Detecting Structured Motifs From DNA Sequences

Sündüz Keleş

Division of Biostatistics, U.C. Berkeley

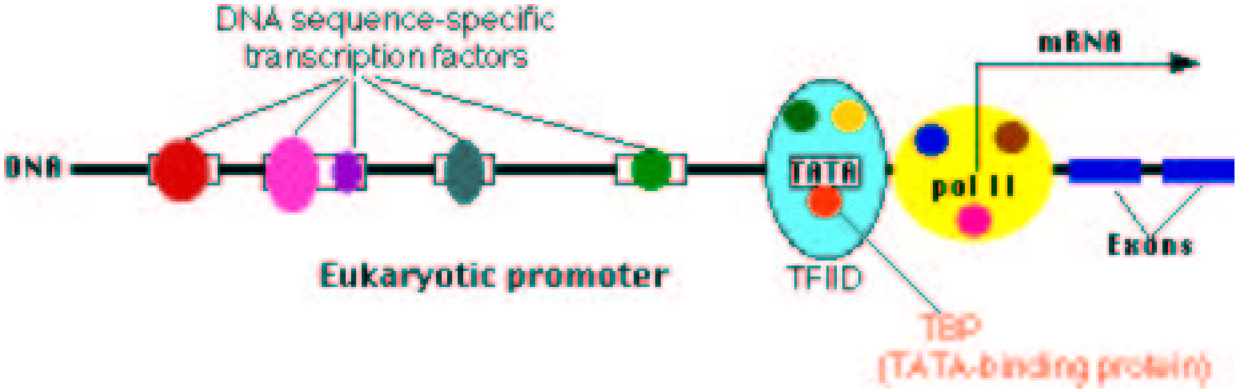
Joint work with Mark van der Laan, Sandrine Dudoit, Michael B. Eisen, and Biao Xing

Outline

- Introduction.
- One motif per sequence model (oops).
- Motivation for our approach.
- Our approach: Constraint entropy model (c.oops,c.zoops).
- Results.
- Conclusions.

Eukaryotic Gene Regulation

- Transcription is regulated by regulatory proteins (transcription factors) binding to elements (motifs) of upstream regions (or promoter regions).

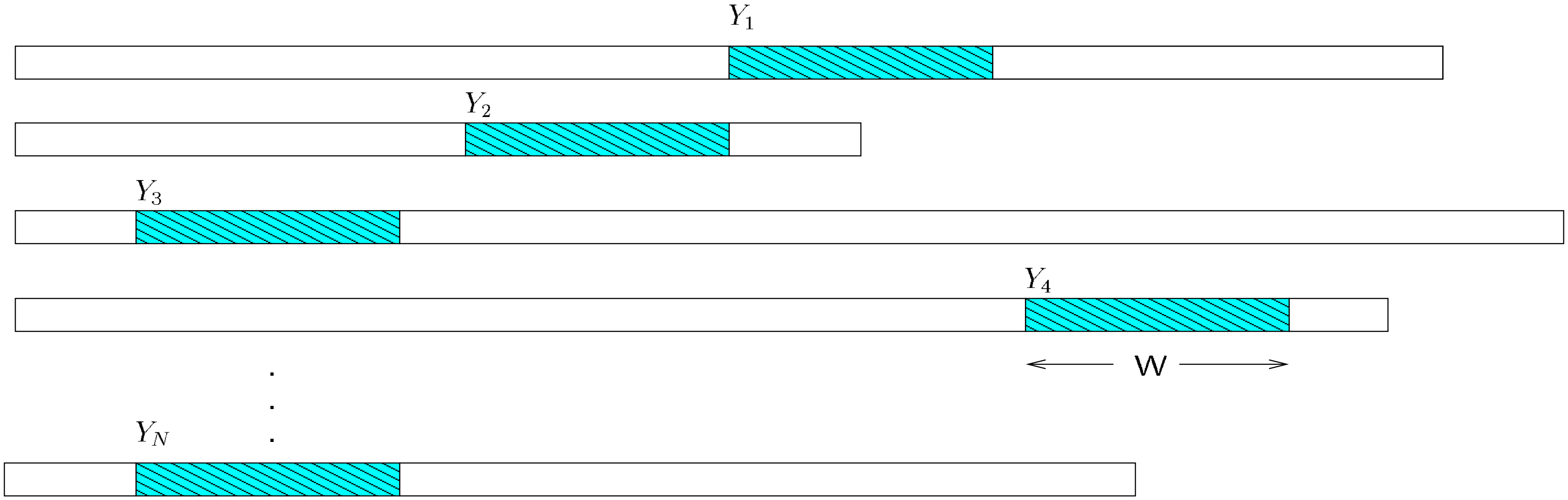

DNA sequence-specific transcription factors
mRNA
DNA
TATA
pol II
Exons
Eukaryotic promoter
TFIID
TBP (TATA-binding protein)

GOAL: Finding motifs (binding sites) from a set of potentially co-regulated genes.

