

Bioinformatika – Cv. 2

Motivation

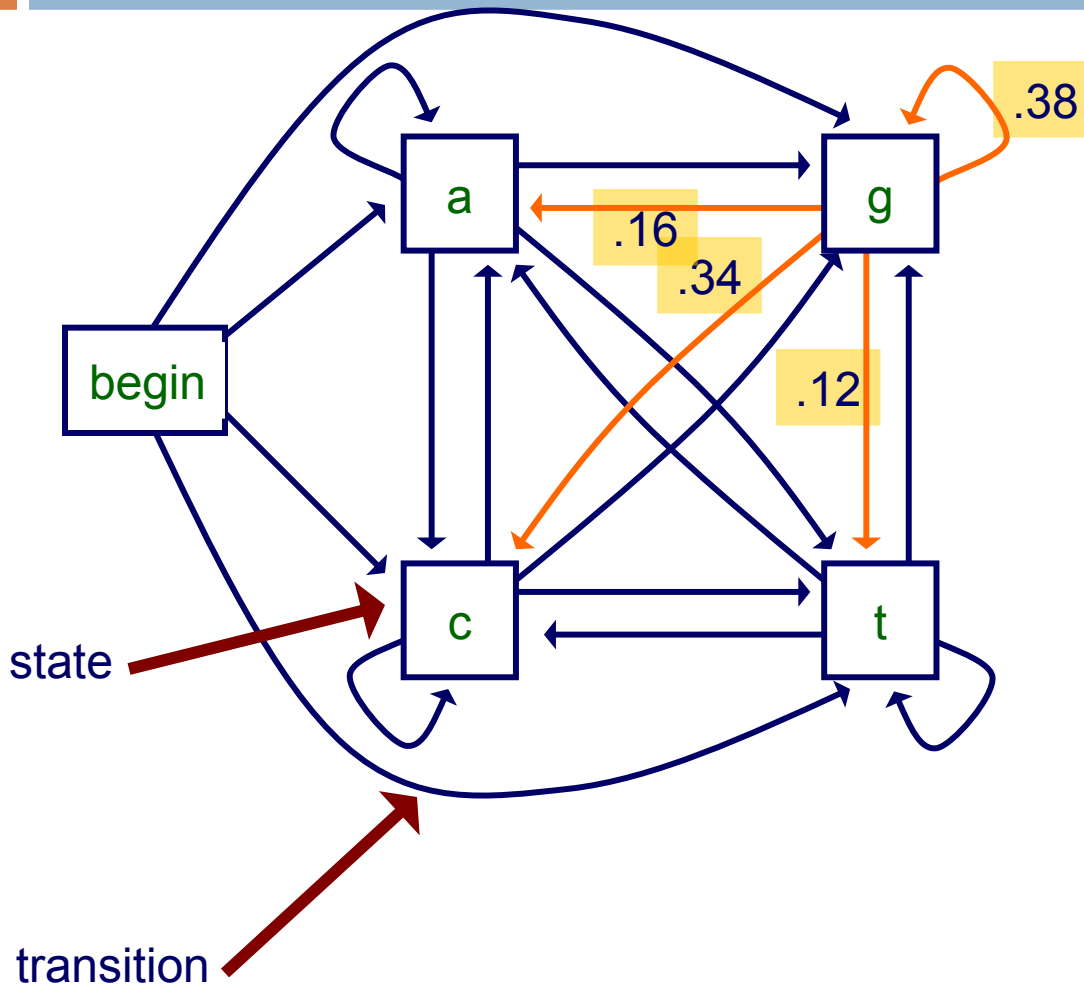


- CpG islands detection
- Protein binding regions detection (case study, BioProspector)
- Gene finding in DNA (... HMM, further)

Refresh Questions

- What is a Markov chain?
- What is Markovian property?
- How many parameters does the model have?
-

