

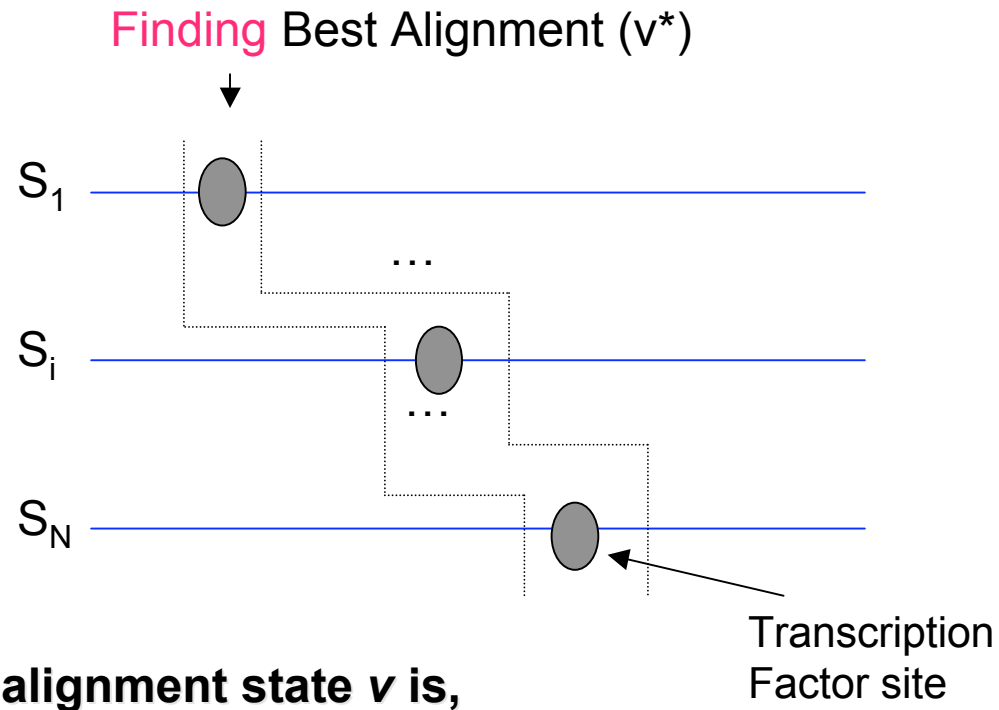
Rocky'06: #26

# ***Monte Carlo EM Algorithm for Sequence Motif-finding***

December 2, 2006

Chengpeng "Charlie" Bi, Ph.D.  
**Children's Mercy Hospitals & Clinics**

# Multiple Local Alignment – Motif-finding



Probability of alignment state  $v$  is,

$$p^{(v)} = e^{-E[\mathcal{S}, \mathcal{A}^{(v)}] / k_B T} / Z$$

Partition function:

$$Z = \int \dots \int_{\mathbf{v}} e^{-E[\mathcal{S}, \mathcal{A}^{(v)}] / k_B T} d\mathbf{v} = \sum_{a_1} \dots \sum_{a_i} \dots \sum_{a_N} e^{-E[\mathcal{S}, \mathcal{A}^{(v)}] / k_B T}$$

# EM Algorithms and its Limitations

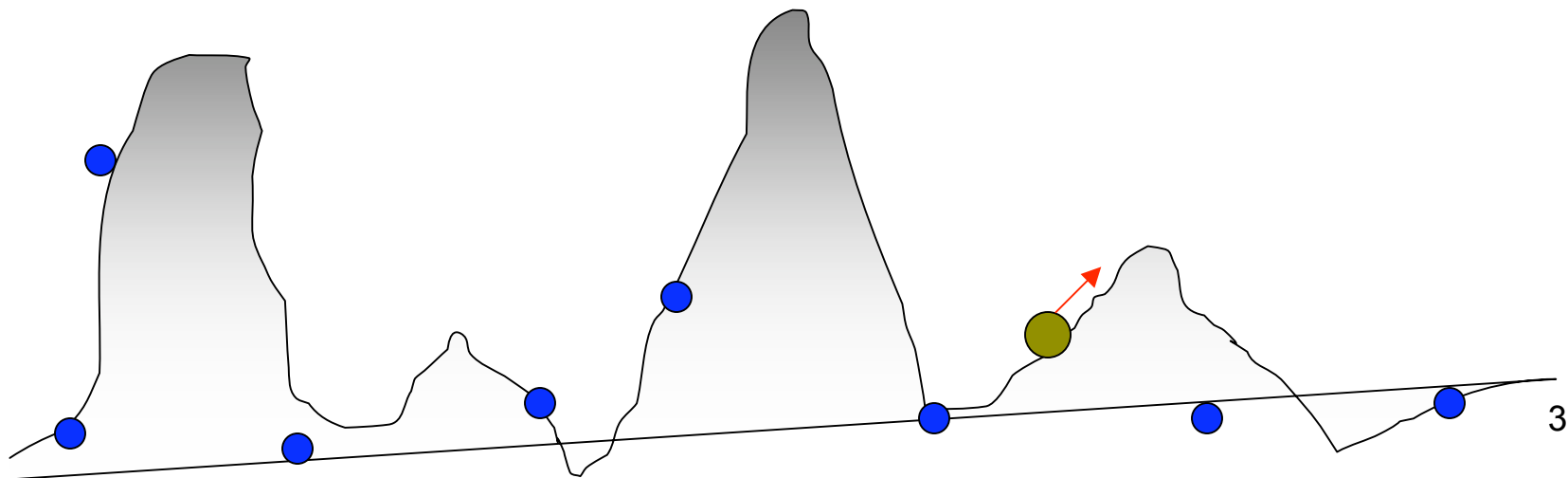


**E-step:**

$$p(a_i = l | S_i, \Theta^{(t)}) = \frac{p(S_i | a_i = l, \Theta^{(t)})}{\sum_{j=1}^{L_i - w + 1} p(S_i | a_i = j, \Theta^{(t)})}$$