- Some ways of determining potentially co-regulated genes are using (a) scientific knowledge (b) gene expression profiles (c) cross-species comparisons.

Data

N *unaligned* sequences $\overrightarrow{X}_i = (X_{i,1}, \cdots, X_{i,L_i})$, $i = 1, \cdots, N$, where L_i is the length of the ith sequence.



$\Longrightarrow$ Find the common patten. We don't know the start positions Y_i!

If we knew the start positions: Motif Representation

- Example of an aligned motif: ABF1 (from SCPD)

```
TCTCTCGCAACG
TCTCTCGCAACG
TCACGTCACACG
TCACCGCGAACG
TCATAAAGCACG
TCACTAAAGACG
TCAAAATTAACG
TCACTGTACACG
TCACTAACGACG
TCCCCATTAACG
TCACGATACACG
TCATGCGCTACG
TCATGCGCTACG
TCAAATAACAGA
```

- Assume that each position (1) is independent and (2) has a multinomial distribution with $P_w, w = 1, \cdots, W$.

- Position specific probability matrix with motif width $W = 12$.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
A	0	0	0.79	0.14	0.21	0.50	0.21	0.36	0.36	1	0.00	0.07
T	1	0	0.14	0.07	0.43	0.14	0.50	0.21	0.00	0	0.00	0.00
C	0	0	0.00	0.00	0.14	0.21	0.07	0.14	0.21	0	0.07	0.93
G	0	1	0.07	0.79	0.21	0.14	0.21	0.29	0.43	0	0.93	0.00

One motif per sequence model

(Lawrence and Reily (1990); OOPS model of Bailey and Elkan's MEME (1994))

Assumptions

- Sites are distributed independently with

$$\begin{aligned} P_0 &= (p_{01}, \cdots, p_{04}) \quad \text{for background site,} \\ P_w &= (p_{w1}, \cdots, p_{w4}) \quad \text{for position w in the motif, } w \in \{1, \cdots, W\}. \end{aligned}$$

- Unknown start site.

$$Y_{il} = \begin{cases} 1 & \text{if motif starts at position } l \text{ in sequence } i \\ 0 & \text{o.w.} \end{cases}$$

where $l \in \{1, \cdots, L_i - 1 + W\}$. Allow only one motif per sequence $\Sigma_l Y_{il} = 1$.

- Uniform start site distribution.

$$P(Y_{il} = 1) = 1/(L_i - W + 1).$$

Full data log-likelihood

$$\sum_{i=1}^{N} \sum_{l=1}^{L_i - W + 1} I(Y_{il} = 1) \left[\log p(Y_{il} = 1) + \sum_{h \in T_{nl}} \sum_{a=1}^{4} I(S_{n,h} = a) \log p_{a0} + \sum_{w=1}^{W} \sum_{a=1}^{4} I(X_{i,l+w-1} = a) \log p_{aw} \right],$$

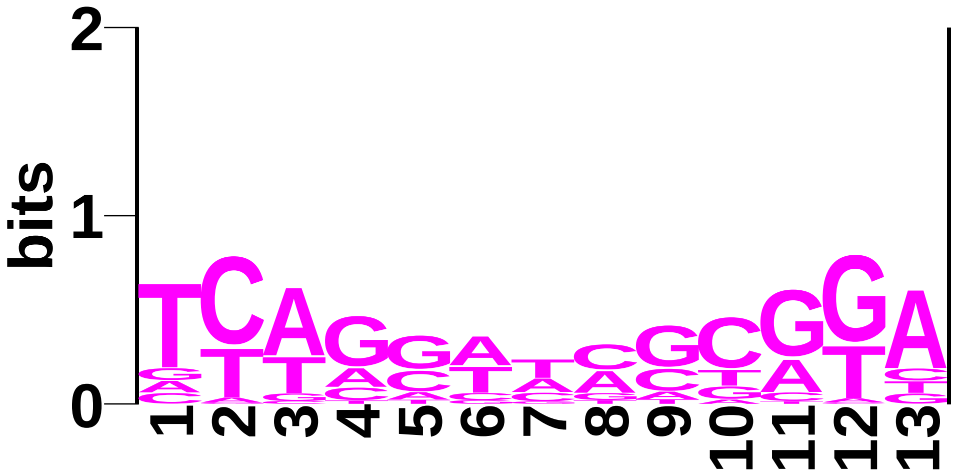
where $T_{il} = \{1, \cdots, L_i\} - \{l, l+1, \cdots, l+W-1\}$, $l \in \{1, \cdots, L_i - W + 1\}$.

