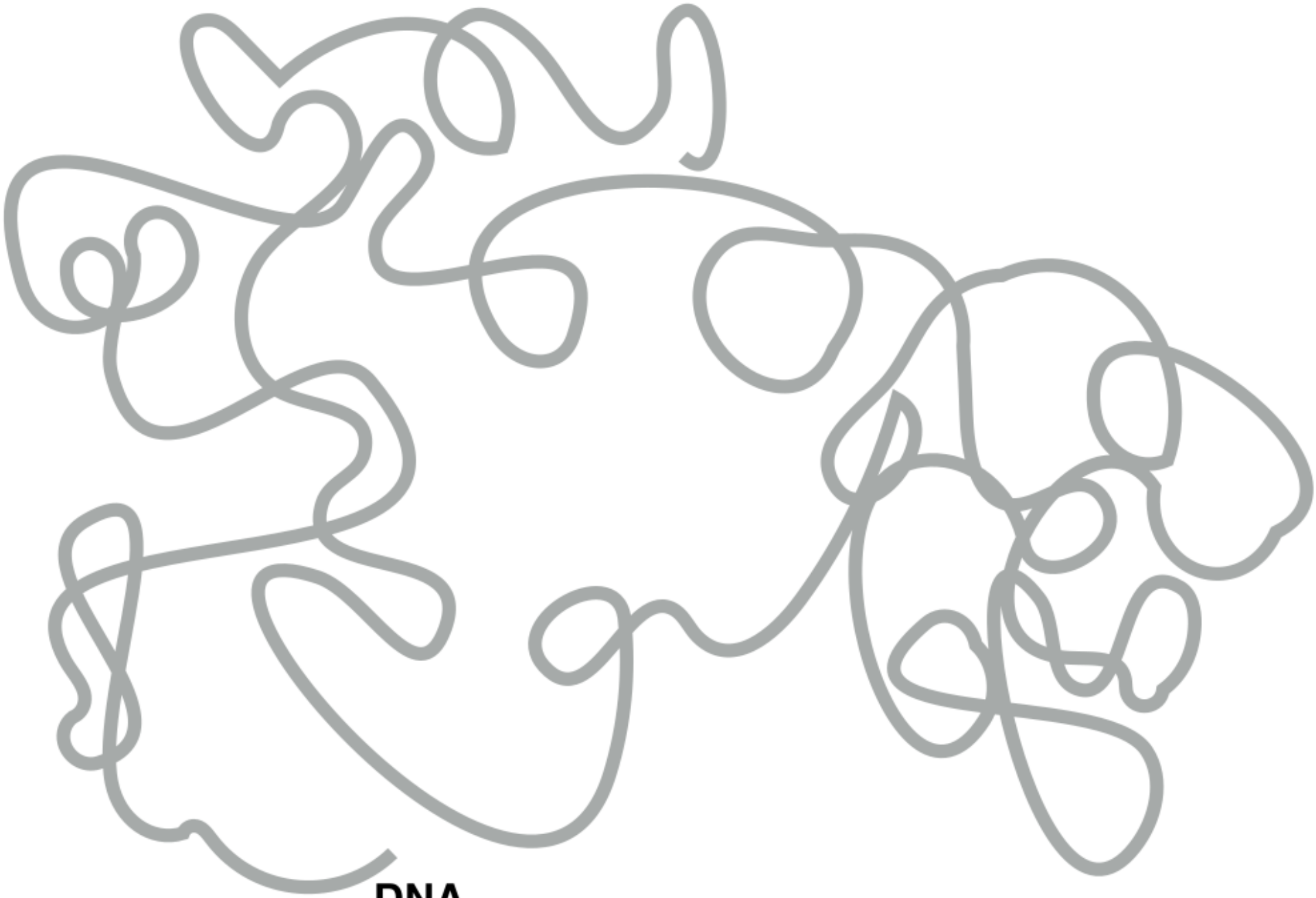
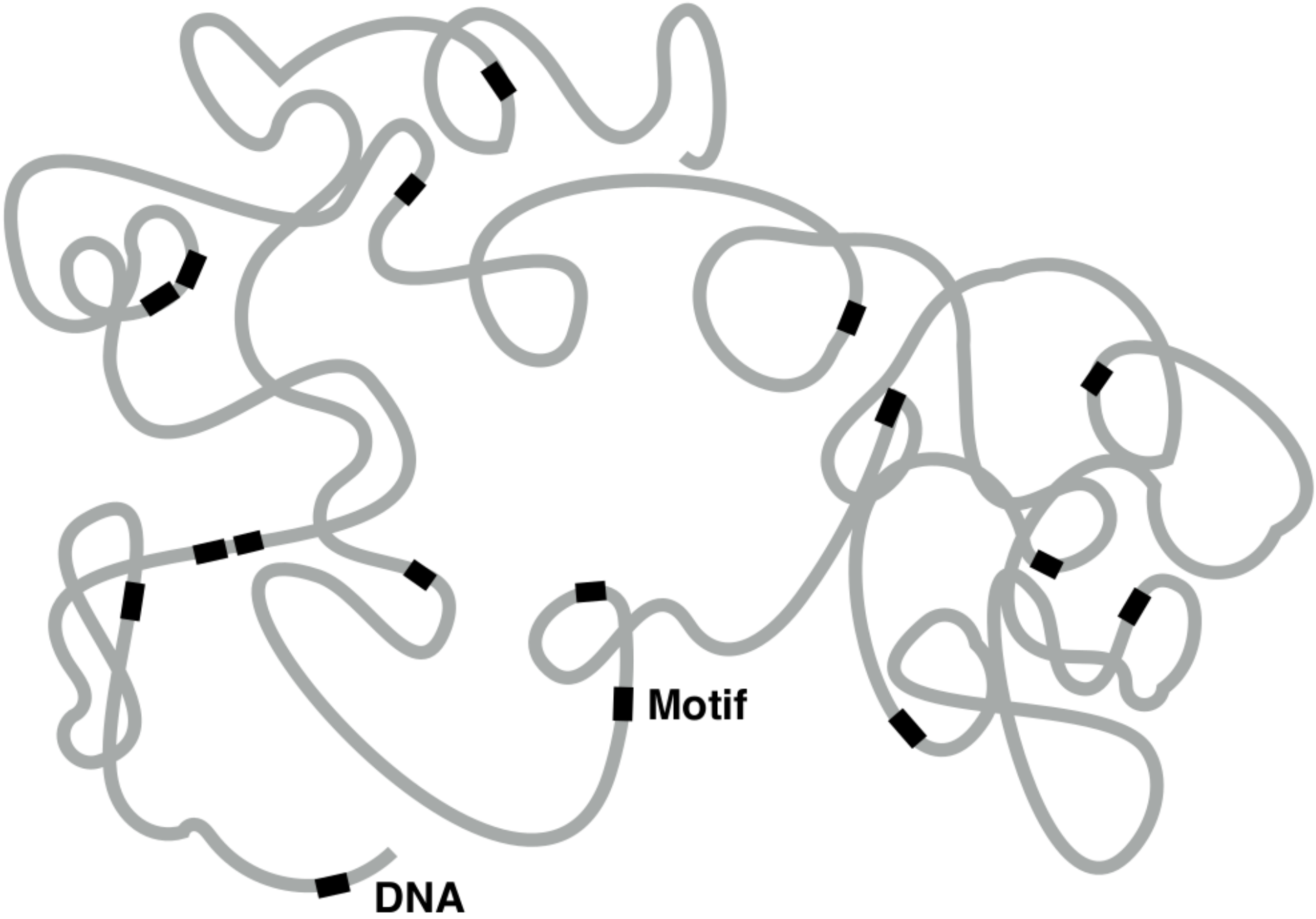
Motif Analysis