Markov Chain Model



transition probabilities

$$P(x_i = a | x_{i-1} = g) = 0.16$$

$$P(x_i = c | x_{i-1} = g) = 0.34$$

$$P(x_i = g | x_{i-1} = g) = 0.38$$

$$P(x_i = t | x_{i-1} = g) = 0.12$$

Refresh Questions

- to estimate a 1st order parameter, such as $P(c|g)$, we count the number of times that g follows the history c in our given sequences
- using Laplace estimates with the sequences

gccgcgcttg

gcttggtggc

tggccgttgc

$$P(a | g) = \frac{0 + 1}{12 + 4} \quad P(a | c) = \frac{0 + 1}{7 + 4}$$

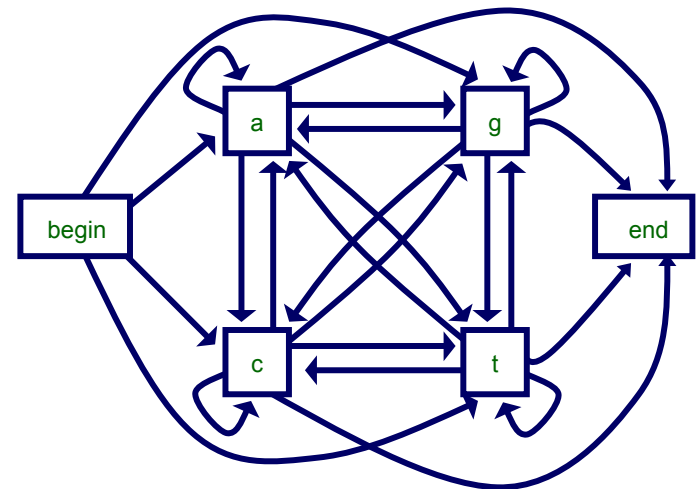
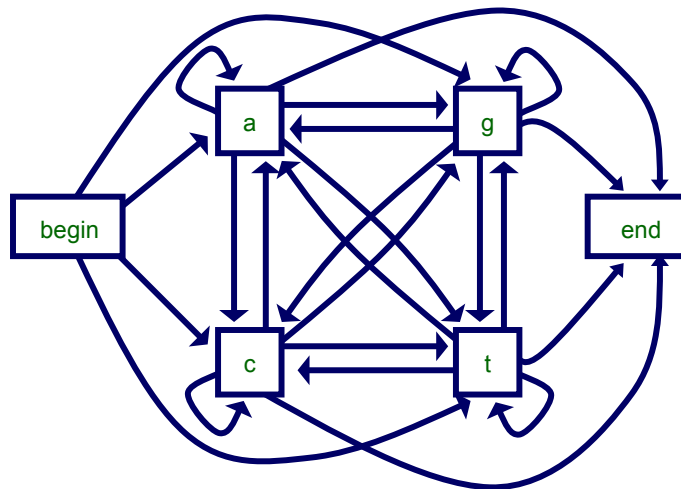
$$P(c | g) = \frac{7 + 1}{12 + 4} \quad M$$

$$P(g | g) = \frac{3 + 1}{12 + 4}$$

$$P(t | g) = \frac{2 + 1}{12 + 4}$$

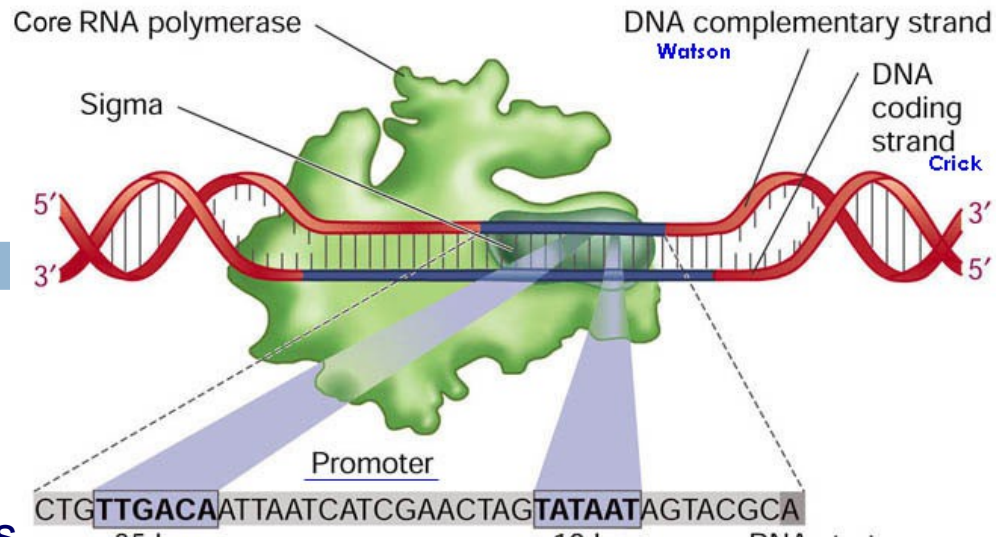
CpG islands classification

1. train two Markov models: one to represent CpG island sequence regions, another to represent other sequence regions (*null*)