Parameter estimation is done with EM algorithm or Gibbs Sampling.

How do methods based on this model perform?

- Pretty good if the motif is well represented in the data in terms of its
  - Frequency,
  - Information content: $IC(w) = 2 -$ Entropy at position w. $Entropy(w) = H(w) = -\Sigma_{j=1}^{4} p_{wj} \log p_{wj}$. $IC(w)$ measures how conserved that position of the motif is.
- Not very successful otherwise (Pevzner & Sze, 2001).
- They also fail when there are other *uninteresting competing* motifs.

→ Example of a weak signal motif:



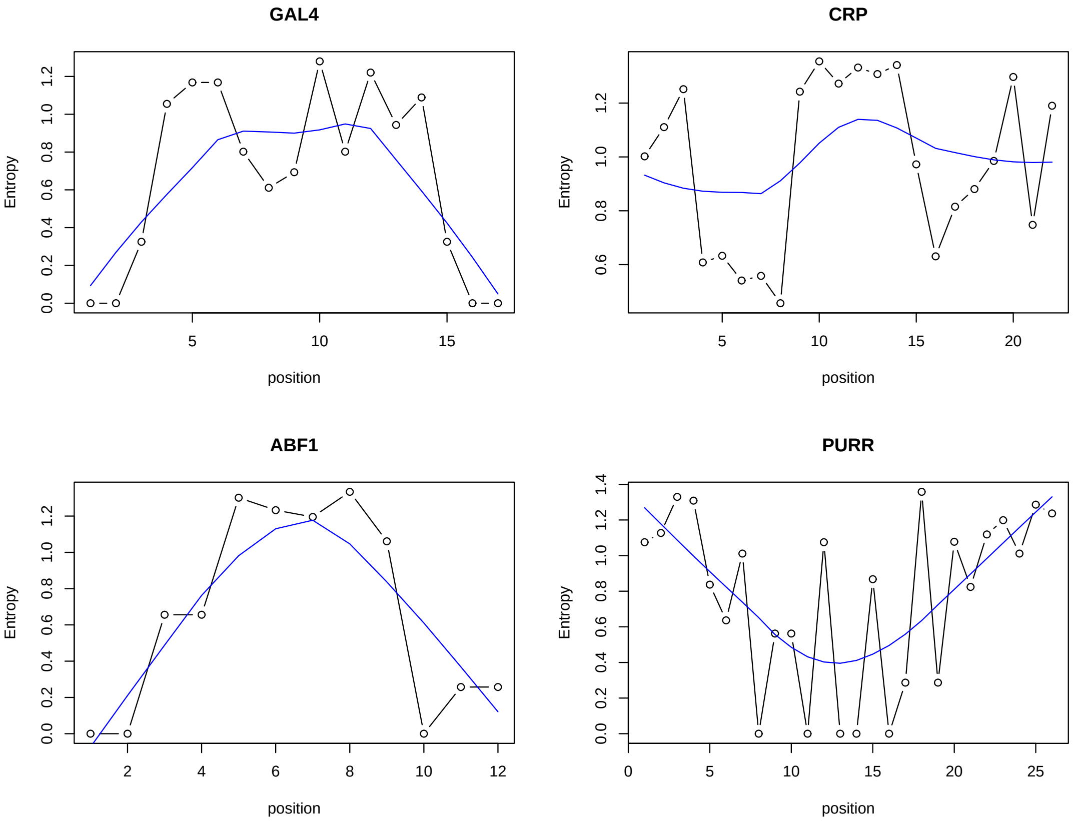
- Anything interesting about this entropy structure?

Structured (Regular) Motifs

Mirny & Gelfand (2002):

- "Base pairs that are have more interaction with the protein are more conserved."

