Improved Sequence Motif Finding

Axel Wintsche

February 18, 2010

Sequence motifs

Definition

Sequence motifs are recurring patterns in a set of sequences. [1]

Biological meaning

- Sequence motifs in a set of *related* sequences are presumed to have a biological function
- Example 1: genes with similar expression patterns
- Example 2: proteins with similar functional annotation

Sequence motif representation

- $\mathcal{A}$ sequence alphabet, m sequence motif
- simplest case: $m = a_1 \ldots a_m$, $a_i \in \mathcal{A}$
- Example 1: *Eco*R1: GAATTC
- Example 2: *Hin*dII: GTCAAC or GTTAAC
- consensus sequence: $m = u_1 \ldots u_m$, $u_i \in IUPAC$[1]
- Example 3: *Hin*dII GTYAAC

[1] http://www.chem.qmul.ac.uk/iubmb/misc/naseq.html#tab1

Sequence motif representation

Sequence motif as 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

- independence between positions
- total or average frequencies
- Example: Rox1 TF (*Saccharomyces cerevisiae*)

Assumption and the problem

Assumption

The sequences in the set share one or more common sequence motifs that are responsible for the relatedness.

Problem

In a set of related sequences, find one or more sequence motifs so that the motifs are reliable (i.e. fulfill particular constraints).

Available solutions

D'haeseleer [2]:

- enumerative algorithms
- probabilistic optimization
- deterministic optimization

Motivation

We like to find motifs that are responsible for observed behavior.

- assumption: not every motif contributes to relatedness
- additional information to distinguish is needed
- second set of sequences (negative set), different behavior
- motifs not assumed in second set
- How to utilize additional information?

Expectation Maximization (EM) in MEME

- MEME developed by Bailey and Elkan [3]
- model Θ_m (PFM), bg-model Θ_{bg} (frequency vector)
- $\Theta = (\Theta_m, \Theta_{bg})$, dataset X
- ML estimation by iterations of E-step and M-step
- missing data Z, measure if generated by Θ_m or Θ_{bg}
- Z has to be estimated

EM in MEME

$$L(\Theta|X,Z) = P(X,Z|\Theta) = P(X|Z,\Theta) \cdot P(Z|\Theta)$$

E-step

- finds the expected value for $P(Z|\Theta)$
- $Z_{s,k} = score(\Theta_m, X_s[k, k+w-1]) \cdot \frac{1}{c}$

EM in MEME

$$L(\Theta|X,Z) = P(X,Z|\Theta) = P(X|Z,\Theta) \cdot P(Z|\Theta)$$

E-step

- finds the expected value for $P(Z|\Theta)$
- $Z_{s,k} = score(\Theta_m, X_s[k, k+w-1]) \cdot \frac{1}{c}$

M-step

- finds the Θ_m that maximizes $L(\Theta|X,Z)$
- $\Theta_m(i,j) = \left(\sum_{s=1}^{|X|} \sum_{k=1}^{|X_s|-w+1} Z_{s,k} \delta_{a_j}^{X_s[k+i]} \right) \cdot \frac{1}{c}$

Alternative EM iteration

original MEME:

1. $Z = \text{e-step}(\Theta_m, X)$
2. $\Theta_m = \text{m-step}(Z, X)$

Alternative EM iteration

original MEME:

1. $Z = \text{e-step}(\Theta_m, X)$
2. $\Theta_m = \text{m-step}(Z, X)$

MEME with negative data:

1. $Z_{neg} = \text{e-step}(\Theta_{pos}, X_{neg})$
2. $\Theta_{neg} = \text{m-step}(Z_{neg}, X_{neg})$
3. $Z_{pos} = \text{e-step'}(\Theta_{pos}, \Theta_{neg}, X_{pos})$
4. $\Theta_{pos} = \text{m-step}(Z_{pos}, X_{pos})$

Alternative EM iteration

original MEME:

1. $Z = \text{e-step}(\Theta_m, X)$
2. $\Theta_m = \text{m-step}(Z, X)$

MEME with negative data:

1. $Z_{neg} = \text{e-step}(\Theta_{pos}, X_{neg})$
2. $\Theta_{neg} = \text{m-step}(Z_{neg}, X_{neg})$
3. $Z_{pos} = \text{e-step'}(\Theta_{pos}, \Theta_{neg}, X_{pos})$
4. $\Theta_{pos} = \text{m-step}(Z_{pos}, X_{pos})$

e-step'

- $\widehat{Z}_{s,k} = Z_{s,k} \cdot (1 - score(\Theta_{neg}, X_s[k, k + w - 1])) \cdot \frac{1}{c}$

Preliminary results

Simulation:

- motif1 = TGAAAA, motif2 = CCGTTT
- 100 random sequences, equal letter frequency
- X_{pos} : 50 sequences, all contain motif1, 45 also motif2
- X_{neg} : 50 sequences, all contain motif1, 5 also motif2

Preliminary results

Simulation:

- motif1 = TGAAAA, motif2 = CCGTTT
- 100 random sequences, equal letter frequency
- X_{pos} : 50 sequences, all contain motif1, 45 also motif2
- X_{neg} : 50 sequences, all contain motif1, 5 also motif2

Results:

- MEME:
  1. TGAAAA
  2. CCGTTT
- MEME with negative data:
  1. CCGTTT
  2. GAAAAN

Remarks

- fine-tuning of MEME parameters
- test method with real data
- weighted negative data integration
- drawback: works only on small data sets

Literature


P. D'haeseleer.
What are DNA sequence motifs?
Nature biotechnology, 24(4):423–426, 2006.


P. D'haeseleer.
How does DNA sequence motif discovery work?
Nature biotechnology, 24(8):959–962, 2006.


T.L. Bailey and C. Elkan.
Fitting a mixture model by expectation maximization to discover motifs in biopolymers.
2(1553-0833):28–36, 1994.

Thanks to Sonja

Thank you for your attention!