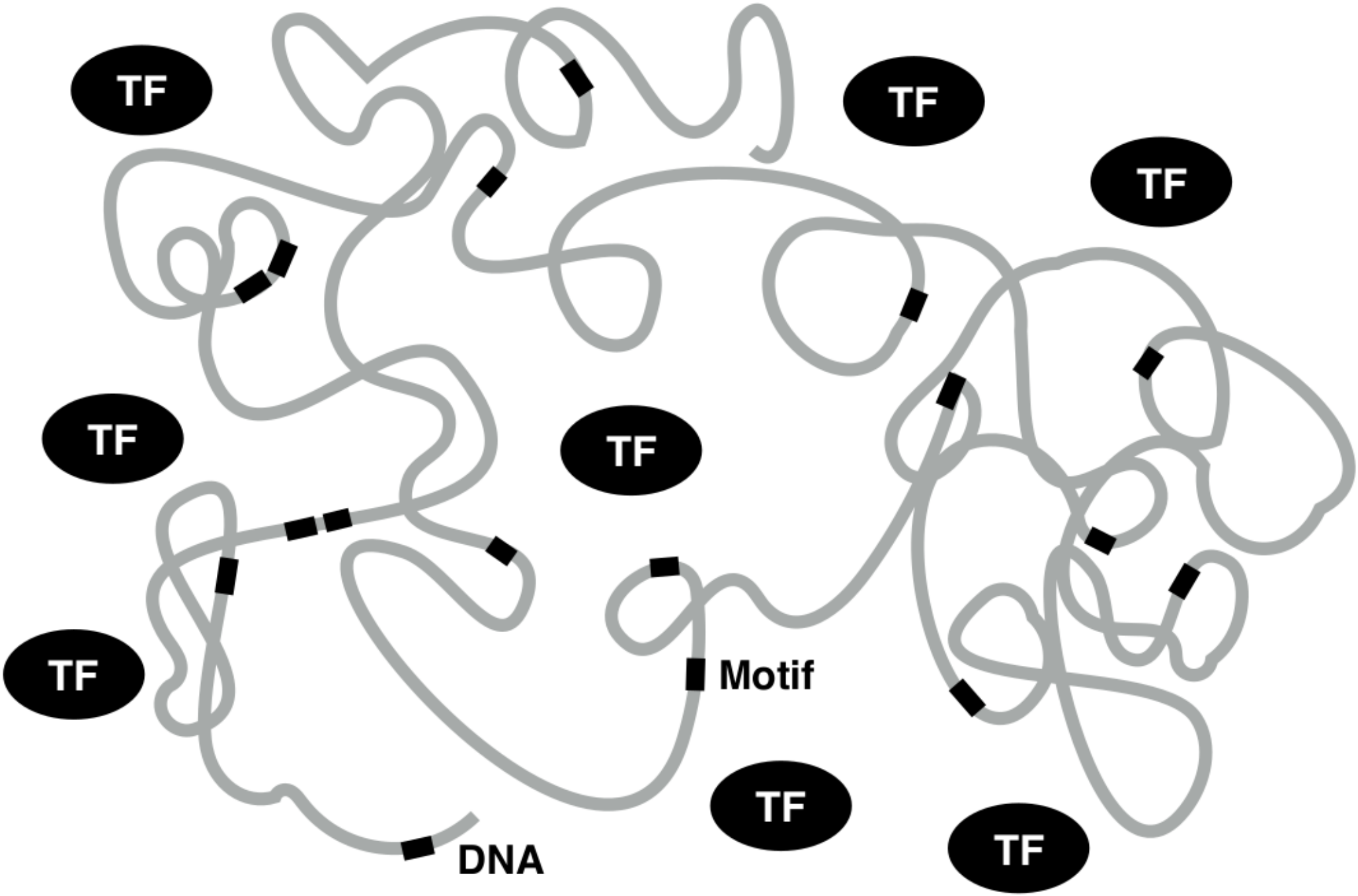
Alessandro Romanel

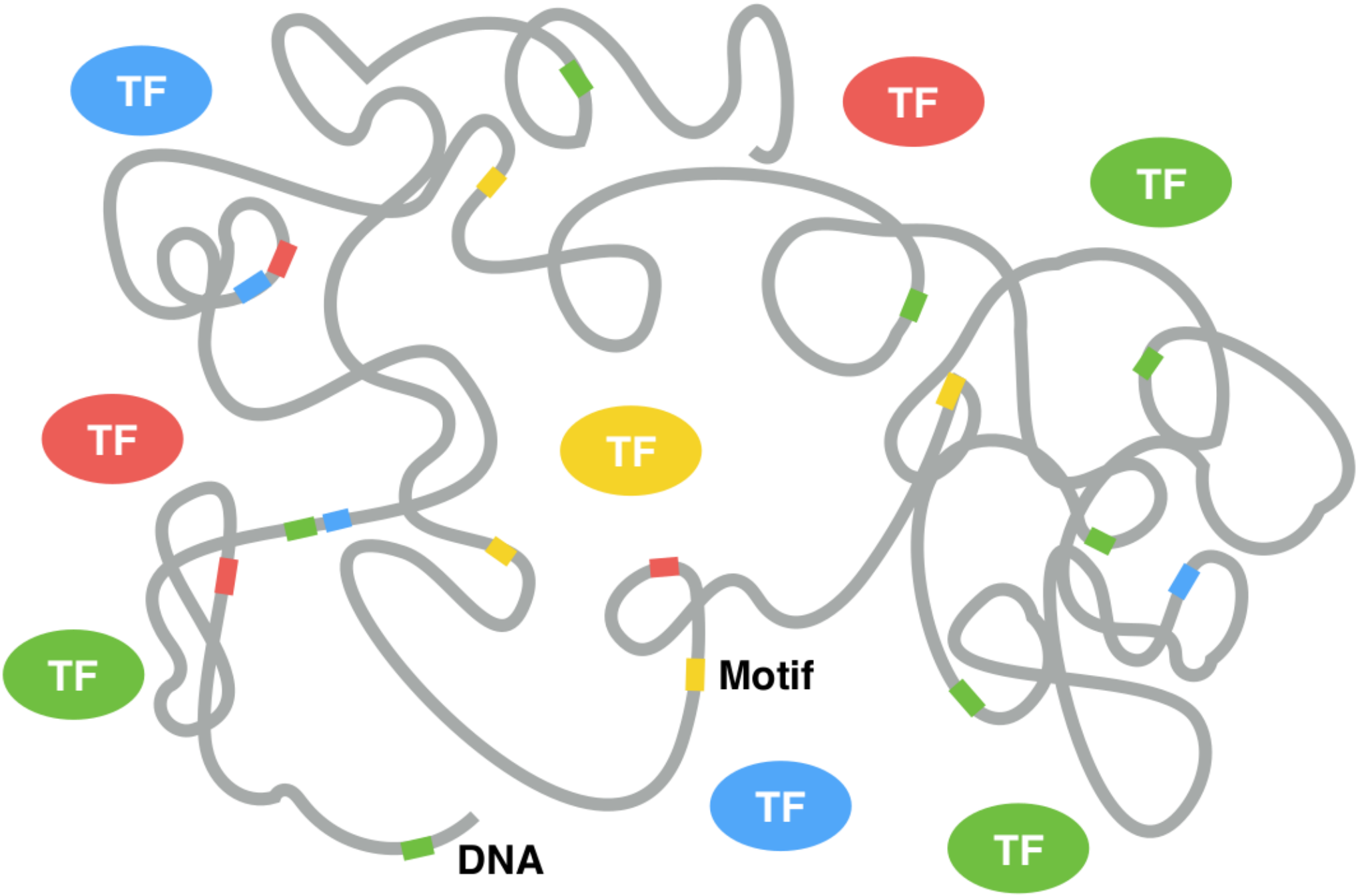
Bioinformatics Resources
2019-2020

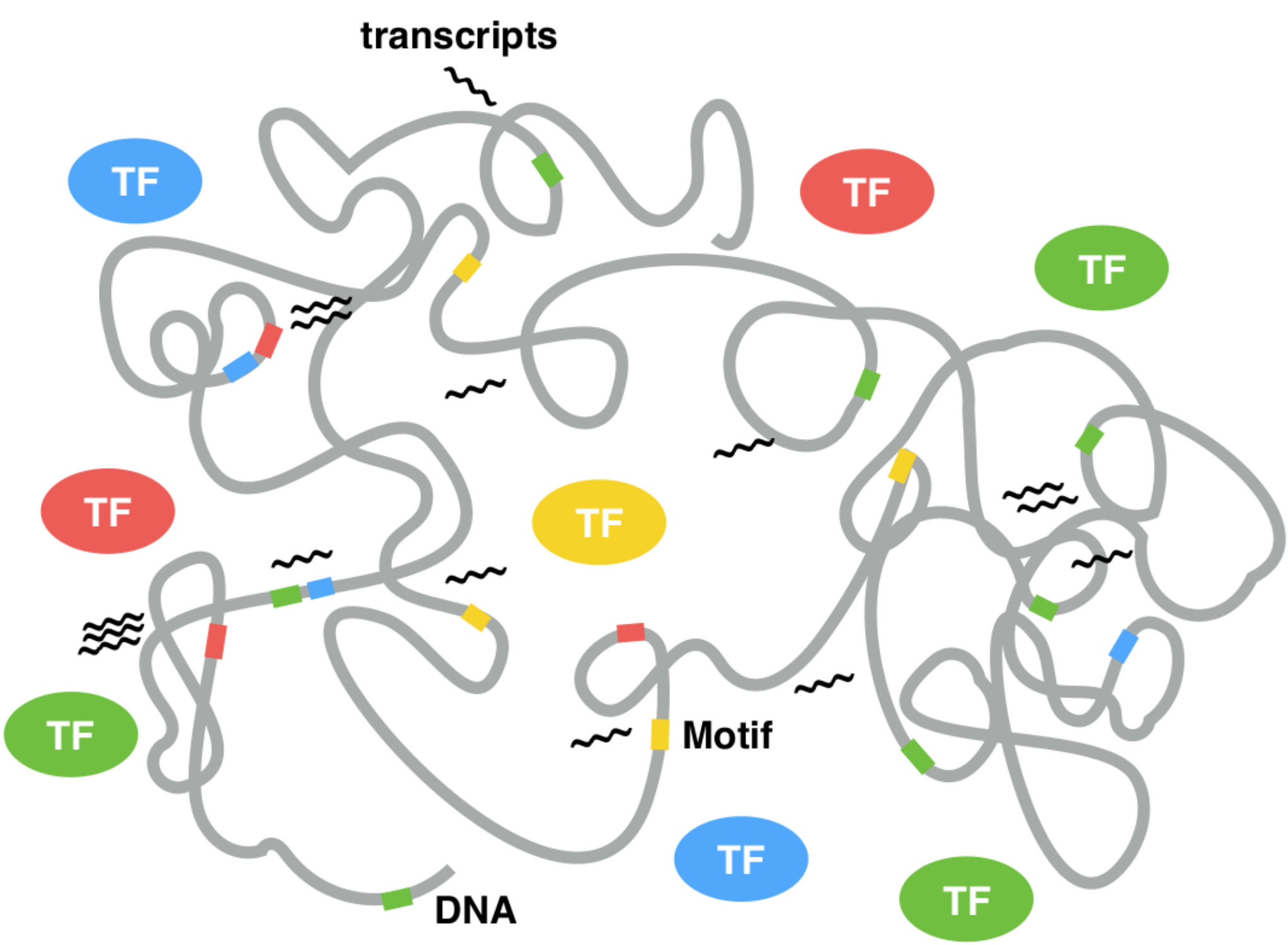


DNA


Motif
DNA


TF
TF
TF
TF
TF
TF
TF
TF
TF
Motif
DNA


TF
TF
TF
TF
TF
TF
TF
TF
Motif
DNA


transcripts
TF
TF
TF
TF
TF
TF
TF
TF
TF
Motif
DNA

Questions...

- How to find DNA motifs?
- Relations between motifs and transcription factors (TF)?
- Given a motif, which is the corresponding TF?
- Given a set of initial sequences, which are the more represented motifs?
- Given a set of transcripts, which are the motifs (and hence the TFs) in the promoters of the corresponding genes?

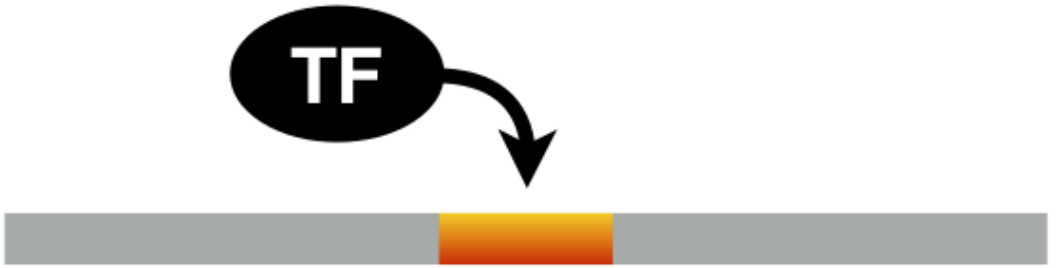
Questions...

- Given a set of motifs, which are the involved signaling pathways?
- Given a set of motifs, which are the signaling pathways that are more represented?
- ...