- "If a protein-DNA complex is available but the recognition motif is unknown, one can compute the number of contacts per base pair and predict the most conserved ones" ⟹ Rough idea about the entropy structure!

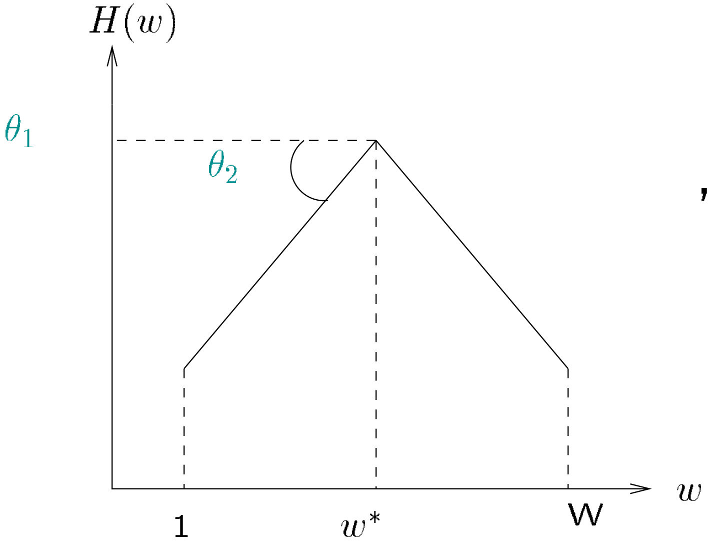


Finding Structured (Regular) Motifs

- Define the regular motif as the motif following a structured entropy. Entropy at position w of the motif equals

$$H(w) = -\sum_{a=1}^{4} p_{aw} \log p_{aw}.$$

- We assume a model $H(w;\theta)$ for $H(w)$, $w = 1, \cdots, W$.
- Different entropy curves define different motif structures e.g.



"Low information in the middle, and higher information towards the ends"

$$H(\theta_1, \theta_2; w) = \theta_1 - |w - w^*| \, \mathrm{ta}^{\mathrm{n}} \, \theta_2$$

$$w = 1, \cdots, W.$$

C.OOPS: M-step of the EM

- Define

$$\zeta_{il} = Pr(Y_{il} = 1 \mid X_i),$$

$$N_{wj} = \begin{cases} \Sigma_{i=1}^{N} \Sigma_{l=1}^{L_i - W + 1} \zeta_{il} I(X_{i,l+w-1} = j), & \text{if} \quad w = 1 \cdots, W, \\ \Sigma_{i=1}^{N} \Sigma_{l=1}^{L_i - W + 1} \zeta_{il} \Sigma_{h \in T_{il}} I(X_{i,h} = j), & \text{if} \quad w = 0. \end{cases}$$

- M-step for the motif parameters is

$$\max \quad \sum_{w=1}^{W} \sum_{j=1}^{4} N_{wj} \log p_{wj}$$

s.t.

$$- \sum_{a=1}^{4} p_{wj} \log p_{wj} = \theta_1 - \delta(w, w^*) \tan \theta_2, \quad w = 1, \cdots, W \qquad (1)$$

$$\sum_{a=1}^{4} p_{wj} = 1, \quad w = 1, \cdots, W$$

$$p_{wj} \geq 0 \quad a = 1, \cdots, 4; w = 1, \cdots, W.$$

where $\delta(w, w^*) = |w - w^*|$.

- Constraint (1) is a entropy structure specific constraint. Maximization w.r.t $\overrightarrow{P_1}, \cdots, \overrightarrow{P_W}$, θ_1 and θ_2 is done with a nonlinear constraint optimization method (Augmented Lagrange Multipliers).

Extended Model: Zero or one motif per sequence (ZOOPS model of MEME)

- Introducing another hiden variable. Let

$$Z_i = \begin{cases} 1 & \text{if sequence i has a copy of the motif,} \\ 0 & \text{o.w.} \end{cases}$$

- Uniform conditional start site distribution,

$$P(Y_{il} = 1 \mid Z_i = 1) = 1/(L_i - W + 1).$$

- Full data log-likelihood equals

$$\begin{aligned} &\sum_{i=1}^{N} I(Z_i = 0)\left[\log(1-\pi) + \sum_{l=1}^{L_i}\sum_{j=1}^{4} I(X_{i,l} = j)\log p_{0j}\right] + \sum_{i=1}^{N} I(Z_i = 1)\log\pi \\ &+ \sum_{i=1}^{N}\sum_{l=1}^{L_i-W+1} I(Z_i = 1, Y_{i,l} = 1)\left[\sum_{h\in T_{il}}\sum_{j} I(X_{i,h} = a)\log p_{0j}\right. \\ &\left. + \sum_{w=1}^{W}\sum_{j=1}^{4} I(X_{i,l+w-1} = j)\log p_{wj} + \log P(Y_{il} = 1 \mid Z_i = 1)\right], \end{aligned}$$

where $\pi = P(Z = 1)$.

