Interaktionen von RNAs und Proteinen

Sonja Prohaska

Computational EvoDevo
Universitaet Leipzig

May 18, 2018

Motif Discovery

Assume...
a DNA-binding protein is known
but its binding motif is NOT...

... ChIP-seq data are provided,
read mapping and peak calling has already been done,
but the peaks are still sequences of lenght 100-200nt.

What to do next?

Can we derive the motif from the sequences?

to discover – entdecken; genau das ist hier gefragt

Yes we can! – MEME

Multiple Expectation Maximization for Motif Elicitation

- **Problem:** Given a set of (unaligned) sequence of variable length
  - characterize – derive a motif model
  - identify – matching positions in the sequences
- **Assumptions:**

  motif is contiguous
  - no instertions or deletions
  - all instances of a motif have the same length

  motif positions are independent

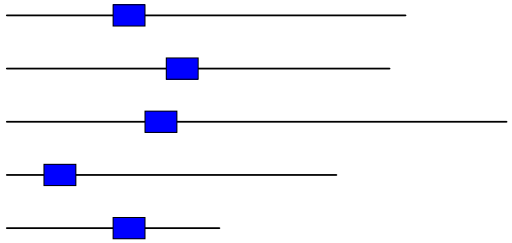
find the most probable non-overlapping shared motif(s).

Types of Possible Models

Motif occurences within a set of sequences

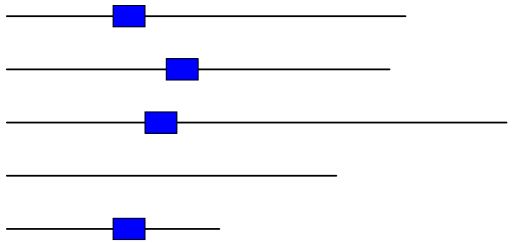
OOPS

- exactly one per sequence
- motif in all sequences



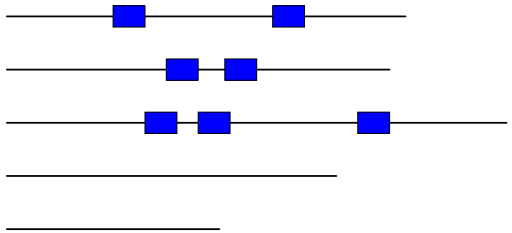
ZOOPS

- zero or one per sequence
- motif in subset of sequences



Two-Component Mixture – TCM

- any number of motifs
- motif in subset of sequences



MEME: an expectation-maximization (EM) algorithm

Initialization

- choose an initial motif model wisely
- e.g. use the most frequent k-mer(s) to build a motif model

Expectation Step

- estimate the probability of finding the site
- at any position in each of the sequences

Maximixation Step

- use the probabilities to improve/update the motif
- do re-normalization

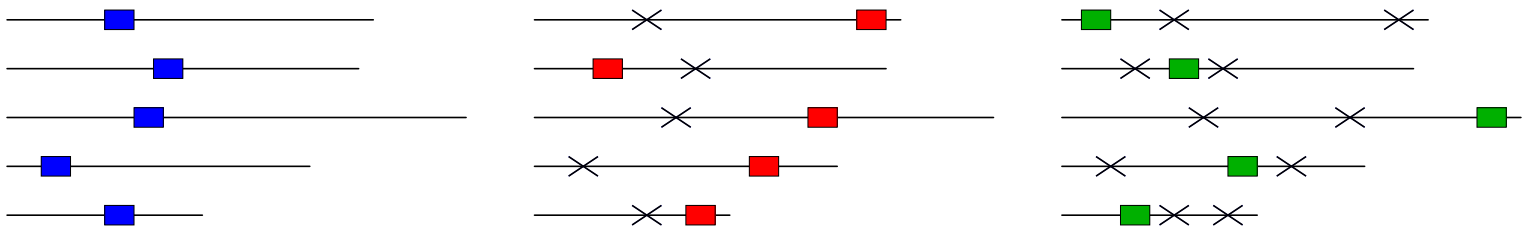
Interation

- Iterate E an M step until motif model does not change anymore.
- Use the model to find the highest scoring motif site for each sequence.

Discovery of multiple motifs

What if you want to find a second or third motif?

- detect **all** sits matching the motif
- x-out the motif sites from all previous run(s)
- run MEME again



This way a ranked list of motifs can be derived.

What needs to be considered when running MEME?

- What's the best initial guess for the motif model?
- What's the best motif size?
- consider the model OOPS, ZOOPS, TCM
- consider forward and reverse direction of non-palindromic motifs
- consider adding a prior probability (additional knowledge) to the probability distribution

  e.g. downweight low complex regions

  e.g. randomize the sequences, predict motifs, downweight common motifs

MEME will always return something! from almost nothing!