Inferring Probabilistic Models of cis-Regulatory Modules

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2015

Colin Dewey

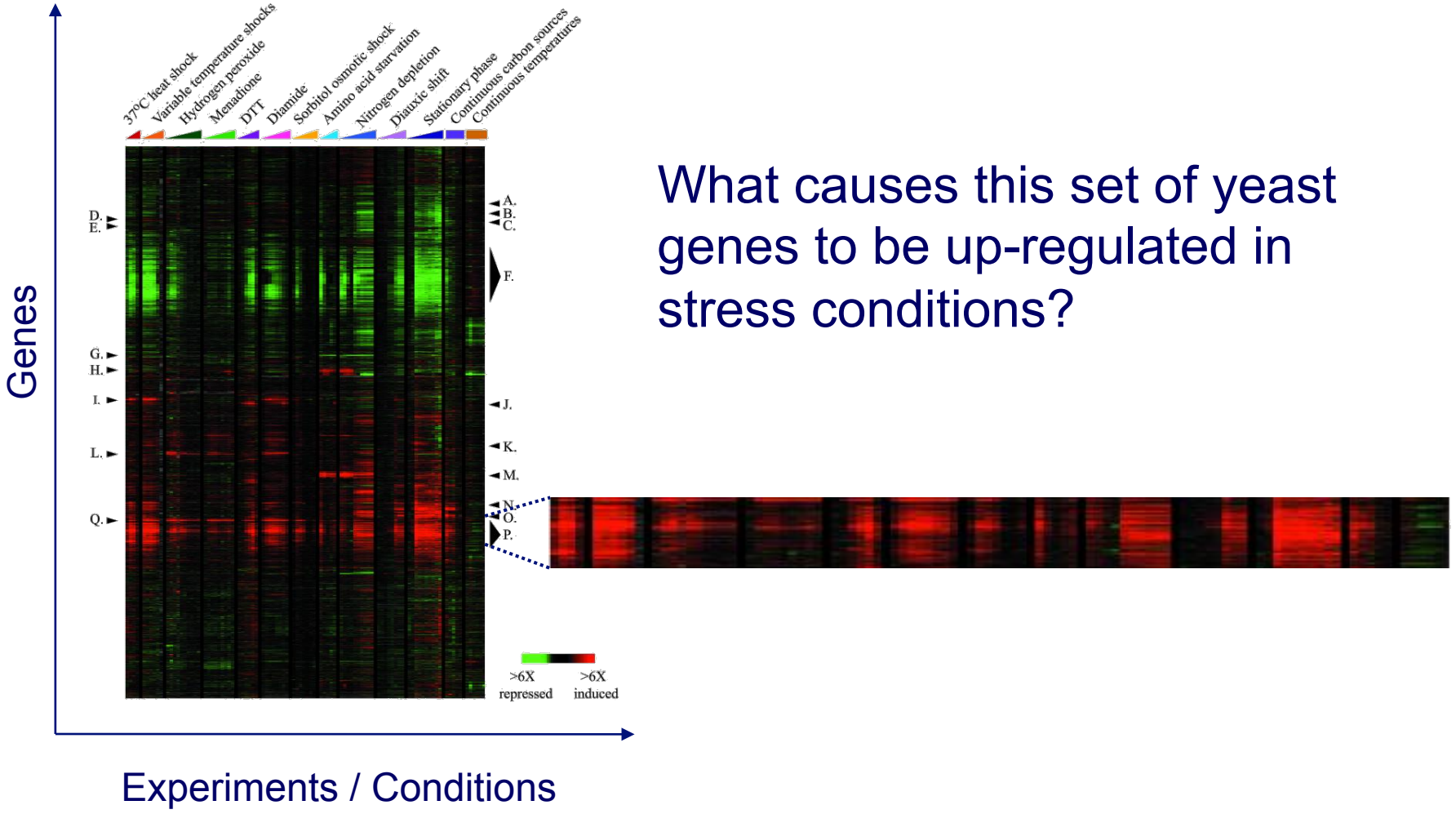
cdewey@biostat.wisc.edu

Goals for Lecture

the key concepts to understand are the following

- cis-regulatory modules
- the CRM finding task
- the structure search problem
- beam search
- sensitivity, specificity, recall, precision
- duration modeling
- semi-Markov models
- time complexity of DP methods for semi-Markov models