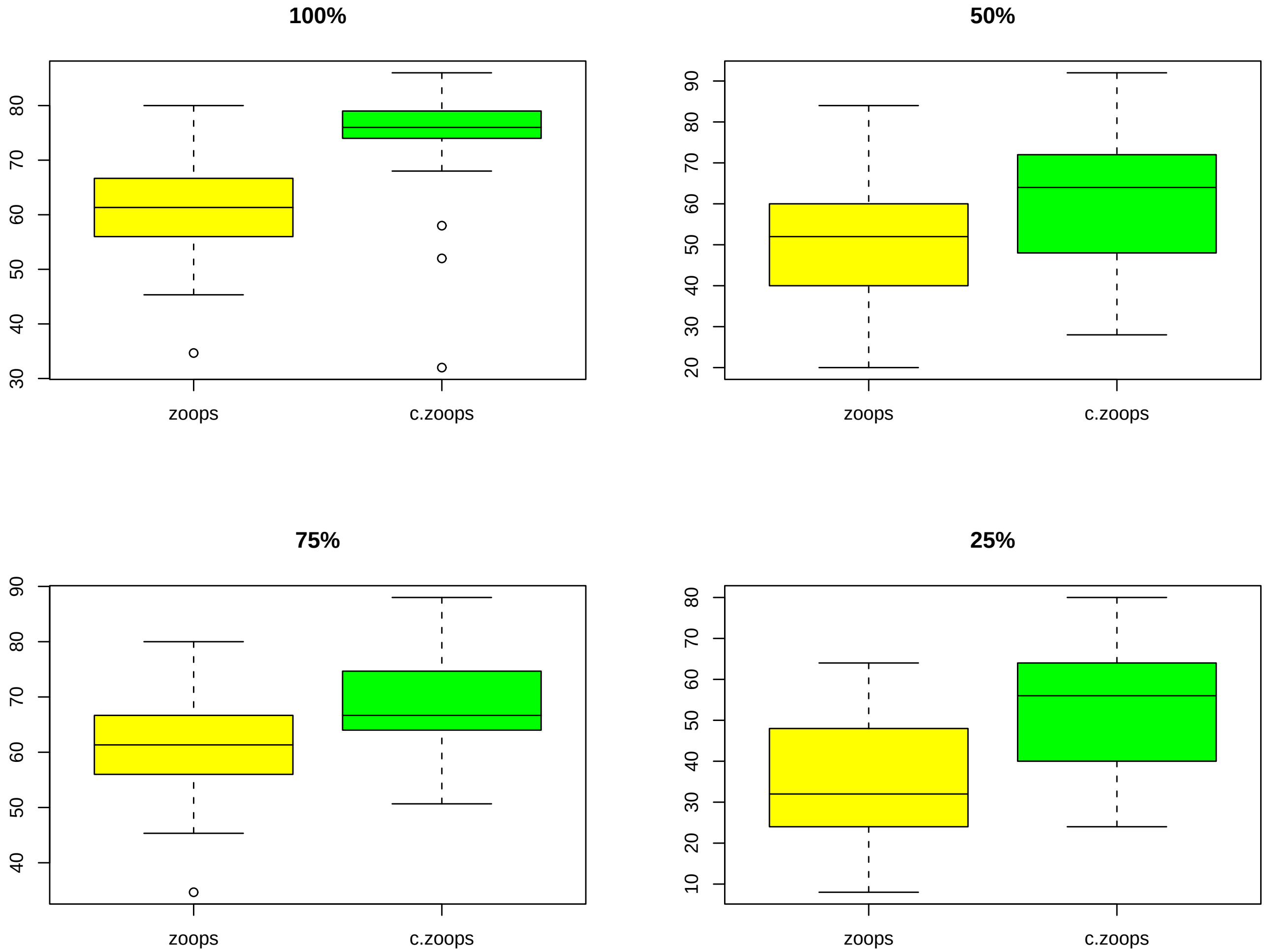
Simulations

- PART 1: Performance in finding the weak regular motif as its frequency varies. Weak Signal.
- PART 2: Performance in finding the regular motif in the existence of a competing irregular motif. Irregular motif is obtained by permuting columns of the regular motif. Model Misspecification.
- PART 3: Bias and relative efficiency comparisons in various scenarios.