DNA motif (definition)

- Pattern of nucleotide sequences
- Usually they are associated to DNA-protein binding sites (regulatory regions)
- Small pattern (5-30bp) that can recur many times in the genome and many times for the same gene
- Standard motifs, palindromes and gapped

DNA motifs (definition)



Standard

Palindrome
(e.g. CACGTG)

Gapped

DNA motifs (functions)

- Sequence specific binding sites
  - TF, Nuclease, Ribosome
- mRNA processing
  - Splicing: Exonic Splicing Enhancer (ESE)
  - Editing: Protospacer adjacent motif (PAM), DNA sequence that immediately follow the target DNA sequence of Cas9 nuclease in the CRISPR system
  - Polyadenilation
  - Transcription termination

Motifs and TFs

- Motifs in regulatory regions are often similar but variable
- Transcription factors are often pleiotropic (1:N)
  - Regulate a lot of genes but need to be expressed at different levels
- An effect of degenerate motifs is the non-specific binding
  - A protein can bind in genomic positions that are different with respect to the one corresponding to the expected functional sites

Motif search (objectives)

- Identify over-represented motifs in the genome
- Identify motifs that are conserved in ortholog sequences
- Identify sequences that can be a candidate for TF binding

How to represent motifs

- Consensus sequences
- Profiles
  - Positional matrix
  - Hidden Markov Model (HMM)

Represent motifs

- TF binding sites are often represented a consensus binding sites
- Consensus sequence
  - Represents the result of multiple sequence alignments with the goal of finding recurrent motifs across the sequences
  - Potentially different from all input sequences
  - Presents only the most conserved sequences for each position

Consensus sequence

T	A	C	G	A	T	A	A	G
T	A	T	A	A	T	A	G	G
T	A	T	A	A	T	A	A	C
T	A	T	A	C	T	A	A	C
T	A	T	G	A	T	A	A	A
T	A	T	G	T	T	A	A	T

Consensus sequence

T	A	C	G	A	T	A	A	G
T	A	T	A	A	T	A	G	G
T	A	T	A	A	T	A	A	C
T	A	T	A	C	T	A	A	C
T	A	T	G	A	T	A	A	A
T	A	T	G	T	T	A	A	T
T	**A**	**T**	**G**	**A**	**T**	**A**	**A**	**G**

Remember IUPAC notation

Symbol[2]	Description	Bases represented				
A	**A**denine	A				1
C	**C**ytosine		C			
G	**G**uanine			G		
T	**T**hymine				T	
U	**U**racil				U	
W	**W**eak	A			T	2
S	**S**trong		C	G		
M	a**M**ino	A	C			
K	**K**eto			G	T	
R	pu**R**ine	A		G		
Y	p**Y**rimidine		C		T	
B	not A (**B** comes after A)		C	G	T	3
D	not C (**D** comes after C)	A		G	T	
H	not G (**H** comes after G)	A	C		T	
V	not T (**V** comes after T and U)	A	C	G		
N or **-**	any **N**ucleotide (not a gap)	A	C	G	T	4

Consensus sequence

T	A	C	G	A	T	A	A	G
T	A	T	A	A	T	A	G	G
T	A	T	A	A	T	A	A	C
T	A	T	A	C	T	A	A	C
T	A	T	G	A	T	A	A	A
T	A	T	G	T	T	A	A	T
T	**A**	**T**	**G**	**A**	**T**	**A**	**A**	**G**

Sequenza consenso IUPAC?

Symbol[2]	Description	Bases represented				
A	**A**denine	A				1
C	**C**ytosine		C			
G	**G**uanine			G		
T	**T**hymine				T	
U	**U**racil				U	
W	**W**eak	A			T	2
S	**S**trong		C	G		
M	a**M**ino	A	C			
K	**K**eto			G	T	
R	pu**R**ine	A		G		
Y	p**Y**rimidine		C		T	
B	not A (**B** comes after A)		C	G	T	3
D	not C (**D** comes after C)	A		G	T	
H	not G (**H** comes after G)	A	C		T	
V	not T (**V** comes after T and U)	A	C	G		
N or **-**	any **N**ucleotide (not a gap)	A	C	G	T	4

https://en.wikipedia.org/wiki/Nucleic_acid_notation

Consensus sequence

T	A	C	G	A	T	A	A	G
T	A	T	A	A	T	A	G	G
T	A	T	A	A	T	A	A	C
T	A	T	A	C	T	A	A	C
T	A	T	G	A	T	A	A	A
T	A	T	G	T	T	A	A	T
T	**A**	**T**	**G**	**A**	**T**	**A**	**A**	**G**
T	**A**	**Y**	**R**	**H**	**T**	**A**	**R**	**N**

Symbol[2]	Description	Bases represented				
A	**A**denine	A				1
C	**C**ytosine		C			
G	**G**uanine			G		
T	**T**hymine				T	
U	**U**racil				U	
W	**W**eak	A			T	2
S	**S**trong		C	G		
M	a**M**ino	A	C			
K	**K**eto			G	T	
R	pu**R**ine	A		G		
Y	p**Y**rimidine		C		T	
B	not A (**B** comes after A)		C	G	T	3
D	not C (**D** comes after C)	A		G	T	
H	not G (**H** comes after G)	A	C		T	
V	not T (**V** comes after T and U)	A	C	G		
N or **-**	any **N**ucleotide (not a gap)	A	C	G	T	4

https://en.wikipedia.org/wiki/Nucleic_acid_notation

Positional matrix

- Alternative to consensus
- The elements in the matrix represent all possible bases at each position

- Position Frequency Matrix (PFM) (PSWM)
- Position Probability Matrix (PPM) (PFM)
- Position Weight Matrix (PWM) (PSSM)

PFM matrix

T	A	C	G	A	T	A	A	G
T	A	T	A	A	T	A	G	G
T	A	T	A	A	T	A	A	C
T	A	T	A	C	T	A	A	C
T	A	T	G	A	T	A	A	A
T	A	T	G	T	T	A	A	T

	1	2	3	4	5	6	7	8	9
A	0								
C	0								
G	0								
T	6								

PFM matrix

T	A	C	G	A	T	A	A	G
T	A	T	A	A	T	A	G	G
T	A	T	A	A	T	A	A	C
T	A	T	A	C	T	A	A	C
T	A	T	G	A	T	A	A	A
T	A	T	G	T	T	A	A	T

	1	2	3	4	5	6	7	8	9
A	0	6	0	3	4	0	6	5	1
C	0	0	1	0	1	0	0	0	2
G	0	0	0	3	0	0	0	1	2
T	6	0	5	0	1	6	0	0	1

Matrix

PFM

	1	2	3	4	5	6	7	8	9
A	0	6	0	3	4	0	6	5	1
C	0	0	1	0	1	0	0	0	2
G	0	0	0	3	0	0	0	1	2
T	6	0	5	0	1	6	0	0	1

Matrix

PFM

	1	2	3	4	5	6	7	8	9
A	0	6	0	3	4	0	6	5	1
C	0	0	1	0	1	0	0	0	2
G	0	0	0	3	0	0	0	1	2
T	6	0	5	0	1	6	0	0	1

PPM

	1	2	3	4	5	6	7	8	9
A	0								
C	0								
G	0								
T	1								

Matrix

PFM

	1	2	3	4	5	6	7	8	9
A	0	6	0	3	4	0	6	5	1
C	0	0	1	0	1	0	0	0	2
G	0	0	0	3	0	0	0	1	2
T	6	0	5	0	1	6	0	0	1

PPM

	1	2	3	4	5	6	7	8	9
A	0	1	0	0.5	0.67	0	1	0.83	0.17
C	0	0	0.17	0	0.17	0	0	0	0.34
G	0	0	0	0.5	0	0	0	0.17	0.34
T	1	0	0.83	0	0.17	1	0	0	0.17

Matrix

PFM

	1	2	3	4	5	6	7	8	9
A	0	6	0	3	4	0	6	5	1
C	0	0	1	0	1	0	0	0	2
G	0	0	0	3	0	0	0	1	2
T	6	0	5	0	1	6	0	0	1

PPM

	1	2	3	4	5	6	7	8	9
A	0	1	0	0.5	0.67	0	1	0.83	0.17
C	0	0	0.17	0	0.17	0	0	0	0.34
G	0	0	0	0.5	0	0	0	0.17	0.34
T	1	0	0.83	0	0.17	1	0	0	0.17

PWM

	1	2	3	4	5	6	7	8	9
A	-inf	1.38	-inf	0.69	0.99	-inf	1.38	1.20	-0.39
C	-inf	-inf	-0.39	-inf	-0.39	-inf	-inf	-inf	0.31
G	-inf	-inf	-inf	0.69	-inf	-inf	-inf	-0.39	0.31
T	1.38	-inf	1.20	-inf	-0.39	1.38	-inf	-inf	-0.39

PPM matrix

	1	2	3	4	5	6	7	8	9
A	0	1	0	0.5	0.67	0	1	0.83	0.17
C	0	0	0.17	0	0.17	0	0	0	0.34
G	0	0	0	0.5	0	0	0	0.17	0.34
T	1	0	0.83	0	0.17	1	0	0	0.17

$$M_{k,j} = \frac{1}{N} \sum_{i=1}^{N} I(X_{i,j} = k)$$

- *k* is the set of all symbols in the alphabet (A,C,G,T)
- *N* is the number of aligned sequences
- *I* is an indicator function (1 if Xi,j=k, 0 otherwise)
- *j* ranges from the 1 to the length of the sequences

PPM matrix

	1	2	3	4	5	6	7	8	9
A	0	1	0	0.5	0.67	0	1	0.83	0.17
C	0	0	0.17	0	0.17	0	0	0	0.34
G	0	0	0	0.5	0	0	0	0.17	0.34
T	1	0	0.83	0	0.17	1	0	0	0.17

- Probabilities are calculated for each position independently
  - ?

PPM matrix

	1	2	3	4	5	6	7	8	9
A	0	1	0	0.5	0.67	0	1	0.83	0.17
C	0	0	0.17	0	0.17	0	0	0	0.34
G	0	0	0	0.5	0	0	0	0.17	0.34
T	1	0	0.83	0	0.17	1	0	0	0.17

- Probabilities are calculated for each position independently
  - We assume there is no statistical dependence between position in the pattern

PPM matrix

	1	2	3	4	5	6	7	8	9
A	0	1	0	0.5	0.67	0	1	0.83	0.17
C	0	0	0.17	0	0.17	0	0	0	0.34
G	0	0	0	0.5	0	0	0	0.17	0.34
T	1	0	0.83	0	0.17	1	0	0	0.17

- Given an input sequence, which is the probability to *belong* to a PPM?

  S: TACACTAGT

  P(S|M) = ?

PPM matrix

	1	2	3	4	5	6	7	8	9
A	0	1	0	0.5	0.67	0	1	0.83	0.17
C	0	0	0.17	0	0.17	0	0	0	0.34
G	0	0	0	0.5	0	0	0	0.17	0.34
T	1	0	0.83	0	0.17	1	0	0	0.17

- Given an input sequence, which is the probability to *belong* to a PPM?