A Common Type of Question

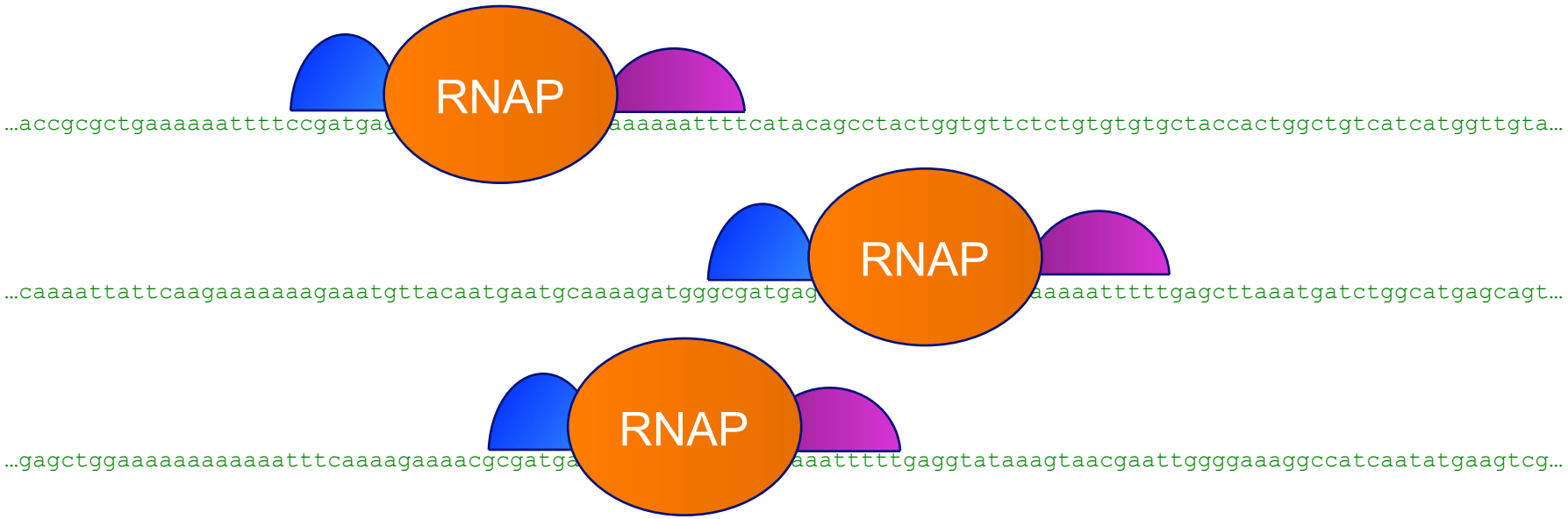


What causes this set of yeast genes to be up-regulated in stress conditions?

Figure from Gasch *et al*., *Mol. Biol. Cell*, 2000

cis-Regulatory Modules (CRMs)

- co-expressed genes are often controlled by specific configurations of binding sites