PART 1: Finding Weak Regular Motifs

$N = 30$ sequences generated from an i.i.d. background model and an instance of the weak motif is inserted in varying percentage of the sequences $(L = 100)$. Let

- $K_i = \{$set of true motif sites in sample i$\}$
- $\hat{K}_i = \{$set of predicted motif sites in sample i$\}$
- $s\widehat{en}s = \Sigma_{i=1}^{50} \frac{|K_i \cap \hat{K}_i|}{|K_i|}$. $100 \times s\widehat{en}s$ is reported.



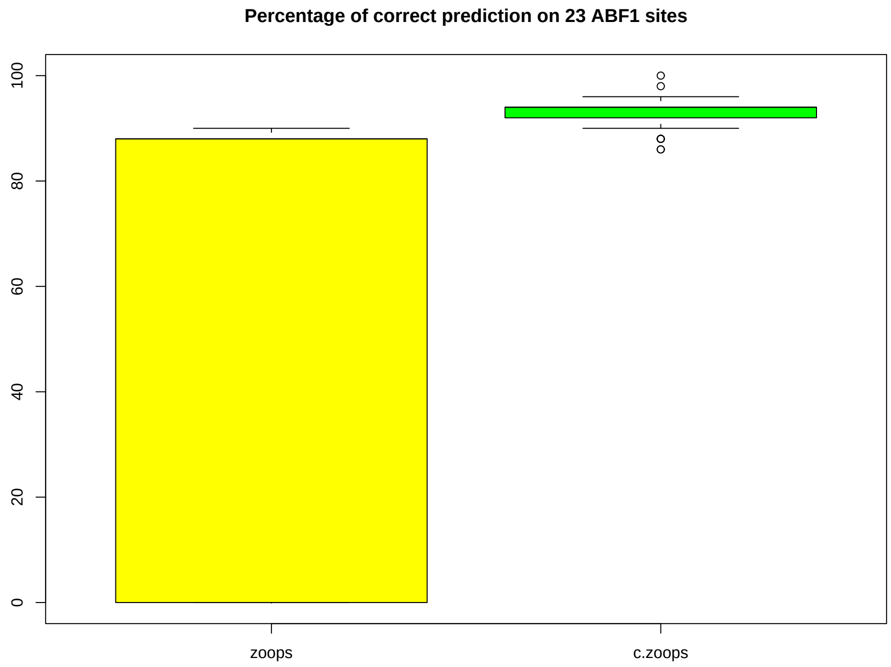
PART II: Performance in the Presence of Competing Irregular Motif

- $N = 23$ sequences are generated from an i.i.d. background model. All sequences have an irregular motif and a known ABF1 site $(L = 100)$.