  S: TACACTAGT

  P(S|M) = 1*1*0.17*0.5*0.17*1*1*0.17*0.17

  = 0.000417605

PPM matrix

	1	2	3	4	5	6	7	8	9
A	0	1	0	0.5	0.67	0	1	0.83	0.17
C	0	0	0.17	0	0.17	0	0	0	0.34
G	0	0	0	0.5	0	0	0	0.17	0.34
T	1	0	0.83	0	0.17	1	0	0	0.17

- Given an input sequence, which is the probability to *belong* to a PPM?

  S: TAC**C**CTAGT

  P(S|M) = ?

PPM matrix

	1	2	3	4	5	6	7	8	9
A	0	1	0	0.5	0.67	0	1	0.83	0.17
C	0	0	0.17	0	0.17	0	0	0	0.34
G	0	0	0	0.5	0	0	0	0.17	0.34
T	1	0	0.83	0	0.17	1	0	0	0.17

- Given an input sequence, which is the probability to *belong* to a PPM?

  S: TAC**C**CTAGT

  P(S|M) = ?

Laplace smoothing (pseudocounts): Allows to estimate probabilities in case of few observations

Pseudocounts

- A *pseudocount* is an amount (integer or double) added to the number of observed cases in order to change the expected probability
  - When values not known to be 0

$$p_i, \text{ empirical} = \frac{x_i}{N} \qquad p_i, \alpha\text{-smoothed} = \frac{x_i + \alpha}{N + \alpha d}$$

d is the number of observations

Pseudocounts

	1	2	3	4	5	6	7	8	9
A	0	6	0	3	4	0	6	5	1
C	0	0	1	0	1	0	0	0	2
G	0	0	0	3	0	0	0	1	2
T	6	0	5	0	1	6	0	0	1

S: TACACTAGT
P(S|M) = ?

S: TACCCTAGT
P(S|M) = ?

	1	2	3	4	5	6	7	8	9
A	1	7	1	4	5	1	7	6	2
C	1	1	2	1	2	1	1	1	3
G	1	1	1	4	1	1	1	2	3
T	7	1	6	1	2	7	1	1	2

PWM matrix

M

	1	2	3	4	5	6	7	8	9
A	-inf	1.38	-inf	0.69	0.99	-inf	1.38	1.20	-0.39
C	-inf	-inf	-0.39	-inf	-0.39	-inf	-inf	-inf	0.31
G	-inf	-inf	-inf	0.69	-inf	-inf	-inf	-0.39	0.31
T	1.38	-inf	1.20	-inf	-0.39	1.38	-inf	-inf	-0.39

$$M_{k,j} = \log_2 \left(M_{k,j} / b_k\right)$$

b represents a background model

$$b_k = 1/|k|$$

- b=0.25 for nucleotides (n=4) and 0.05 for amino acids (n=20)
- b can vary across nucleotide for organisms with high GC content

PWM matrix

M

	1	2	3	4	5	6	7	8	9
A	-inf	1.38	-inf	0.69	0.99	-inf	1.38	1.20	-0.39
C	-inf	-inf	**-0.39**	-inf	-0.39	-inf	-inf	-inf	0.31
G	-inf	-inf	-inf	0.69	-inf	**-inf**	-inf	-0.39	0.31
T	**1.38**	-inf	1.20	-inf	-0.39	1.38	-inf	-inf	-0.39

$$M_{T,1} = \ln(1/0.25) = 1.38$$

PWM matrix

M

	1	2	3	4	5	6	7	8	9
A	-inf	1.38	-inf	0.69	0.99	-inf	1.38	1.20	-0.39
C	-inf	-inf	**-0.39**	-inf	-0.39	-inf	-inf	-inf	0.31
G	-inf	-inf	-inf	0.69	-inf	**-inf**	-inf	-0.39	0.31
T	**1.38**	-inf	1.20	-inf	-0.39	1.38	-inf	-inf	-0.39

$$M_{T,1} = \ln (1/0.25) = 1.38$$

$$M_{C,3} = \ln (0.17/0.25) = -0.39$$

PWM matrix

M

	1	2	3	4	5	6	7	8	9
A	-inf	1.38	-inf	0.69	0.99	-inf	1.38	1.20	-0.39
C	-inf	-inf	**-0.39**	-inf	-0.39	-inf	-inf	-inf	0.31
G	-inf	-inf	-inf	0.69	-inf	**-inf**	-inf	-0.39	0.31
T	**1.38**	-inf	1.20	-inf	-0.39	1.38	-inf	-inf	-0.39

$$M_{T,1} = \ln (1/0.25) = 1.38$$

$$M_{C,3} = \ln (0.17/0.25) = -0.39$$

$$M_{G,6} = \ln (0/0.25) = -\inf$$

PWM matrix

M

	1	2	3	4	5	6	7	8	9
A	-inf	1.38	-inf	0.69	0.99	-inf	1.38	1.20	-0.39
C	-inf	-inf	-0.39	-inf	-0.39	-inf	-inf	-inf	0.31
G	-inf	-inf	-inf	0.69	-inf	-inf	-inf	-0.39	0.31
T	1.38	-inf	1.20	-inf	-0.39	1.38	-inf	-inf	-0.39

- Given an input sequence, which is the probability to *belong* to a PPM?
  S: TACACTAGT
  Score = ?

PWM matrix

M

	1	2	3	4	5	6	7	8	9
A	-inf	1.38	-inf	0.69	0.99	-inf	1.38	1.20	-0.39
C	-inf	-inf	-0.39	-inf	-0.39	-inf	-inf	-inf	0.31
G	-inf	-inf	-inf	0.69	-inf	-inf	-inf	-0.39	0.31
T	1.38	-inf	1.20	-inf	-0.39	1.38	-inf	-inf	-0.39

- Given an input sequence, which is the probability to *belong* to a PPM?
  S: TACACTAGT
  Score = 1.38+1.38-0.39+0.69-0.39+1.38+1.38-0.39-0.39
  = 4.65
- The score indicates how much the sequence is different from a random sequence

PWM matrix

M

	1	2	3	4	5	6	7	8	9
A	-inf	1.38	-inf	0.69	0.99	-inf	1.38	1.20	-0.39
C	-inf	-inf	-0.39	-inf	-0.39	-inf	-inf	-inf	0.31
G	-inf	-inf	-inf	0.69	-inf	-inf	-inf	-0.39	0.31
T	1.38	-inf	1.20	-inf	-0.39	1.38	-inf	-inf	-0.39

- Given an input sequence, which is the probability to *belong* to a PPM?
  S: TAC***C***CTAGT
  Score = ?

PWM matrix

M

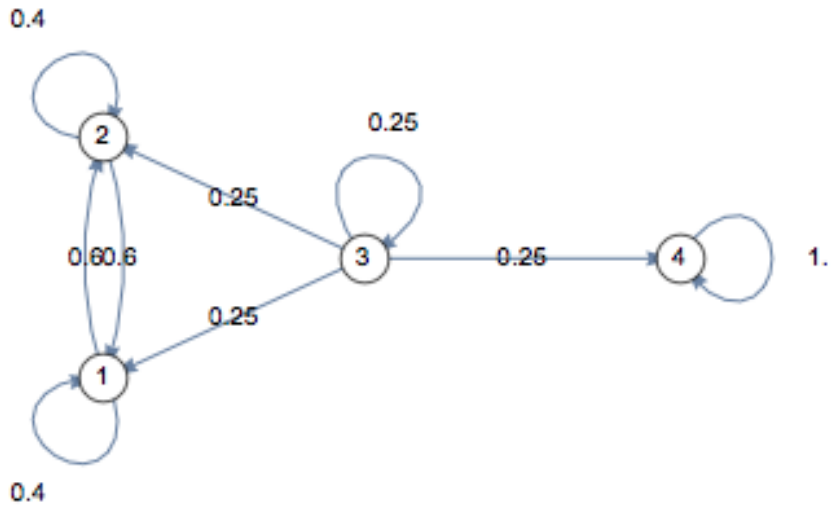
	1	2	3	4	5	6	7	8	9
A	-inf	1.38	-inf	0.69	0.99	-inf	1.38	1.20	-0.39
C	-inf	-inf	-0.39	-inf	-0.39	-inf	-inf	-inf	0.31
G	-inf	-inf	-inf	0.69	-inf	-inf	-inf	-0.39	0.31
T	1.38	-inf	1.20	-inf	-0.39	1.38	-inf	-inf	-0.39

- Given an input sequence, which is the probability to *belong* to a PPM?
  S: TAC***C***CTAGT
  Score = ?

  PSEUDOCOUNTS!!!!

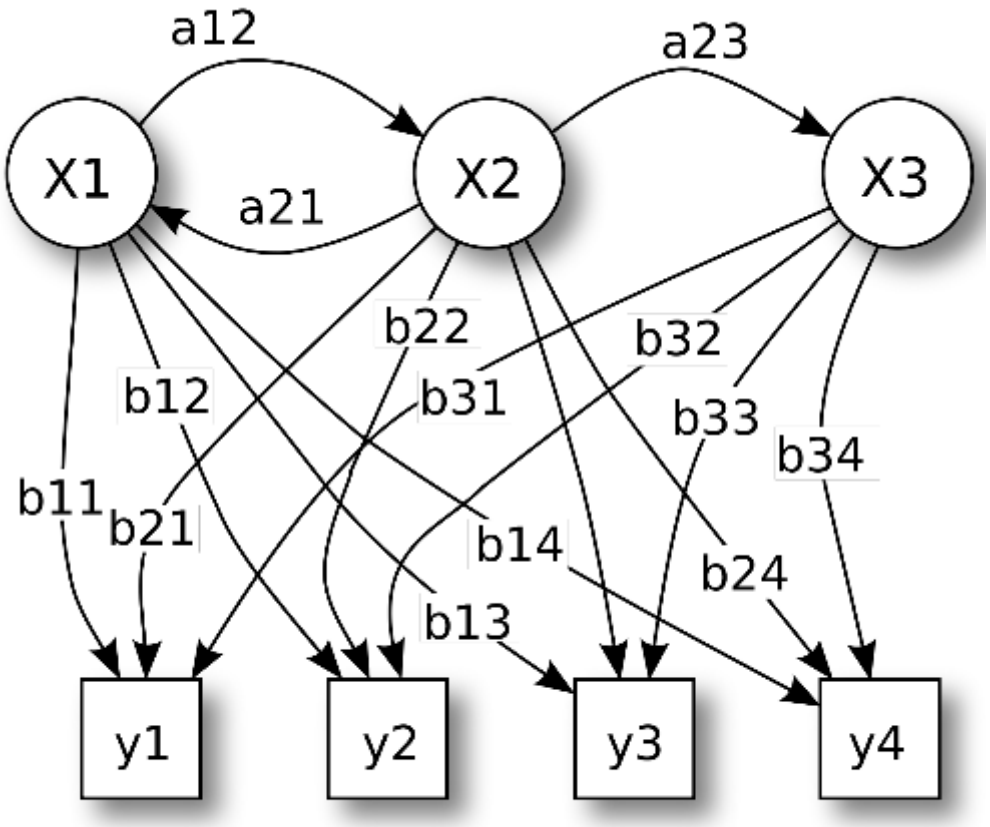
HMM profile

- HMM: Hidden Markov Model
- A Markov chain is a mathematical system that experiences transitions from one state to another according to certain probabilistic rules
  - no matter how the process arrived at its present state, the possible future states are fixed