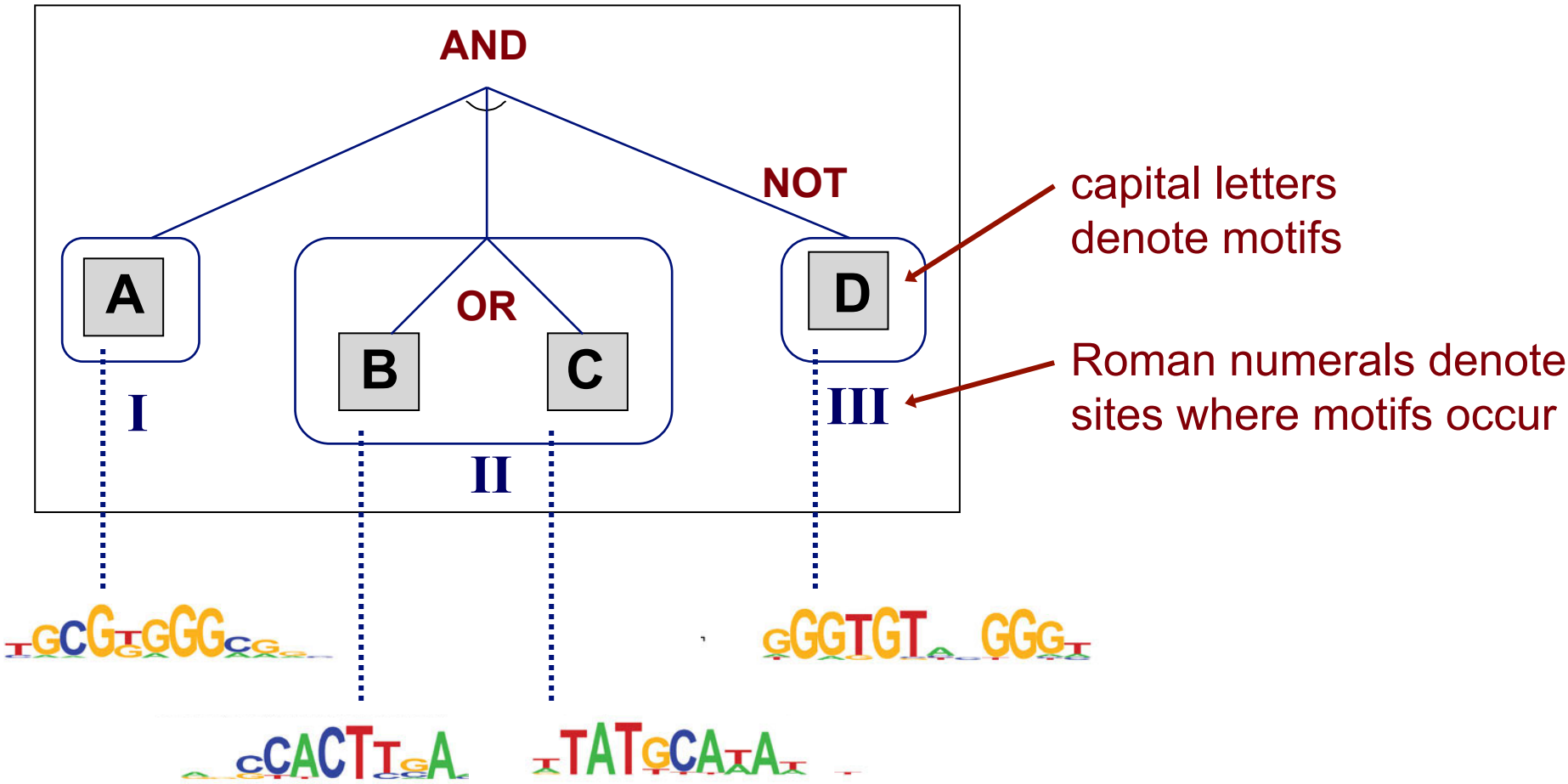


CRM Learning Task

- **Given:** sequences believed to share a CRM, and "negative" sequences believed not to contain the CRM
- **Learn:** models of the binding-site motifs and the spatial and logical relationships that characterize the CRM

