Motif finding as an application of the EM-algorithm

Axel Wintsche

November 25, 2011

Sequences and probabilities

given sequence	S=`ACCAGAT`
probability	$P(S) = ?$

Sequences and probabilities

given sequence	S=$\texttt{ACCAGAT}$
probability	$P(S) = ?$
for example	$P(S) = P(\texttt{A})P(\texttt{C})P(\texttt{C})P(\texttt{A})P(\texttt{G})P(\texttt{A})P(\texttt{T})$

Sequences and probabilities

given sequence	S=ACCAGAT
probability	$P(S) = ?$
for example	$P(S) = P(\mathtt{A})P(\mathtt{C})P(\mathtt{C})P(\mathtt{A})P(\mathtt{G})P(\mathtt{A})P(\mathtt{T})$

if $P(\mathtt{A}) = P(\mathtt{C}) = P(\mathtt{G}) = P(\mathtt{T}) = 0.25$
then $P(S) = 0.25^7 = 6.1035 \times 10^{-5}$

Sequences and probabilities

given sequence	S=$\mathtt{ACCAGAT}$
probability	$P(S) = ?$
for example	$P(S) = P(\mathtt{A})P(\mathtt{C})P(\mathtt{C})P(\mathtt{A})P(\mathtt{G})P(\mathtt{A})P(\mathtt{T})$

if $P(\mathtt{A}) = P(\mathtt{C}) = P(\mathtt{G}) = P(\mathtt{T}) = 0.25$ $\Rightarrow$ Distribution θ

then $P(S) = 0.25^7 = 6.1035 \times 10^{-5}$ $\Rightarrow P(S|\theta)$

Sequences and probabilities

given sequence	$S = \texttt{ACCAGAT}$
probability	$P(S) = ?$
for example	$P(S) = P(\texttt{A})P(\texttt{C})P(\texttt{C})P(\texttt{A})P(\texttt{G})P(\texttt{A})P(\texttt{T})$

if $P(\texttt{A}) = P(\texttt{C}) = P(\texttt{G}) = P(\texttt{T}) = 0.25$ $\Rightarrow$ Distribution θ
then $P(S) = 0.25^7 = 6.1035 \times 10^{-5}$ $\Rightarrow P(S|\theta)$

if $P(\texttt{A}) = P(\texttt{T}) = 0.2$ and $P(\texttt{C}) = P(\texttt{G}) = 0.3$
then $P(S) = \ldots$

PFM as probability model

PFM:

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

PFM as probability model

PFM:

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

$\Leftarrow$

Alignment:

CCCATTGTTCTC
TTTCTGGTTCTC
TCAATTGTTTAG
CTCATTGTTGTC
TCCATTGTTCTC
CCTATTGTTCTC
TCCATTGTTCGT
CCAATTGTTTTG

PFM as probability model

PFM:

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

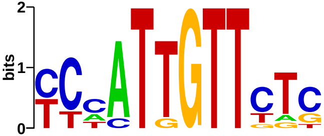
$\Leftarrow$

Alignment:

CCCATTGTTCTC
TTTCTGGTTCTC
TCAATTGTTTAG
CTCATTGTTGTC
TCCATTGTTCTC
CCTATTGTTCTC
TCCATTGTTCGT
CCAATTGTTTTG

$\Downarrow$

Sequence logo:



Sequences with a motif

given:

- sequence S
- S contains exactly one motif m
- distribution θ_S for the sequence
- PFM θ_{PFM} for the motif

$P(S|\theta_{PFM}, \theta_S) = ?$

if $S = \mathtt{A}\underline{\mathtt{AAB}}\mathtt{B}$ and $m = \mathtt{AAB}$

Sequences with a motif

given:

- sequence S
- S contains exactly one motif m
- distribution θ_S for the sequence
- PFM θ_{PFM} for the motif

$P(S|\theta_{PFM}, \theta_S) = ?$

if $S = \mathtt{A}\underline{\mathtt{AAB}}\mathtt{B}$ and $m = \mathtt{AAB}$

then $P(S|\theta_{PFM}, \theta_S) = P(\mathtt{A}|\theta_S) \times P(\mathtt{AAB}|\theta_{PFM}) \times P(\mathtt{B}|\theta_S)$

Sequences with a motif

now for a set of n sequences $S = S_1, S_2, \ldots, S_n$

$P(S|\theta_{PFM}, \theta_S) =$

Sequences with a motif

now for a set of n sequences $S = S_1, S_2, \ldots, S_n$

$P(S|\theta_{PFM}, \theta_S) =$

$P(S_1|\theta_{PFM}, \theta_S) \times P(S_2|\theta_{PFM}, \theta_S) \times \ldots \times P(S_n|\theta_{PFM}, \theta_S)$

