The Abdus Salam
International Centre
for Theoretical Physics
50th Anniversary 1964-2014

2584-2

Spring College on the Physics of Complex Systems

26 May - 20 June, 2014

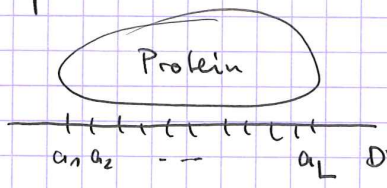
Modeling DNA Motifs

Martin Weigt
Université Pierre et Marie Curie
Paris
France

Modeling DNA Motifs

transcription factor binding sites
(TF)BS

Ingredients



⇒ sequence specific binding

⇒ BS ~ 10 nt

⇒ fuzzy motif → statistical modeling

⇒ additive ~~contrib.~~ binding energies for double mutants
→ statistically independent contribution of sites

⇒ **Simplest modeling**

Position-specific weight matrix (PWM)

$$P(a_1,\dots,a_L \mid w) = \prod_{i=1}^{L} w_i(a_i)$$

statistical indep. of sites

position specific

$= \frac{1}{Z}\exp\{\sum_i \mu_i(a_i)\}$

a \ i	1	2	3	4	..
A	0.7	0.1	0.01		
C	0.1	0.2	0.01		
G	0.06	0.1	0.3		
T	0.04	0.6	0.68		
Σ:	1				

consensus A T T

most likely sequence.

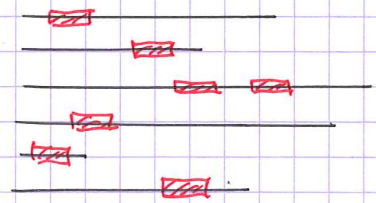
Two questions:

1) How to determine w from sequence data?
2) Given w for some TF, how can we find its BS?

For 1: How look data like?

→ collection of sequences which (possibly) contain TFBS

- promoter regions of potential targets of a TF
- ChIP experiments



Here: idealized situation: exactly TFBS seq.

$$\Rightarrow \quad D = \begin{pmatrix} a_1^1 & \cdots & a_i^1 & \cdots & a_L^1 \\ \vdots & & \vdots & & \vdots \\ a_1^m & \cdots & a_i^m & \cdots & a_L^m \\ \vdots & & \vdots & & \vdots \\ a_1^M & \cdots & a_i^M & \cdots & a_L^M \end{pmatrix}$$

$n_i(A) + n_i(C) + n_i(G) + n_i(T) = M$

M ... number of seqs.

L ... length of seqs.

a_i^m .. nt in pos. i of seq. m

$(\in \{A, C, G, T\})$

For any given w, we can determine the probability of our data D:

$$P(D \mid w) = \prod_{m=1}^{M} P(a_1^m, \dots, a_L^m \mid w)$$

↑ assume independence of sequences

$$= \prod_{m=1}^{M} \prod_{i=1}^{L} w_i(a_i^m)$$

$$= \prod_{i=1}^{L} \prod_{a \in \{ACGT\}} w_i(a)^{n_i(a)}$$

with

$n_i(a)$ = number of occurrences of nt a in column i of D.

Maximum-likelihood inference

→ choose w such that ~~$P(D)$~~ the data are as probable as possible!

$$\hat{w} = \underset{w}{\operatorname{argmax}}\, P(D \mid w)$$

→ can be solved analytically and gives (no surprise)

$$w_i(a) = \frac{w_i(a) + \mu}{M + 4\mu}$$

pseudo count μ

→ avoids $w_i(a) = 0$ only because a never seen in pos. i → small M.

But we need

$$P(\omega \mid D)$$

→ have data, but not the PWM

Bayes theorem:

$$P(\omega \mid D) = \frac{P(D \mid \omega) \cdot P(\omega)}{P(D)}$$

a priori distribution

$$P(\omega) = \prod_{i=1}^{L} P(\omega_i)$$

$$= \prod_{i=1}^{L} c_i \prod_{a \in \{A,C,G,T\}} \omega_i(a)^{\gamma - 1}$$

- $\gamma = 1$ ⇒ flat prior
- $\gamma > 1$ ⇒ favors $\omega_i = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$ → not realistic
- $\gamma < 1$ ⇒ favors $\omega_i \in \{ (1,0,0,0), (0,1,0,0), (0,0,1,0), (0,0,0,1) \}$