**M-step:**

$$\theta_{jk}^{(t+1)} = \frac{N_{jk}}{\sum_{k=1}^K N_{jk}}$$



# Monte Carlo EM Algorithm

**Initialize:**  $A^{(0)}$  and Evaluate  $\Theta^{(0)}, Q^{(0)}$

$t \leftarrow 0, Q_{\max} \leftarrow Q^{(0)}$

**repeat:**  $t \leftarrow t + 1$

**for**  $i \leftarrow 1$  to  $N$    // **Simulation**

    compute conditional prob. distribution ( $p$ ) and its cdf ( $U$ )

    draw sample:  $\mathbf{a}_i \sim U(\mathbf{a}_i | \mathcal{S}_i, \Theta^{(t-1)})$

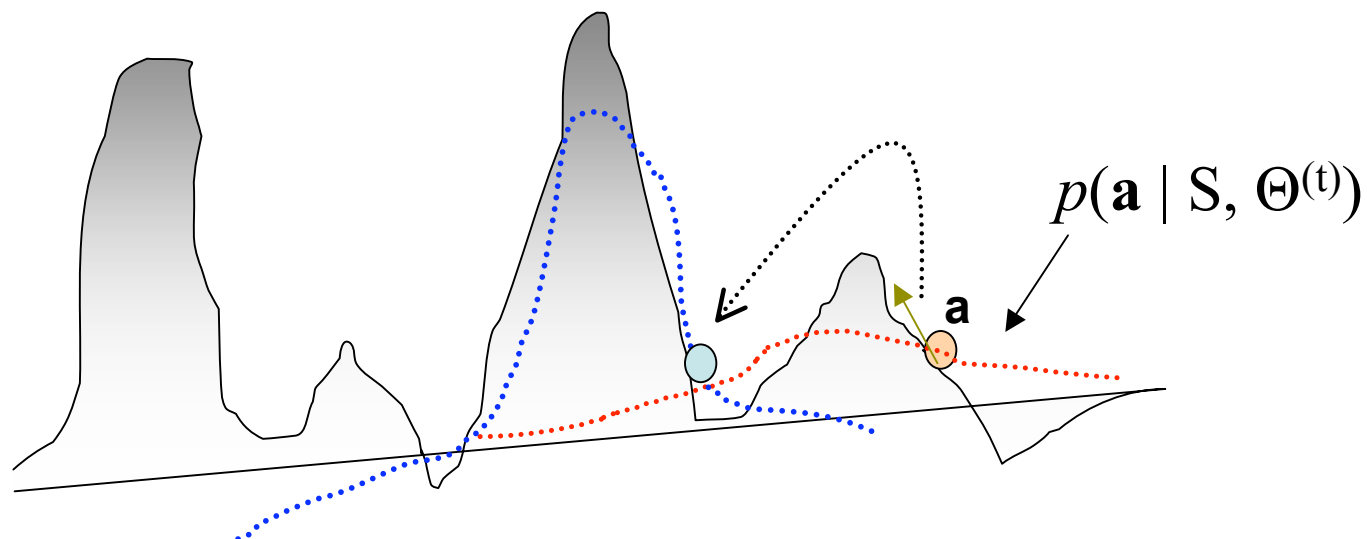
$A^{(t)} = [\mathbf{a}_1, \dots, \mathbf{a}_i, \dots, \mathbf{a}_N]^T$

**Update**  $\Theta^{(t)}, Q^{(t)}$

**if**  $Q^{(t)} > Q_{\max}$

$Q_{\max} \leftarrow Q^{(t)}, A^* \leftarrow A^{(t)}, \Theta^* \leftarrow \Theta^{(t)}$

**output:** optimal alignment ( $A^*$ ) and motif model ( $\Theta^*$ )



# Algorithm Comparison

PROKARYOTE

TF	Algorithms	$w$	$ A $	Precision $\uparrow$	Recall $\dagger$	F-score $\ddagger$
CRP	MCEMDA	22	18	18/18	18/23	0.878
	BioPros	22	9	9/9	9/23	0.563
	MEME	24	13	12/13	12/23	0.667
FNR	MCEMDA	14	67	64/67	64/67	0.955
	BioPros	14	67	39/67	39/67	0.575
	MEME	20	43	43/43	43/67	0.782
LexA	MCEMDA	20	24	24/24	24/24	1.000
	BioPros	20	24	15/24	15/24	0.625
	MEME	21	24	23/24	23/24	0.958
ERE	MCEMDA	15	25	22/25	22/25	0.880
	BioPros	13	16	14/16	14/25	0.683
	MEME	15	17	15/17	15/25	0.714
E2F	MCEMDA	13	25	22/25	22/27	0.846
	BioPros	11	21	11/21	11/27	0.564
	MEME	13	23	19/23	19/27	0.760

EUKARYOTE

# Summary

	<b>EM</b>	<b>MCEM</b>
Dependent on initializing	yes	no
Local optimum	yes	Not always
Convergence	Quick (most of times)	Slow (specified)
Iteration	<b>E</b> -step and <b>M</b> -step	<b>S</b> -step and <b>U</b> -step