Motif finding as an application of the EM-algorithm

Axel Wintsche

November 25, 2011
Sequences and probabilities

given sequence \( S = \text{ACCAGAT} \)

probability \( P(S) = ? \)
Sequences and probabilities

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

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

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

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

\[ P(S|\theta) \]
Sequences and probabilities

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

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

if \( P(A) = P(T) = 0.2 \) and \( P(C) = P(G) = 0.3 \)
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

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>0 2 7 0 0 0 0 0 1 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>4 6 4 1 0 0 0 0 5 0 5</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0 0 0 0 0 1 8 0 0 1 1 2</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>4 2 2 0 8 7 0 8 8 2 6 1</td>
<td></td>
</tr>
</tbody>
</table>

Alignment:

```
CCCATTGTTCTC
TTTCTGGTTCTC
TCAATTGTTTAG
CTCATTGGTC
TCCATTGGTCTC
CCTATTGGTCTC
TCCATTGGTCGT
CCAATTGTTTGT
```
PFM as probability model

PFM:

\[
\begin{align*}
A & \quad 002700000010 \\
C & \quad 464100000505 \\
G & \quad 000001800112 \\
T & \quad 422087088261 \\
\end{align*}
\]

\[\downarrow\]

Sequence logo:

Alignment:

\[
\begin{align*}
CCCATTGTTCTC \\
TTTCTGGTTCCTC \\
TCAATTGTTTAG \\
CTCAATTGTTGTC \\
TCCATTGTTCTC \\
CTCTATTGTTCTC \\
TCCATTGTTCGT \\
CCCAATTGGTTTG \\
\end{align*}
\]

Axel Wintsche
Motif finding as an application of the EM-algorithm
Sequences with a motif

given:
- sequence $S$
- $S$ contains exactly one motif $m$
- distribution $\theta_S$ for the sequence
- PFM $\theta_{PFM}$ for the motif

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

if $S = \underline{AAABB}$ and $m = \underline{AAB}$
Sequences with a motif

given:

- sequence $S$
- $S$ contains exactly one motif $m$
- distribution $\theta_S$ for the sequence
- PFM $\theta_{PFM}$ for the motif

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

If $S = \underline{AAABB}$ and $m = AAB$

Then $P(S|\theta_{PFM}, \theta_S) = P(A|\theta_S) \times P(AAB|\theta_{PFM}) \times P(B|\theta_S)$
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 $\theta$
e.g., find parameters so that $\hat{P}(S|\theta)$ is maximal
EM-algorithm

Motivation

optimize model parameters $\theta$
e.g., find parameters so that $\hat{P}(S|\theta)$ 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 outcome

the expectation value of $X$ is defined as

$$E[X] = \sum_i x_i P(x_i)$$
if we know:
- all outcomes $x_i$ of a discrete random variable $X$
- the probability $P(x_i)$ of each outcome

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 $\theta = (\theta_{PFM}, \theta_S)$

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

$$E[P(S_1, h|\theta)] = P(\text{AAA}|\theta_{PFM})P(\text{B}|\theta_S)P(\text{B}|\theta_S)$$
$$+ P(\text{A}|\theta_S)P(\text{AAB}|\theta_{PFM})P(\text{B}|\theta_S)$$
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 $\theta = (\theta_{PFM}, \theta_S)$

$$E[P(S_1, h|\theta)] = P(\text{AAA}|\theta_{PFM})P(\text{B}|\theta_S)P(\text{B}|\theta_S) + P(\text{A}|\theta_S)P(\text{AAB}|\theta_{PFM})P(\text{B}|\theta_S) + P(\text{A}|\theta_S)P(\text{A}|\theta_S)P(\text{ABB}|\theta_{PFM})$$
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 $\theta = (\theta_{PFM}, \theta_S)$

$$E[P(S_1, h|\theta)] = P(\text{AAA}|\theta_{PFM})P(\text{B}|\theta_S)P(\text{B}|\theta_S) \times \frac{1}{3}$$
$$+ P(\text{A}|\theta_S)P(\text{AAB}|\theta_{PFM})P(\text{B}|\theta_S) \times \frac{1}{3}$$
$$+ P(\text{A}|\theta_S)P(\text{A}|\theta_S)P(\text{ABB}|\theta_{PFM}) \times \frac{1}{3}$$
M-step

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