HMM profile

- In a Markov chain, the state is directly visible to the observer
  - state transition probabilities are the only parameters
- In a HMM the state is not directly visible, but the output, dependent on the state, is visible



https://en.wikipedia.org/wiki/Hidden_Markov_model

HMM profile

- A HMM of the first order is defined as:
  - A finite set of states S
  - A discrete alphabet of symbols
  - A matrix of transition probabilities

    T = P(i|j)

    probability of transition from state j to i
  - A matrix of emission probabilities

    T = P(X|i)

    probability of X emission in state i

HMM profile

A C A C A A

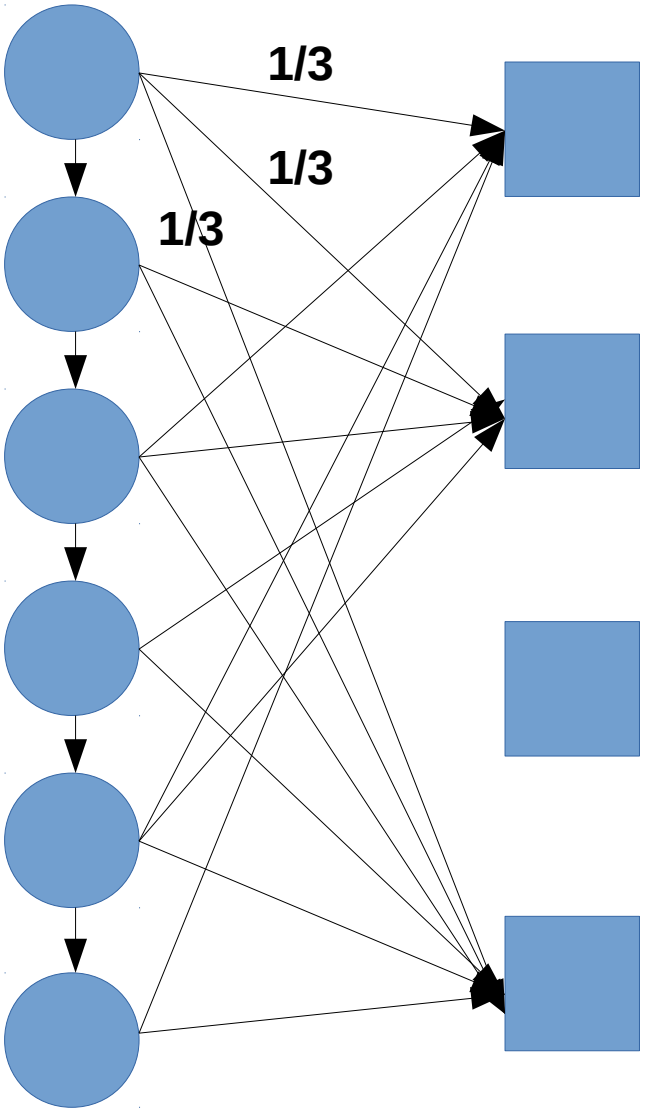
A T A C A A

T T T C T G

T T T C T G

C T C C C G

C T C T C G



HMM profile

A C A C A A

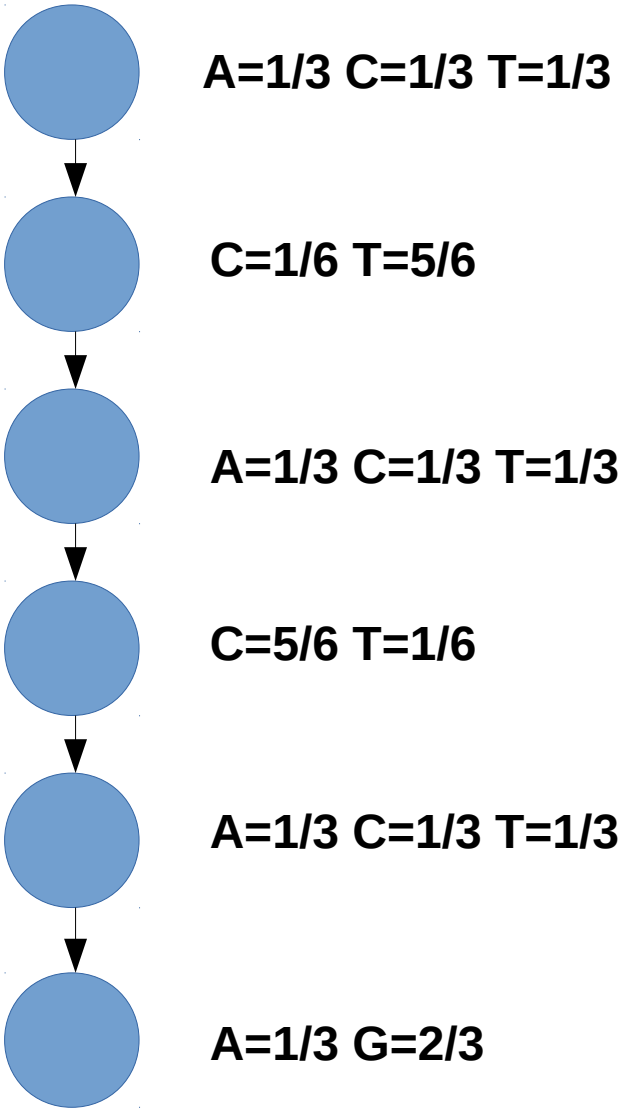
A T A C A A

T T T C T G

T T T C T G

C T C C C G

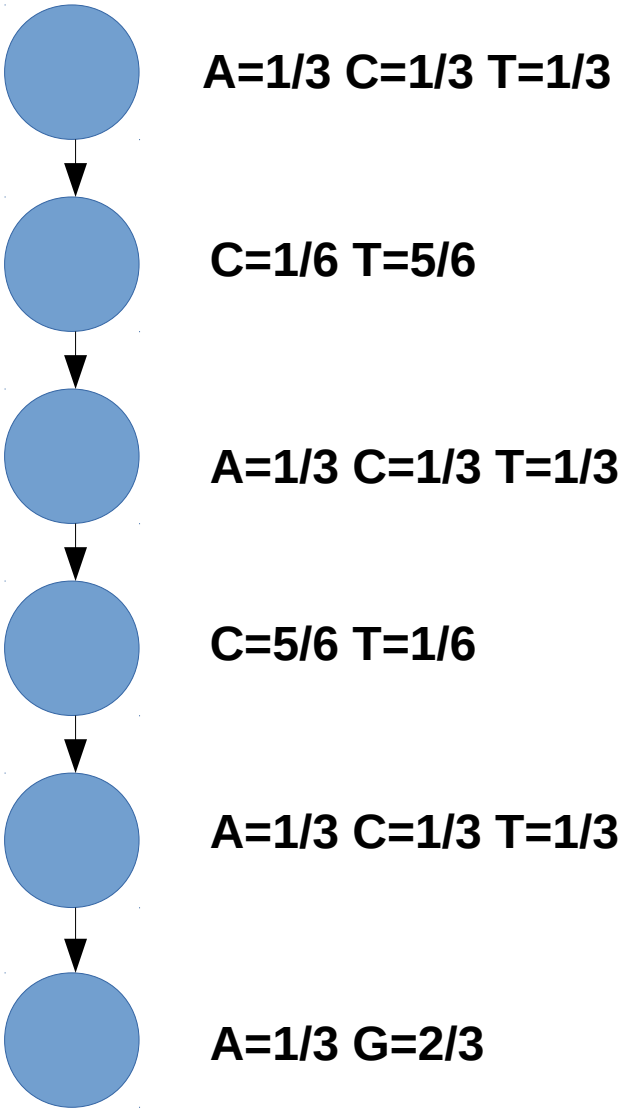
C T C T C G



HMM profile

S = CCATAA

P(S|M) = ?



HMM profile

S = CCATAA

P(S|M)

