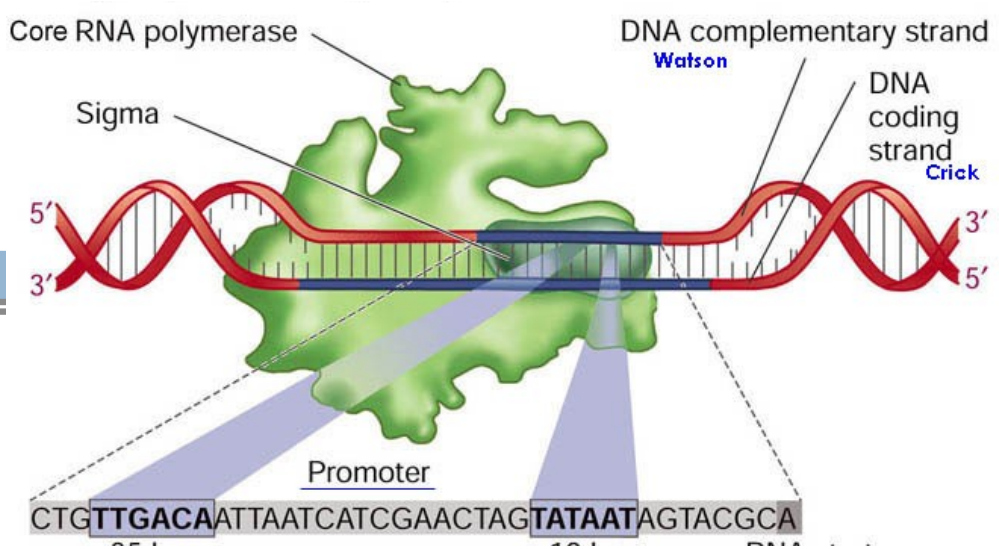
PROMOTOR MOTIF MINNING

Skryté markovské modely
(Some slides are courtesy of Mark Craven, U. of Wisconsin)

Motivation



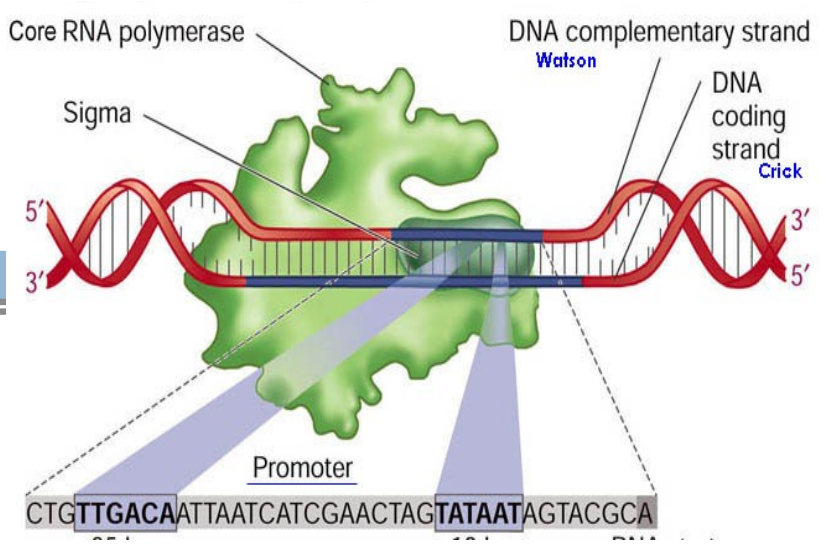
these sequences are E. coli promoters

```
tctgaaatgagctgttgacaattaatcatcgaactagttaactagtacgcaagttca
accggaagaaaaccgtgacattttaacacgtttgttacaaggtaaaggcgacgccgc
aaattaaaatttttattgacttaggtcactaaatactttaaccaatataggcatagcg
ttgtcataatcgacttgtaaaccaaattgaaaagatttaggtttacaagtctacacc
catcctcgcaccagtcgacgacggtttacgctttacgtatagtggcgacaatttttt
tccagtataatttgttggcataattaagtacgacgagtaaaattacatacctgcccg
acagttatccactattcctgtggataaccatgtgtattagagttagaaaacacgagg
```