"consensus" - type ω_i

⇒ ~~$P(\omega \mid D) \propto \prod_{i=1}^{L} \prod$~~

Inference of ω from $P(\omega | D)$

Maximum likelihood (= max posterior prob)

$$\hat{\omega}_{ML} = \arg\max_{\omega} P(\omega | D)$$

$$\Rightarrow \hat{\omega}_i(a) = \frac{n_i(a) + \gamma - 1}{N + \alpha\gamma - 1}$$

But: $P(\omega | D)$ is prob. distr. fct.

e.g. $$\langle \omega_i(a) \rangle_P = \frac{n_i(a) + \gamma}{N + \alpha\gamma}$$

↑ difference for small data sets.

Question 2: How to decide if a NEW sequence is functional site?

⇒ compare to **null model**

- indepl. nucleotides
- no site specificity
- background nucleotide freqs.

$$P_0(\vec{s}) = \prod_{i=1}^{L} p_0(s_i)$$

⇒ **log odds ratio**

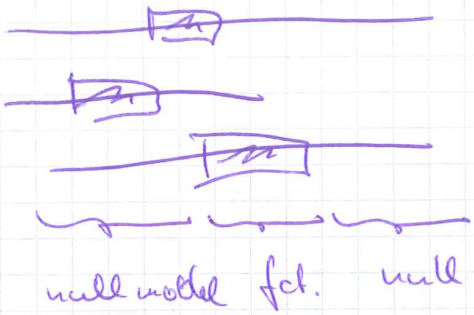
$$S(\vec{s}) = \log \frac{P_f(\vec{s}\,|w)}{P_0(\vec{s})}$$

$$= \sum_{i=1}^{L} \ln \frac{w_i(s_i)}{p_0(s_i)}$$

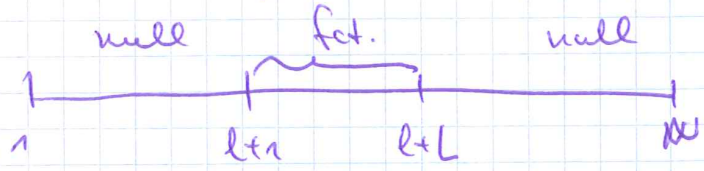
more likely under functional as null model?

remark: threshold zero unreliable because of much better sampling of background sites, many random sites look functional...

more general data situation:



$\vec{S} = (S_1, \dots, S_N)$ $\quad N > L$



$$\Rightarrow P(\vec{S} \mid \omega, \ell) = \overbrace{\prod_{i=1}^{\ell}}^{\ell} p_0(s_i) \overbrace{\prod_{j=1}^{L}}^{L} \omega_j(s_{j+\ell}) \cdot \overbrace{\prod_{i=\ell+L+1}^{N}}^{N} p_0(s_i)$$

$$= \overbrace{\prod_{i=1}^{N}}^{N} p_0(s_i) \overbrace{\prod_{j=1}^{L}}^{L} \frac{\omega_j(s_{j+\ell})}{p_0(s_{j+\ell})}$$

Data set: $D = \{ \vec{S}^\mu \mid \mu = 1 .. M \}$

$\hookrightarrow \vec{S}^\mu = (s_1^\mu, \dots, s_{N_\mu}^\mu)$

$$\Rightarrow P(D \mid \omega, (\ell^\mu)_{\mu=1,\dots,M}) = \overbrace{\prod_{\mu=1}^{M}}^{M} P(\vec{S}^\mu \mid \omega, \ell^\mu)$$

independence of different seqs.

⇒ estimate w and (ℓ^μ) !

two strategies

1) $P(D|w) = \sum_{(\ell^\mu)} P(D|w,(\ell^\mu)) P((\ell^\mu))$

+ max w.r.t. w (MEME)

2) $P(D|(\ell^\mu)) = \int dw\, P(D|w,(\ell^\mu))\, P(w)$

↑ Dirichlet

+ MC sampling of (ℓ^μ)

Strategy 1 : max. of $P(D|w)$ by expectation-maximization (EM)

for given (ℓ^μ)

$$\Rightarrow w_i(a) = \frac{1}{M} \sum_{\mu=1}^{M} \delta_{a, S^\mu_{i+\ell^\mu}}$$