= 1/3*1*

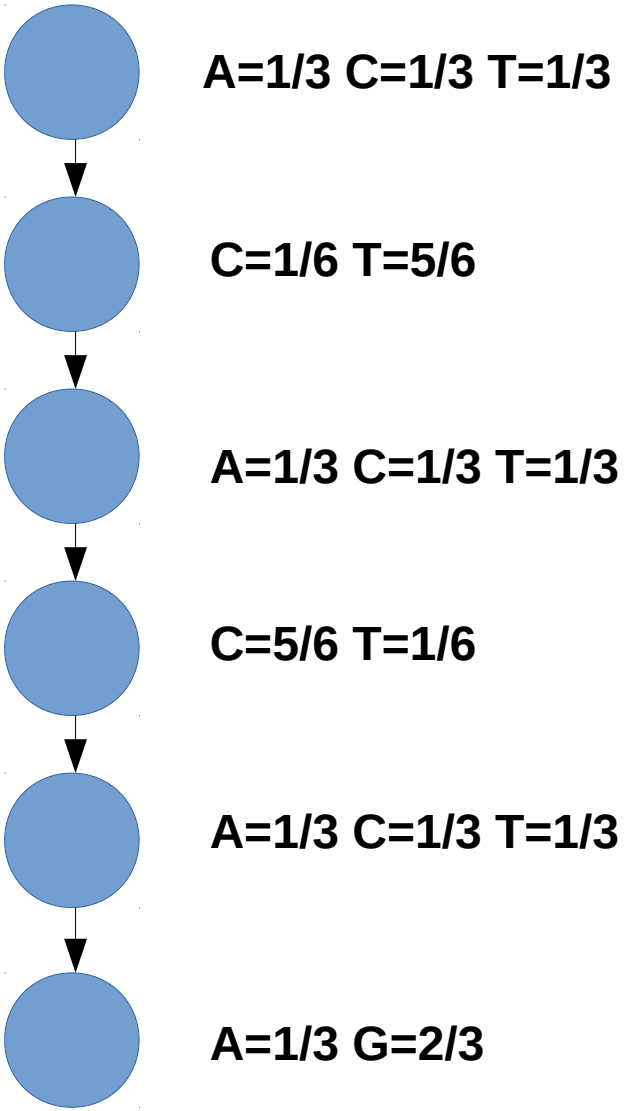
1/6*1*

1/3*1*

1/6*1*

1/3*1*

1/3*1



HMM and score

A C A C A A

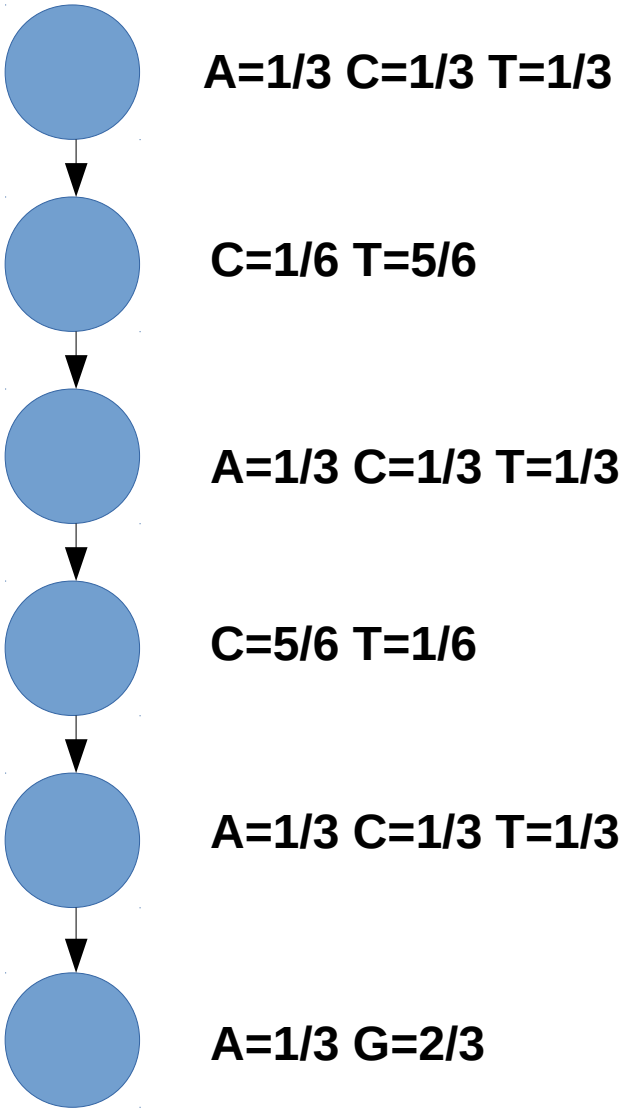
A T A C A A

T T T C T G

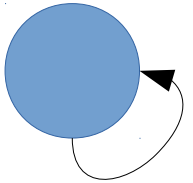
T T T C T G

C T C C C G

C T C T C G



Background model



A=1/4 C=1/4
G=1/4 T=1/4

HMM and score

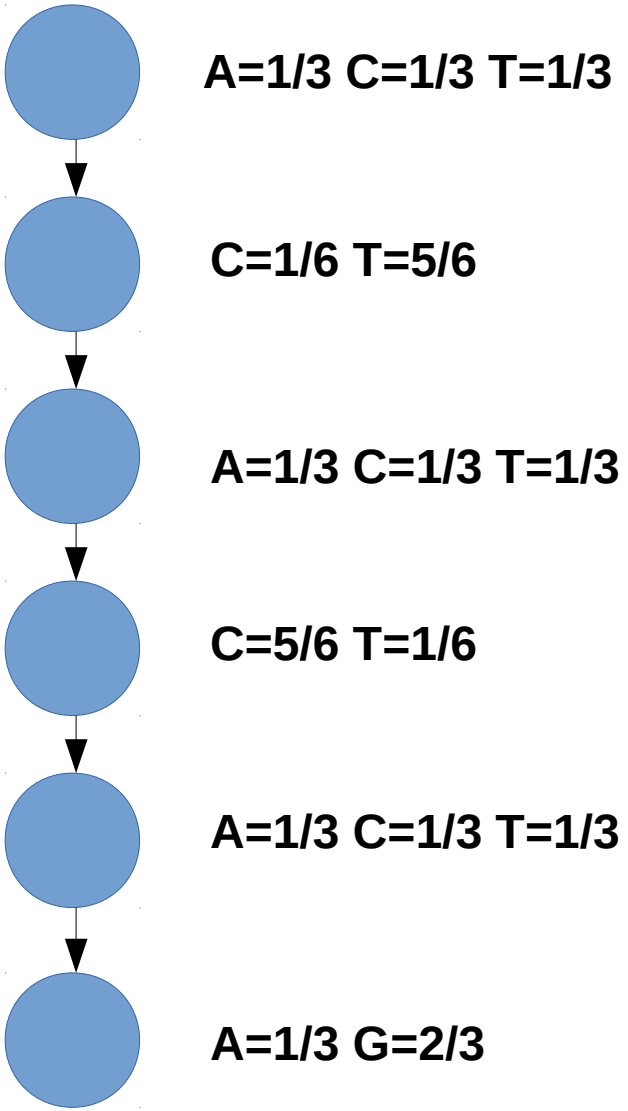
S = CCATAA

Score(S) =

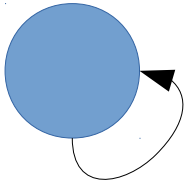
log((1/3)/(1/4))+

log((1/6)/(1/4))+

….



Background model



A=1/4 C=1/4
G=1/4 T=1/4

HMM profile (insertions/deletions)

A C A C A A

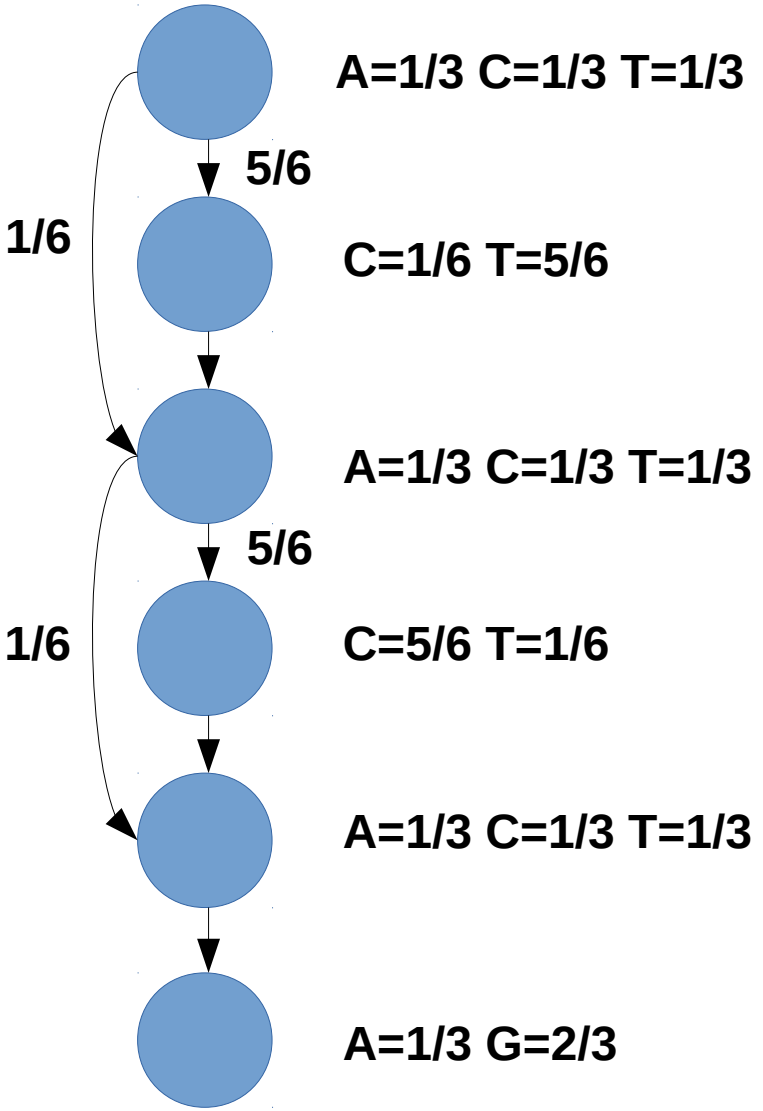
A T A - A A

T T T C T G

T T T C T G

C T C C C G

C - C T C G



HMM profile (insertions/deletions)