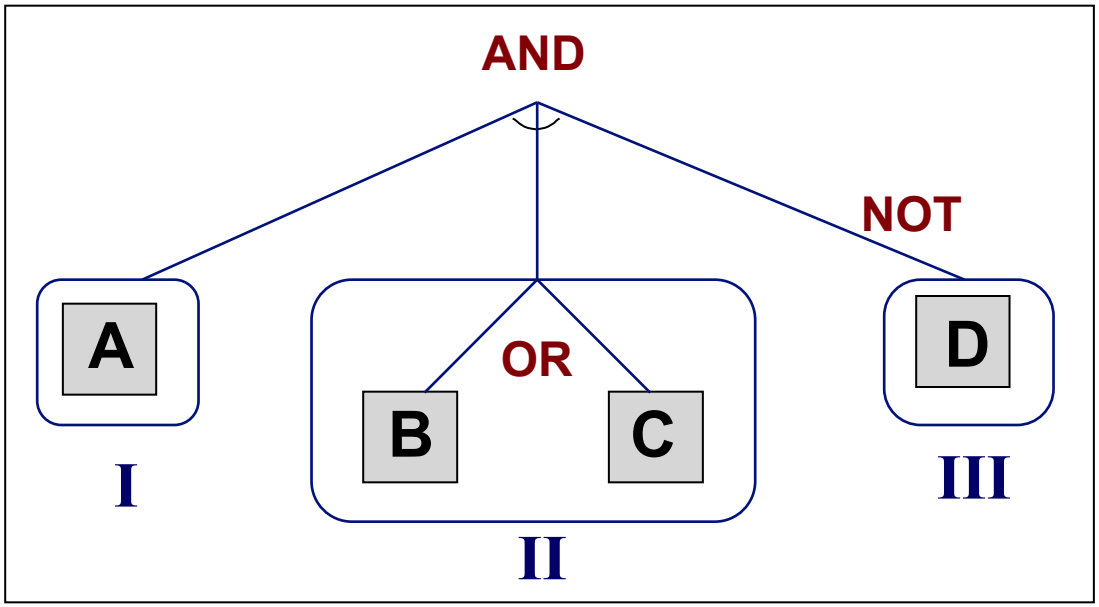
CRM Representation

- to characterize CRMS we want to represent logical relationships among motifs



CRM Representation

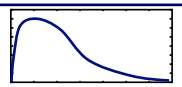
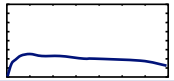
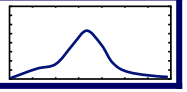
logical aspects: AND, OR, and NOT



- we also want to represent spatial relationships that characterize binding sites

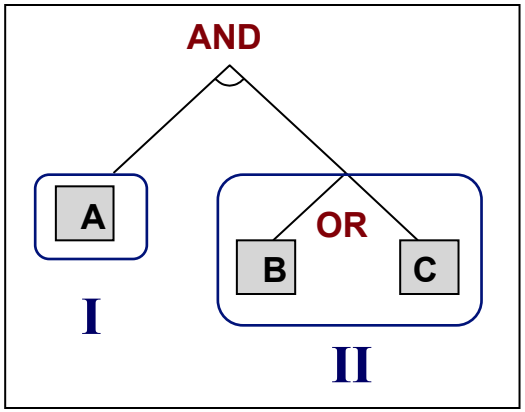
spatial aspects: order, distance, and strand

