http://www.cs.au.dk/~cstorm/courses/ML_f09/slides/hidden-markov-models-2.pdf
   http://www.cs.au.dk/~cstorm/courses/ML_f09/slides/hidden-markov-models-3.pdf
5. Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences.
   http://www.ncbi.nlm.nih.gov/pubmed/2329577
6. TATA-box base frequency table and weigth matrix. http://www.epd.isb-sib.ch/promoter_elements/tata_old.html