A C A C A - A

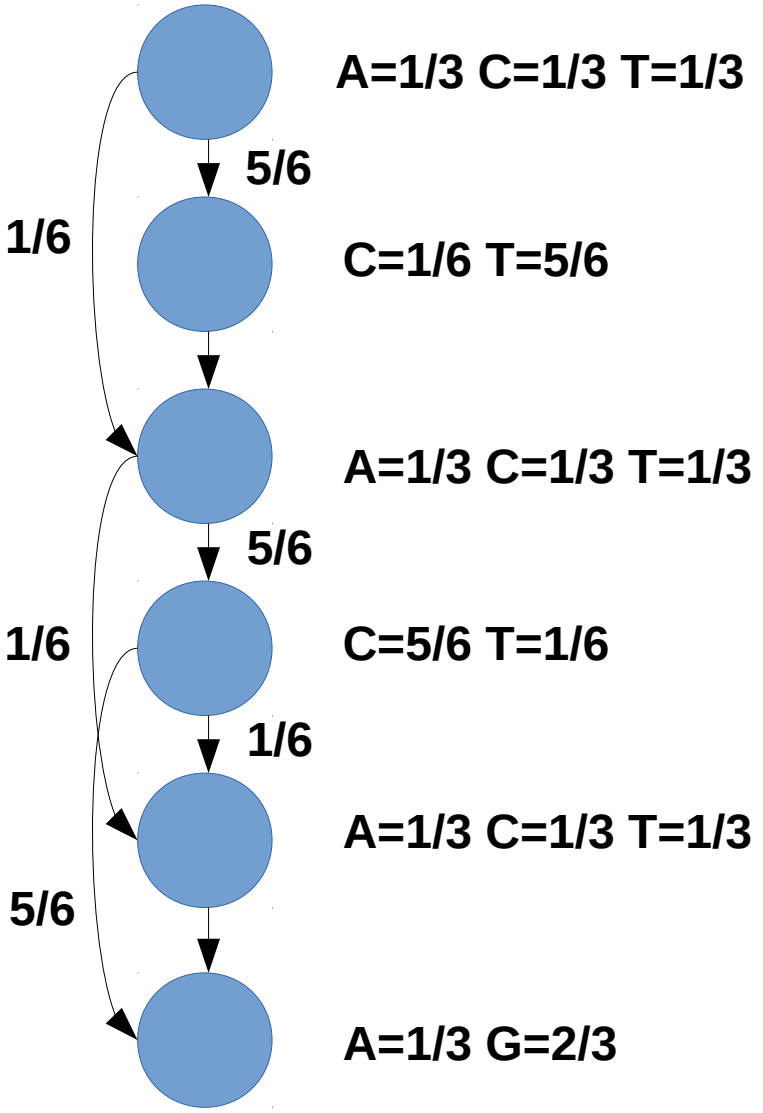
A T A - A - A

T T T C T - G

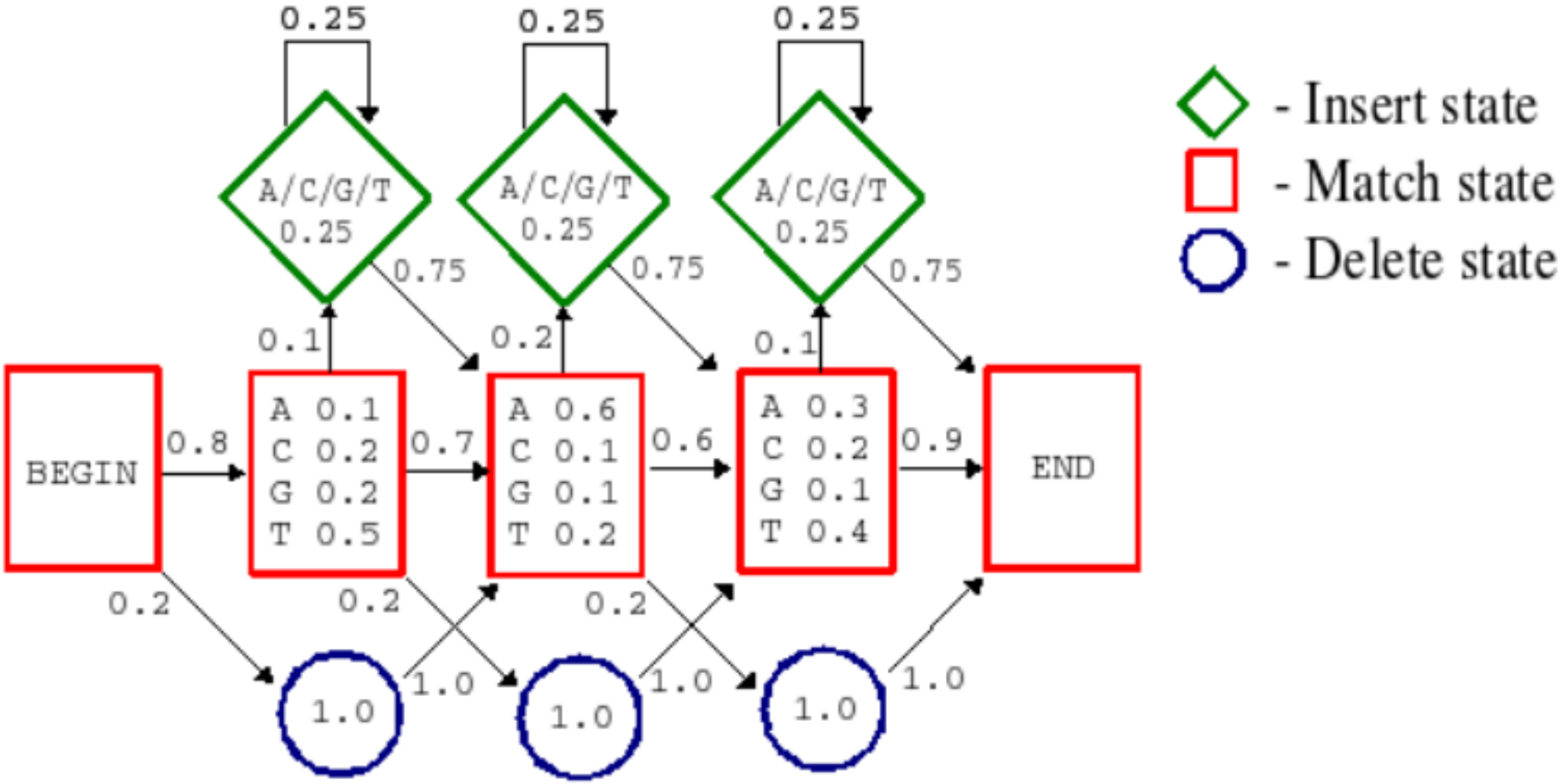
T T T C T - G

C T C C C - G

C - C T C C G

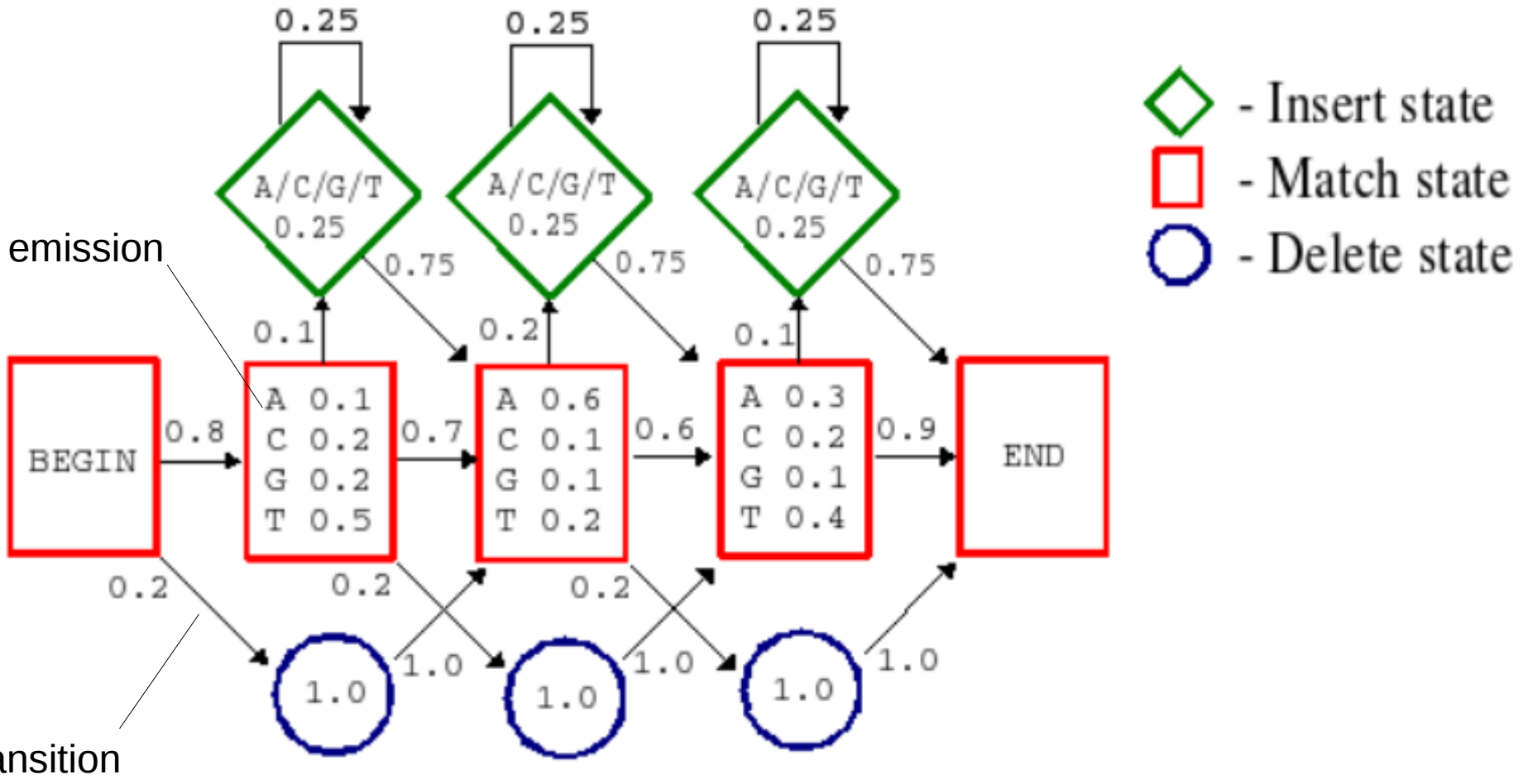


HMM profile



http://www.bioinfo.ifm.liu.se/edu/TFTB29/HT2013/assignment3.html

HMM profile

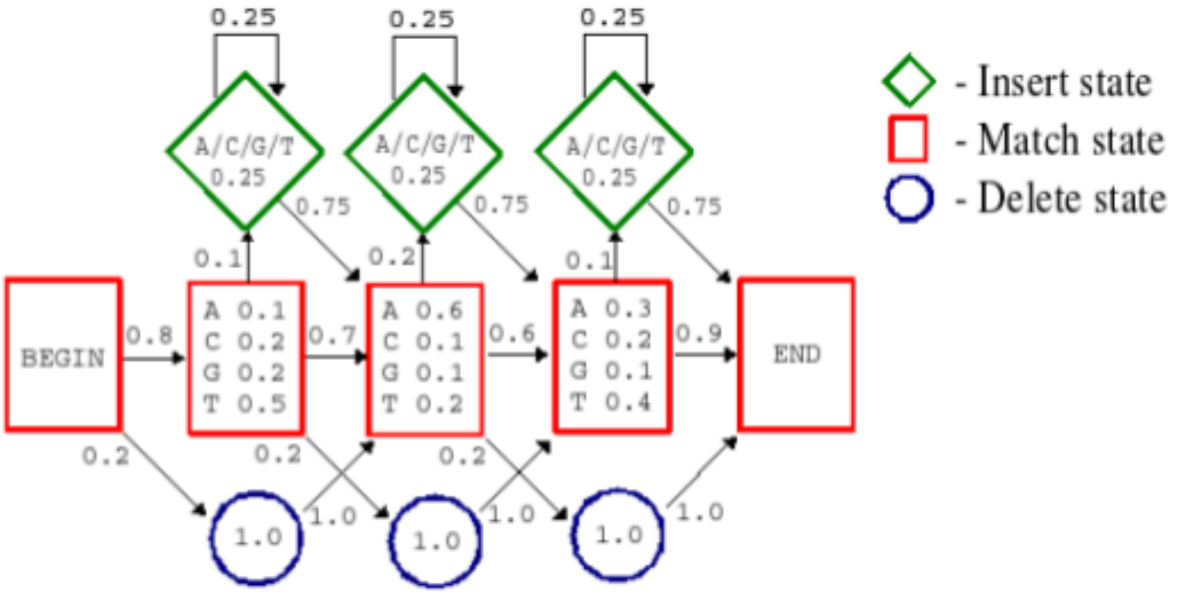


http://www.bioinfo.ifm.liu.se/edu/TFTB29/HT2013/assignment3.html

HMM profile

- Match a sequence to the HMM profile

Model

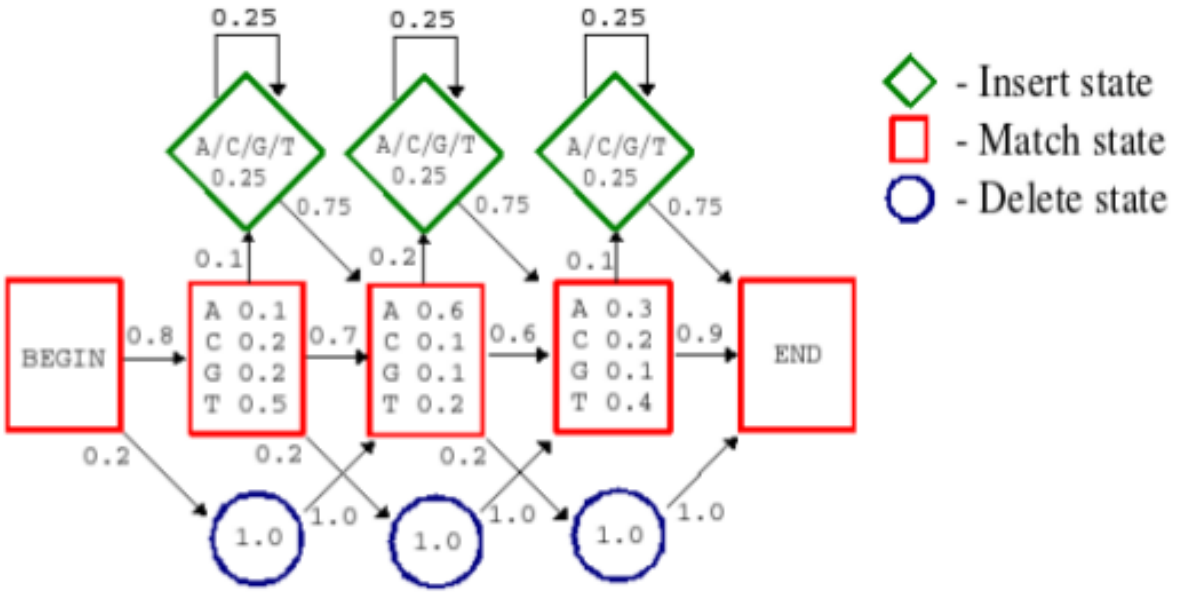


S: ATG → P(S | Model)=?

HMM profile

- Match a sequence to the HMM profile

Model

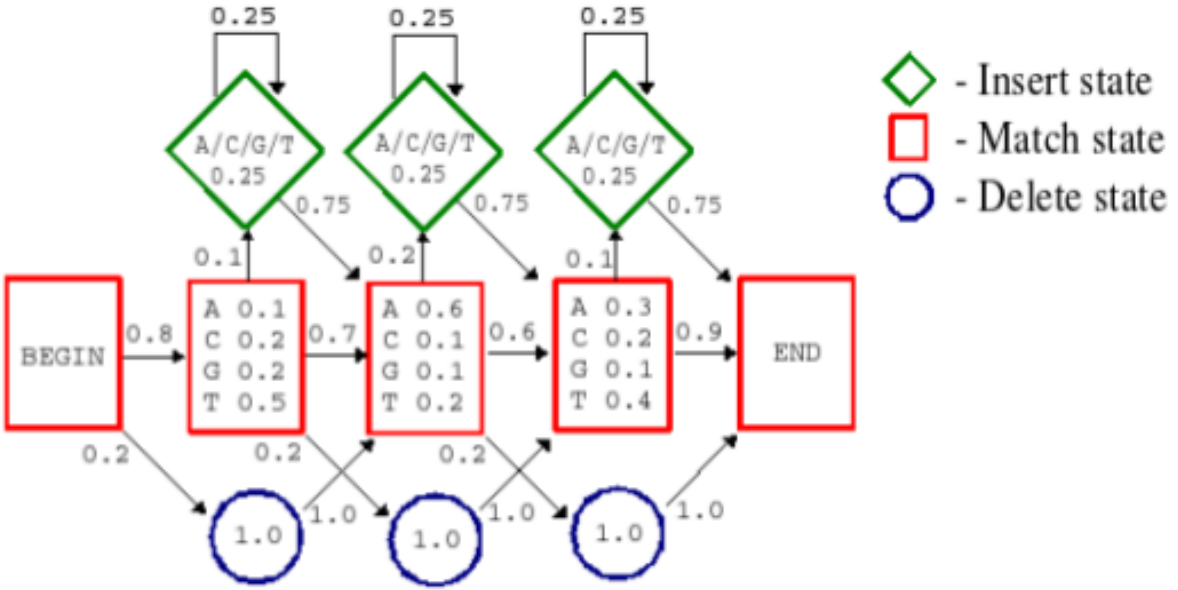


S: ATG → P(S | Model)=0.8*0.1*0.7*0.2*0.6*0.1*0.9 = 0.0006048 BMMME

HMM profile

- Match a sequence to the HMM profile

Model



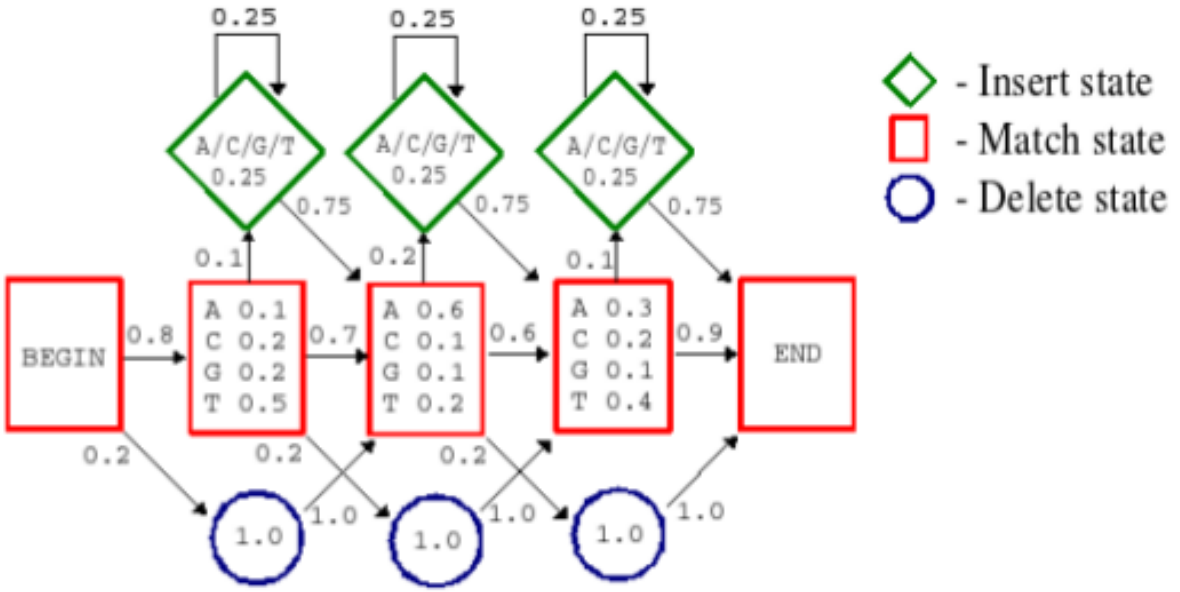
S: ATG → P(S | Model)=0.8*0.1*0.7*0.2*0.6*0.1*0.9 = 0.0006048 BMMME