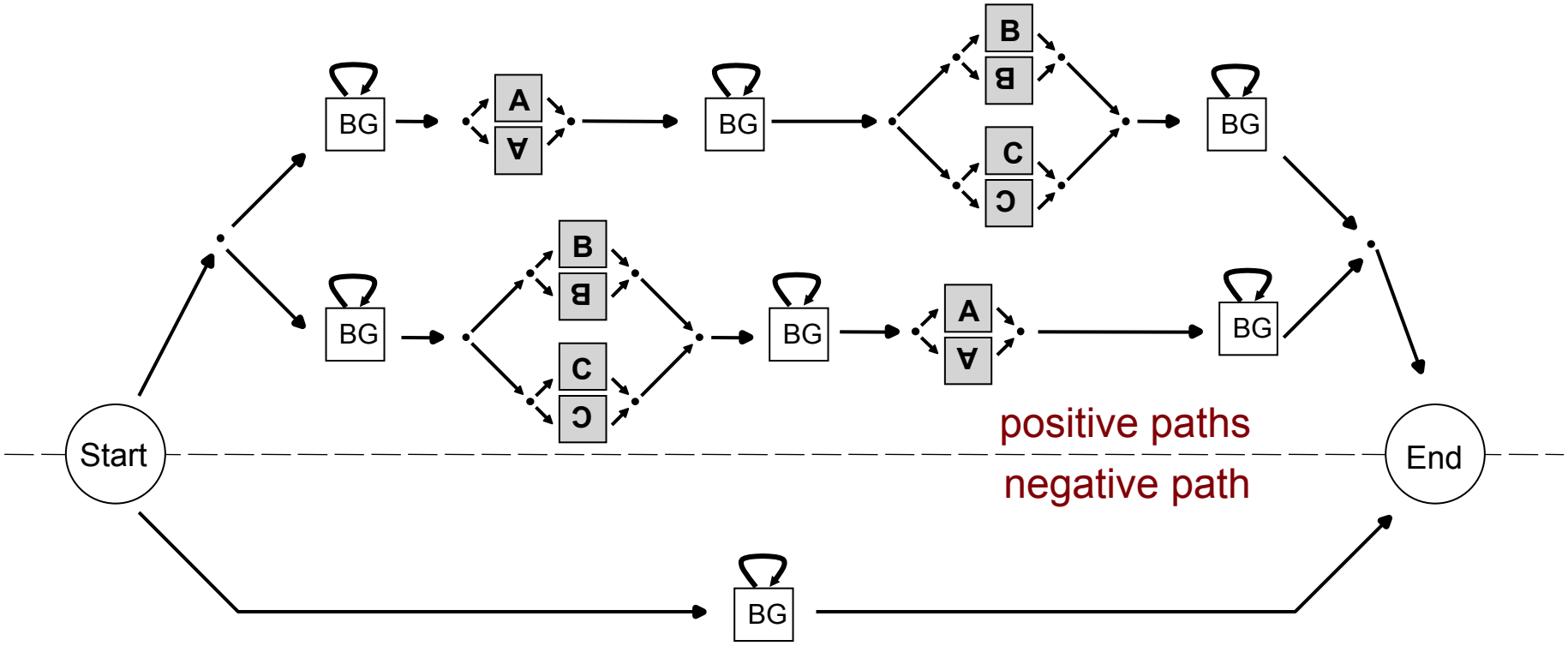
Order	I	II
II	0.5	
III	0.2	0.7

Distance	I	II
II		
III		

Strand	P(template)
I	0.8
II	0.5
III	0.3

The CRM representation viewed as an HMM



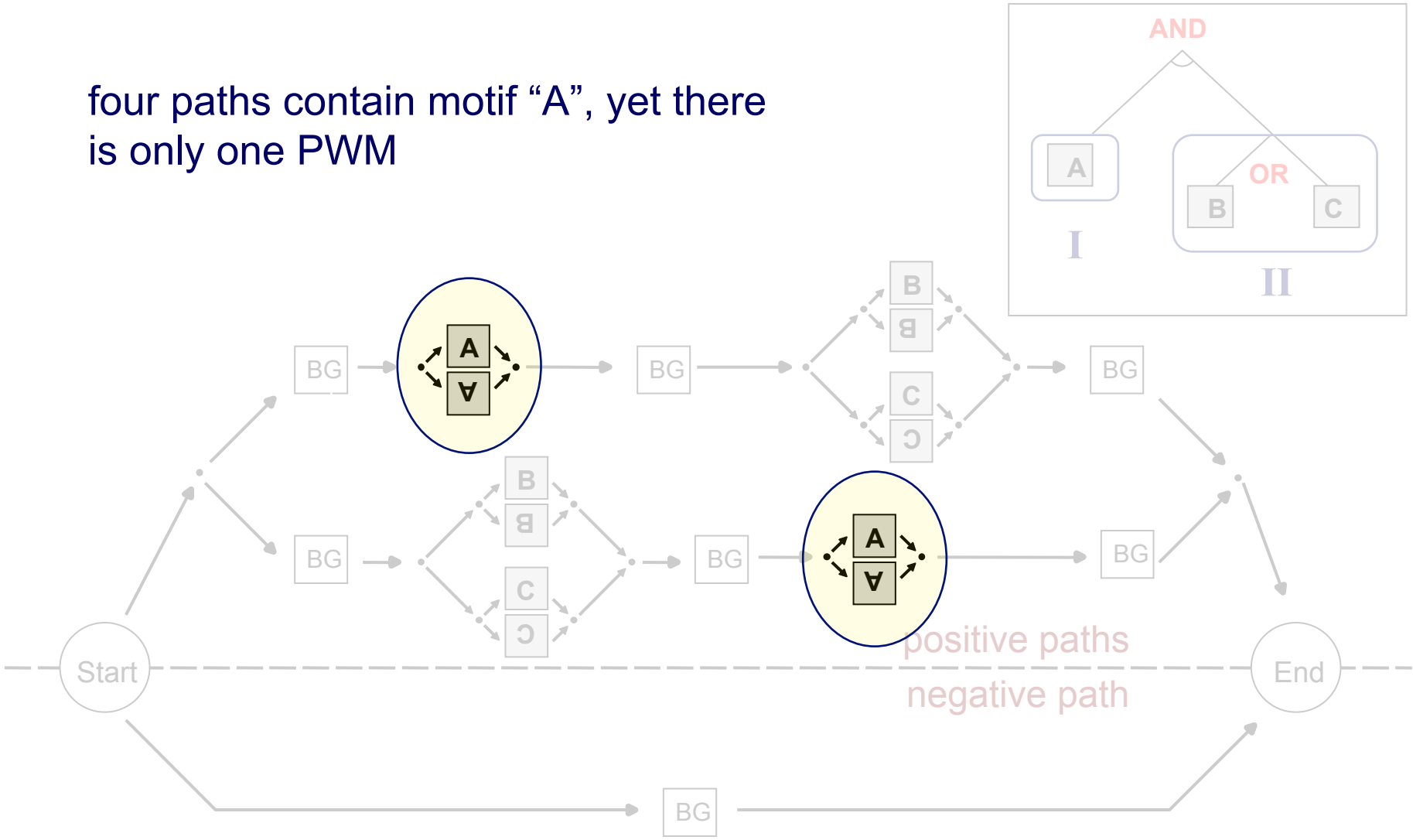


The CRM Representation viewed as an HMM

- each motif block is a sequence of states representing the motif
- the BG states represent “background” sequence
- the negative path(s) accounts for sequences that do not contain the CRM