2. given a test sequence, use two models to
 - determine probability that sequence is a CpG island
 - classify the sequence (*CpG* or *null*)

Case Study:



Mining Promotor Motifs

these sequences are *E. coli* promoters

```
tctgaaatgagctgttgacaattaatcatcgaactagttaactagtacgcaagttca  
accggaagaaaaccgtgacattttaacacgtttggttacaaggtaaaggcgacgccgc  
aattaaaattttattgacttaggtcactaaatactttaaccaatataggcatagcg  
ttgtcataatcgacttgtaaaccaattgaaaagatttaggtttacaagtctacacc  
catcctcgcaccagtcgacgacggtttacgctttacgtatagtggcgacaattttt  
tccagtataatttggtggcataattaagtacgacgagtaaaattacatacctgccc  
acagttatccactattcctgtggataaccatgtgtattagagttagaaaacacgagg
```

these sequences are not promoters

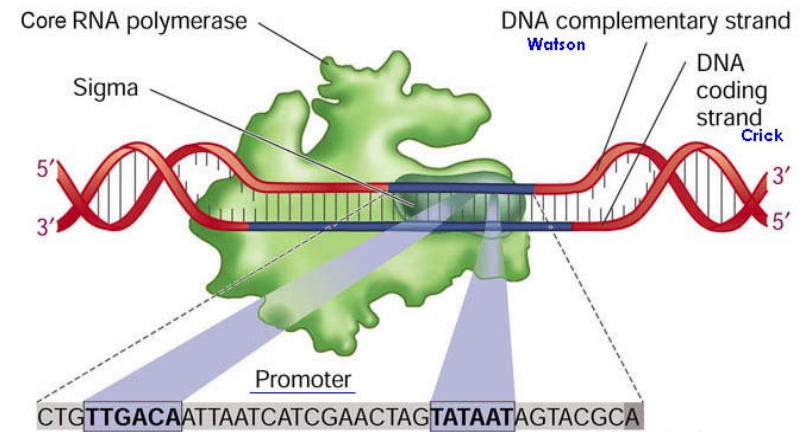
```
atagtctcagagcttgacctaactacgccagcattttggcgggtgtaagctaaccatt  
aactcaaggctgatacggcgagacttgcgagccttgctccttgcggtacacagcagcg  
ttactgtgaacattattcgtctccgcgactacgatgagatgcttgagtgcttccggt  
tattctcaacaagattaaccgacagattcaatctcgtggatggacggtcaacattga  
aacgagtcaatcagaccgctttgactctggtattactgtgaacattattcgtctccg  
aagtgcttagcttcaaggctcacggatacaccgaagcgagcctcgtcctcaatggcc  
gaagaccacgcctcggcaccgagtagacccttagagagcatgtcagcctcgacaact
```

How can we tell the difference? Is this sequence a promoter?

```
ccatcaaaaaaatatttctcaacataaaaaacttttggtgtaataacttgtaacgctacat
```

Mining Promotor Motifs

- Download a local copy of BioProspector
<http://robotics.stanford.edu/~xslu/BioProspector/>
- Find SigA transcription factor binding motif for *Bacillus subtilis*
- Reference genome:
<http://www.ncbi.nlm.nih.gov/nucore/AP012496.1?&feature=CDS>
- Compare with:
<http://dbtbs.hgc.jp/>



Mining Promotor Motifs

- Win: through <http://cygwin.com/>
- Reference genome: bacil_ref.fasta
- 1) Learn BioProspector on 7 promotor sequences only (bacil_red.fasta)
- 2) Learn on all the promotor sequences (bacil.fasta), but without refer. genome
- 3) Learn against only one reference gene (bacil_gene)
- 4) Learn with full information