these sequences are not promoters

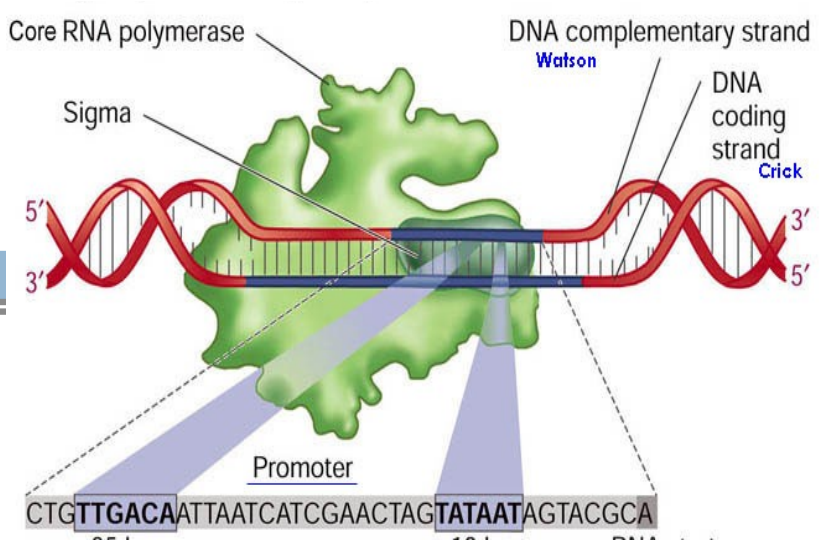
```
atagtctcagagtcttgacctactacgccagcattttggcggtgtaagctaaccatt
aactcaaggctgatacggcgagacttgcgagccttgtccttgcggtacacagcagcg
ttactgtgaacattattcgtctccgcgactacgatgagatgcctgagtgcttccgtt
tattctcaacaagattaaccgacagattcaatctcgtggatggacgttcaacattga
aacgagtcaatcagaccgctttgactctggtattactgtgaacattattcgtctccg
aagtgcttagcttcaaggtcacggatacgaccgaagcgagcctcgtcctcaatggcc
gaagaccacgcctcgccaccgagtagacccttagagagcatgtcagcctcgacaact
```

Bioprospector



- Download a local copy of BioProspector
- Find binding motifs of SigA transcription factor for Bacillus subtilis (ref. genome)
- Compare the motifs with SifA specific motifs found in particular genes in (Transcription Factors → Sigma factors → SigA)
- Deadline: next lesson, award: 5 pt.

How it works



- Promotor motif is a short sequence specific for a tran-scription factor
- BioProspector trains a motif matrix probability matrix Q*(i,n)*, i.e. the probability that nucleotide *n* is at *i*-th position of the motif. The expected likelihood of a non-motif subsequence is generated by 3th order Markov chain.
- Finally, top r consensual motifs are reported

Blk1	A	C	G	T
1	0.00	16.75	19.62	63.63
2	26.32	6.22	38.75	28.71
3	23.92	22.49	18.66	34.93
4	0.00	0.00	0.48	99.52
6	0.00	33.01	14.83	52.15

E.g. motif probability matrix

How?

- Linux: Make BioProspector.linux
- Windows: through http://cygwin.com/

Parameters:

- `-i <seq_file>` : promotor sequences
- `-b <seq_file>` : background sequences, i.e. genetic background
- `-W` : 1st motif width (e.g. `-W 6`)
- `-w` : 2nd motif width (e.g. `-w 6`), need to specify!
- `-o` : output file, where to **look for following**

```
Motif #<number>: (<block1>/<1kcolb>, <block2>/<2kcolb>)
```

Tasks

1. Learn 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 complete information (bacil.fasta + bacil_ref.fasta)
5. Compare the motifs found according to 1) - 4) in terms of dbtbs.hgc.jp/.
   Do they differ? If so, why?