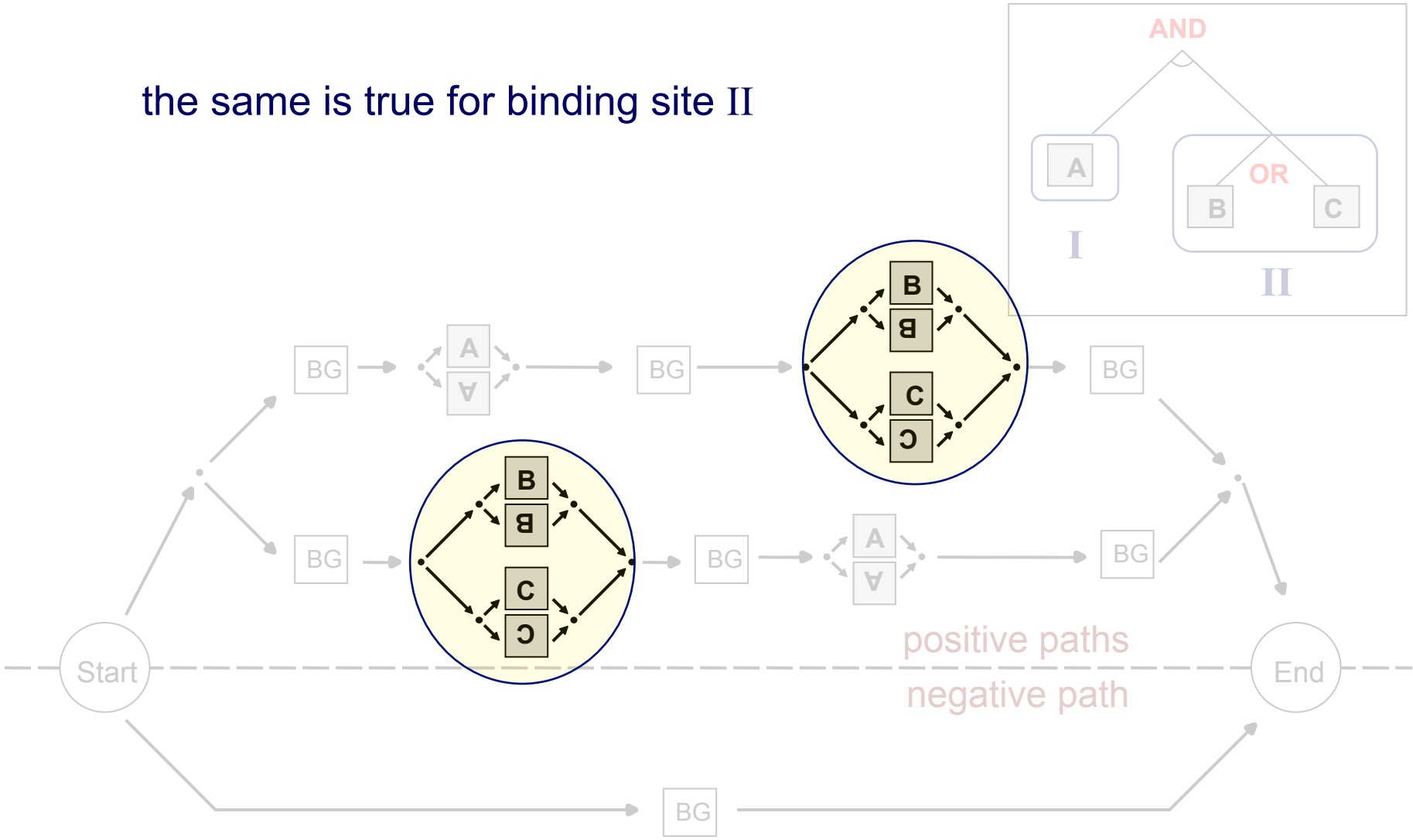
Parameter Tying

four paths contain motif “A”, yet there is only one PWM



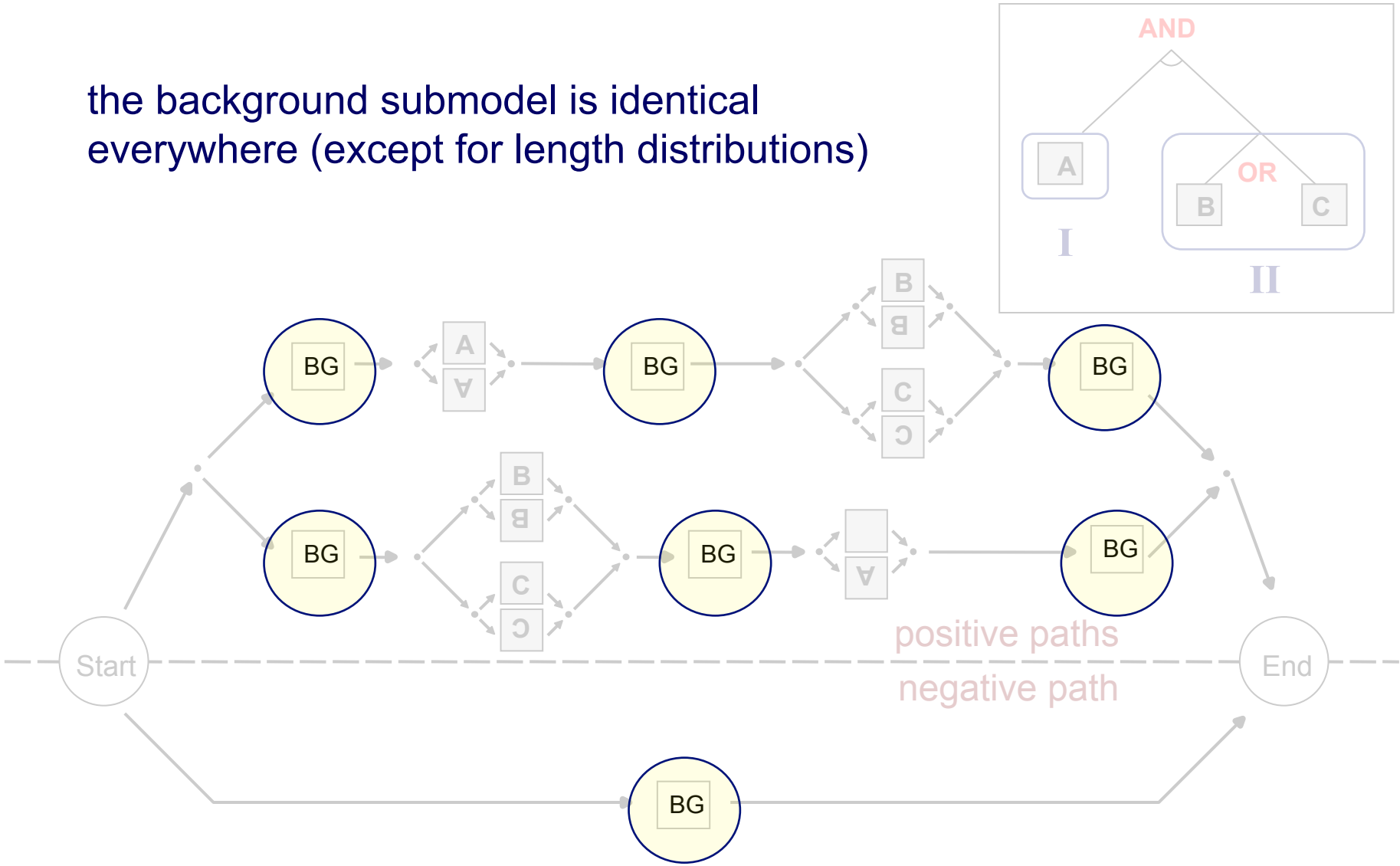
Parameter Tying