Sequences with a motif

now for a set of n sequences $S = S_1, S_2, \ldots, S_n$

$P(S|\theta_{PFM}, \theta_S) =$

$P(S_1|\theta_{PFM}, \theta_S) \times P(S_2|\theta_{PFM}, \theta_S) \times \ldots \times P(S_n|\theta_{PFM}, \theta_S)$

more formal we calculate $P(S, h|\theta_{PFM}, \theta_S)$
h are the positions of the motif

Sequences with a motif

now for a set of n sequences $S = S_1, S_2, \ldots, S_n$

$P(S|\theta_{PFM}, \theta_S) =$

$P(S_1|\theta_{PFM}, \theta_S) \times P(S_2|\theta_{PFM}, \theta_S) \times \ldots \times P(S_n|\theta_{PFM}, \theta_S)$

more formal we calculate $P(S, h|\theta_{PFM}, \theta_S)$
h are the positions of the motif

Problem: we don't know the positions of the motif

EM-algorithm

Motivation

optimize model parameters θ
e.g., find parameters so that $\widehat{P}(S|\theta)$ is maximal

EM-algorithm

Motivation

optimize model parameters θ
e.g., find parameters so that $\widehat{P}(S|\theta)$ is maximal

Concepts:

- observed and hidden data
- iteration of
  1. E-step
  2. M-step

Expectation value

if we know:

- all outcomes x_i of a discrete random variable X
- the probability $P(x_i)$ of each outcomes

the expectation value of X is defined as

$$E[X] = \sum_i x_i P(x_i)$$

Expectation value

if we know:

- all outcomes x_i of a discrete random variable X
- the probability $P(x_i)$ of each outcomes

the expectation value of X is defined as

$$E[X] = \sum_i x_i P(x_i)$$

Example: rolling a dice

E-step

calculates the expectation value of $E[P(S, h|\theta_{PFM}, \theta_S)]$

simplified:

- outcome: a probability for every start position h_i
- probability of $P(h_i)$ is uniformly distributed

E-step

calculates the expectation value of $E[P(S, h|\theta_{PFM}, \theta_S)]$

simplified:

- outcome: a probability for every start position h_i
- probability of $P(h_i)$ is uniformly distributed

Example for S_1 and $\theta = (\theta_{PFM}, \theta_S)$

$$E[P(S_1, h|\theta)] = \quad P(\mathtt{AAA}|\theta_{PFM})P(\mathtt{B}|\theta_S)P(\mathtt{B}|\theta_S)$$

E-step

calculates the expectation value of $E[P(S, h|\theta_{PFM}, \theta_S)]$

simplified:

- outcome: a probability for every start position h_i
- probability of $P(h_i)$ is uniformly distributed

Example for S_1 and $\theta = (\theta_{PFM}, \theta_S)$

$$\begin{aligned} E[P(S_1, h|\theta)] = &\quad P(\mathtt{AAA}|\theta_{PFM})P(\mathtt{B}|\theta_S)P(\mathtt{B}|\theta_S) \\ + &\quad P(\mathtt{A}|\theta_S)P(\mathtt{AAB}|\theta_{PFM})P(\mathtt{B}|\theta_S) \end{aligned}$$

E-step

calculates the expectation value of $E[P(S, h|\theta_{PFM}, \theta_S)]$

simplified:

- outcome: a probability for every start position h_i
- probability of $P(h_i)$ is uniformly distributed

Example for S_1 and $\theta = (\theta_{PFM}, \theta_S)$

$$\begin{aligned} E[P(S_1, h|\theta)] = &\quad P(\texttt{AAA}|\theta_{PFM})P(\texttt{B}|\theta_S)P(\texttt{B}|\theta_S) \\ + &\quad P(\texttt{A}|\theta_S)P(\texttt{AAB}|\theta_{PFM})P(\texttt{B}|\theta_S) \\ + &\quad P(\texttt{A}|\theta_S)P(\texttt{A}|\theta_S)P(\texttt{ABB}|\theta_{PFM}) \end{aligned}$$

E-step

calculates the expectation value of $E[P(S, h|\theta_{PFM}, \theta_S)]$

simplified:

- outcome: a probability for every start position h_i
- probability of $P(h_i)$ is uniformly distributed

Example for S_1 and $\theta = (\theta_{PFM}, \theta_S)$

$$
\begin{aligned}
E[P(S_1, h|\theta)] = \quad & P(\texttt{AAA}|\theta_{PFM})P(\texttt{B}|\theta_S)P(\texttt{B}|\theta_S) && \times 1/3 \\
+ \quad & P(\texttt{A}|\theta_S)P(\texttt{AAB}|\theta_{PFM})P(\texttt{B}|\theta_S) && \times 1/3 \\
+ \quad & P(\texttt{A}|\theta_S)P(\texttt{A}|\theta_S)P(\texttt{ABB}|\theta_{PFM}) && \times 1/3
\end{aligned}
$$

M-step

- maximizes the expected value $E[P(S, h|\theta_{PFM}, \theta_S)]$ over the model parameters of θ_{PFM} and θ_S
- see example...