Results

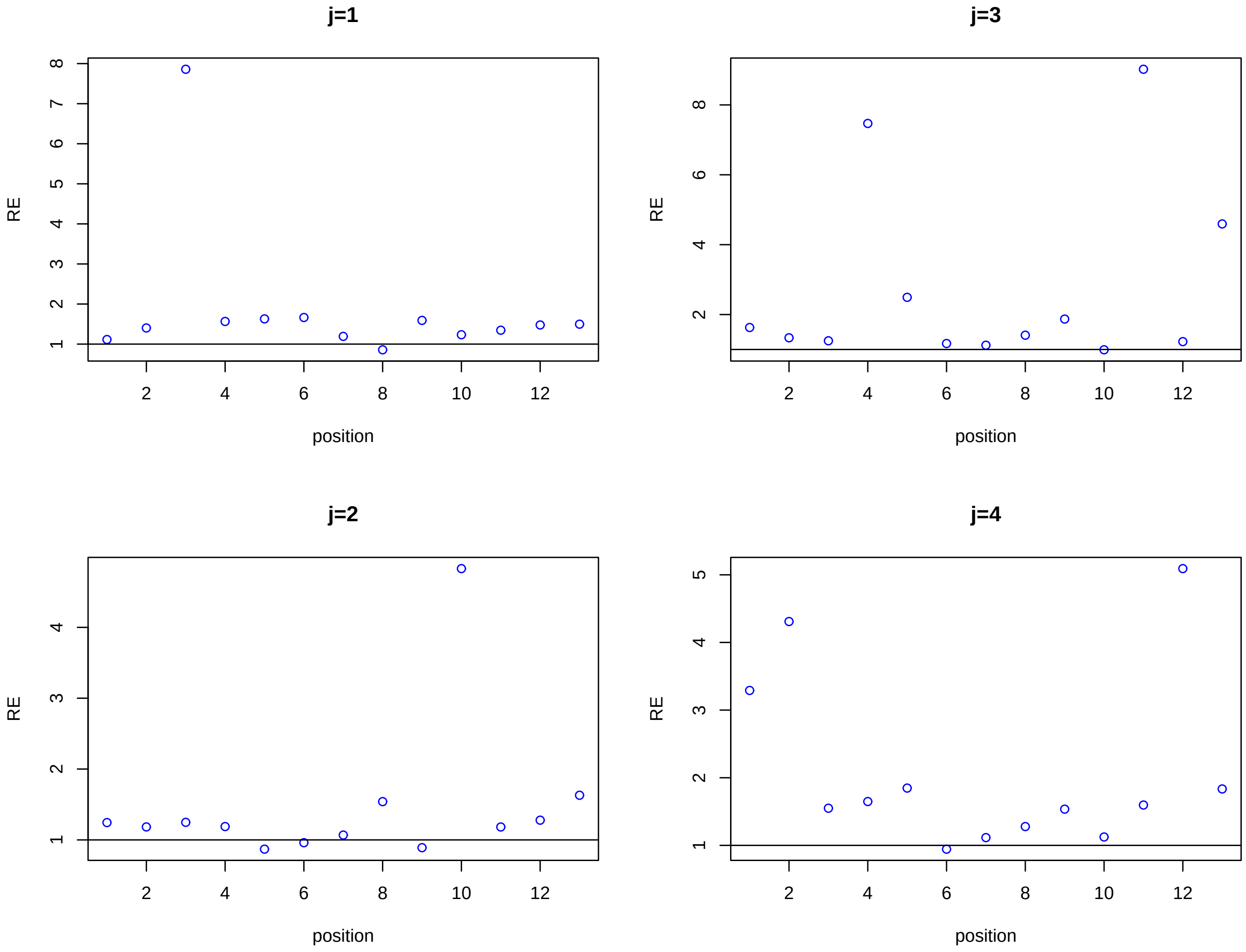
→ OOPS and ZOOPS models perform the same.

→ C.OOPS and C.ZOOPS converge to the regular motif almost all the time, OOPS and ZOOPS converge to the regular motif only 50% of the time.



PART III: Relative Efficiency Comparison

- $N = 50$ sequences are generated from an i.i.d. background model. An instance of the *regular motif* is inserted in a varying percentage of the sequences ($L = 100$).
- Estimated relative efficiencies based on 400 data sets are reported at each (w, j) of the motif matrix at $F = 50\%$.



Starting values for EM Algorithm

- Strategy I (I.k): Compute *initial likelihood* for *al* starting values constructed from length W oligos. Then, run EM till convergence for k of them with the highest initial likelihood.
- Strategy II (OS.k): Run EM one step for *all* starting values constructed from length W oligos. Then, run EM till convergence for k of them with the highest one-step likelihood.

Summary of performances of these two strategies on 10 data sets:

	I.1	I.5	I.10	I.20	I.50	I.100	OS.1	OS.5	OS.10	OS.20
# data sets	1	3	5	9	10	10	6	8	8	10
time required	15.16	22.57	31.24	52.42	119.79	231.28	151.33	157.59	165.45	181.33

datasets: # of data sets for which global maximum is found (out of 10!).

Conclusion

- Constraint entropy motif model
  - improves performance when the signal is weak.
  - improves performance when there is a competing irregular motif.
  - results in more efficient motif estimates.
  - results in more robust motif estimates.
- Extensions we are working on
  - Higher order Markov Chain for the background model.
  - Handling multiple occurrences of the same motif: Iterative Cutting Procedure.
  - Deletion of the high frequency irregular motif to improve accuracy of the regular motif estimate.

Acknowledgments

Eisen Lab

Katerina Kechris

Peter Bickel

Eric van Zwet