P(S | Model)=0.8*0.1*0.2*0.4*0.25*0.75 = 0.0012 BMDMIE

HMM profile

- Match a sequence to the HMM profile

Model



S: ATG → P(S | Model)=0.8*0.1*0.7*0.2*0.6*0.1*0.9 = 0.0006048 BMMME

P(S | Model)=0.8*0.1*0.2*0.4*0.25*0.75 = 0.0012 BMDMIE

P(S | Model)=0.2*0.6*0.2*0.25*0.75*0.1*0.9 = 0.000405 BDMIME

HMM profile

- What is the probability that a given sequence S was generated by the HMM?

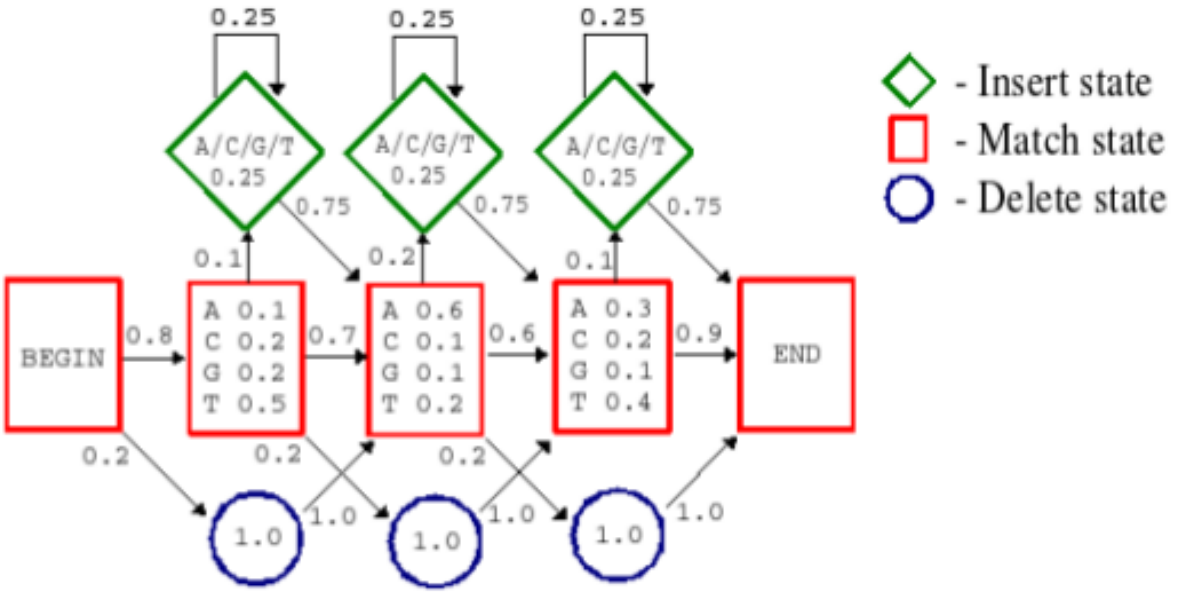
$$P(S \mid w) = \sum_{\pi} P(S, \pi \mid w)$$

  - S = sequence
  - w = parameters (probabilities)
  - π = all possible paths
- Computationally inefficient
  - There are efficient algorithms
  - Forward-Backwad and Viterbi

HMM profile

- Match a sequence to the HMM profile

Model



S: ATG → P(S | Model)=?

Finding the path with highest probability means to find the best alignment to the HMM profile