→ averaging over (ℓ^μ)

$$w_i(a) = \frac{1}{M} \sum_{\mu=1}^{M} \left\langle \delta_{a, S^\mu_{i+\ell^\mu}} \right\rangle_{\ell^\mu}$$

$$= \frac{1}{M} \sum_{\mu=1}^{M} \frac{\sum_{\ell^\mu} \delta_{a, S_{i+\ell^\mu}} P(\vec{S}^\mu | w, \ell^\mu)}{\sum_{\ell^\mu} P(\vec{S}^\mu | w, \ell^\mu)} \quad (*)$$

Problem: w on both sides !

Algo:
- initialize ω randomly
- Plug in rhs of (*), update ω
- iterate until ω remains stationary

→ local max of $P(D|\omega)$
→ reinitialize for chosing good local max (if not global)

Strategy 2
- calculate $P(D|(\ell^M)) = \int d\omega\, P(D|\omega,(\ell^M))P(\omega)$ explicitly
- sample over (ℓ^M)

$$P(D|\omega,|\ell^M)) = \prod_{a\in\{A,C,G,T\}} p_0(a)^{m(a)} \times \prod_{i=1}^{L} \prod_{b\in\{A,C,G,T\}} \left(\frac{\omega_i(b)}{\cancel{p_0(b)}}\right)^{n_i(b)}$$

(backgr. / fct.)

Attention: $m(a), n_i(b)$ depends on (ℓ^M) !

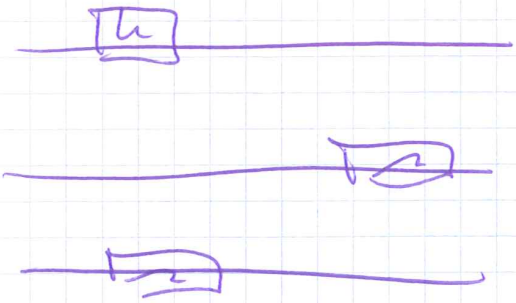
With Dirichlet prior $P(\omega) \sim \prod_{i,a} \omega_i(a)^{\gamma-1}$

$$P(D|(\ell^M)) = \prod_a p_0(a)^{m(a)} * \prod_{i=1}^{L} \left[\frac{\Gamma(4\gamma)}{\Gamma(M+4\gamma)} \prod_b \frac{\Gamma(n_i(a)+\gamma)}{\Gamma(\gamma)}\right]$$

⇒ complicated (ℓ^M)-dependence

⇒ Gibbs Sampling

- initialize $(\ell^{(r)})$

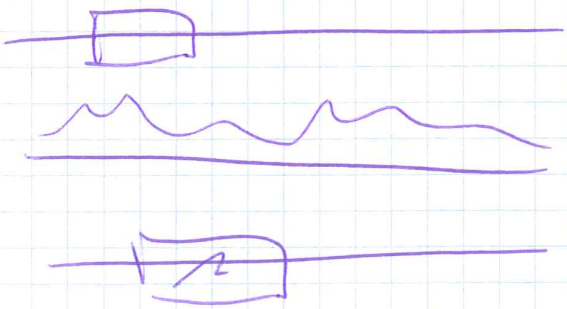


rand.

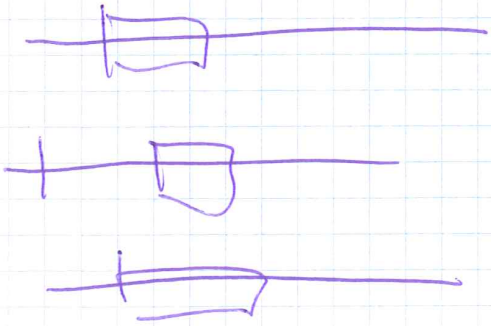
- select seq. $\nu \in \{1, \dots, n\}$
- calculate

$$P(\ell^{\nu} = \ell \mid D, (\ell^{-\nu}))$$

for each $\ell = 0, \dots, N_\nu - L$



- select new value $\ell^{\nu} = \ell$ with this proba.



+ iteration.

Most probable positions
→ simulated annealing

$$P(D \mid (\ell^{\nu}))^{\beta}$$

$\beta = 1 \to +\infty$