the same is true for binding site II

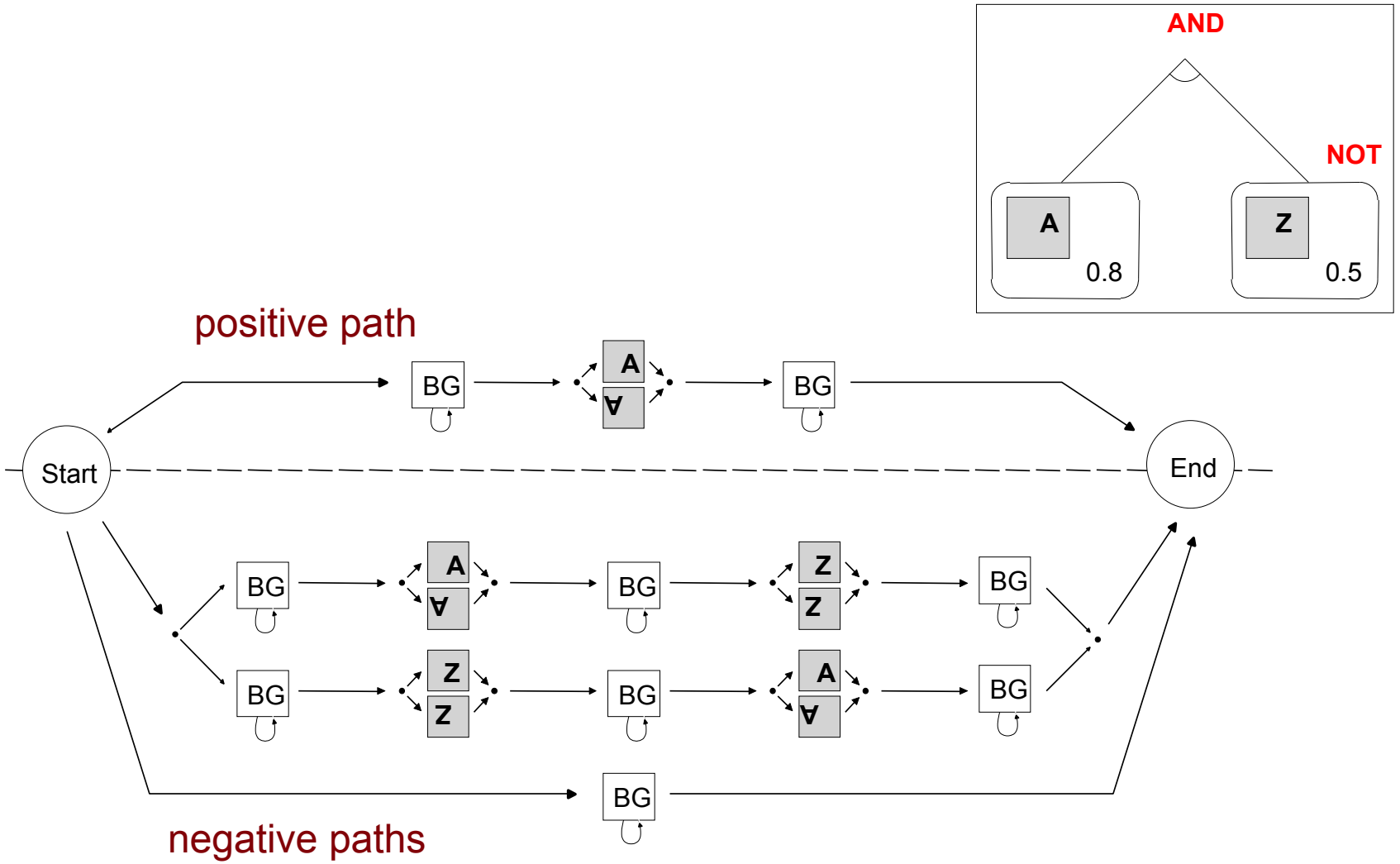


Parameter Tying

the background submodel is identical everywhere (except for length distributions)

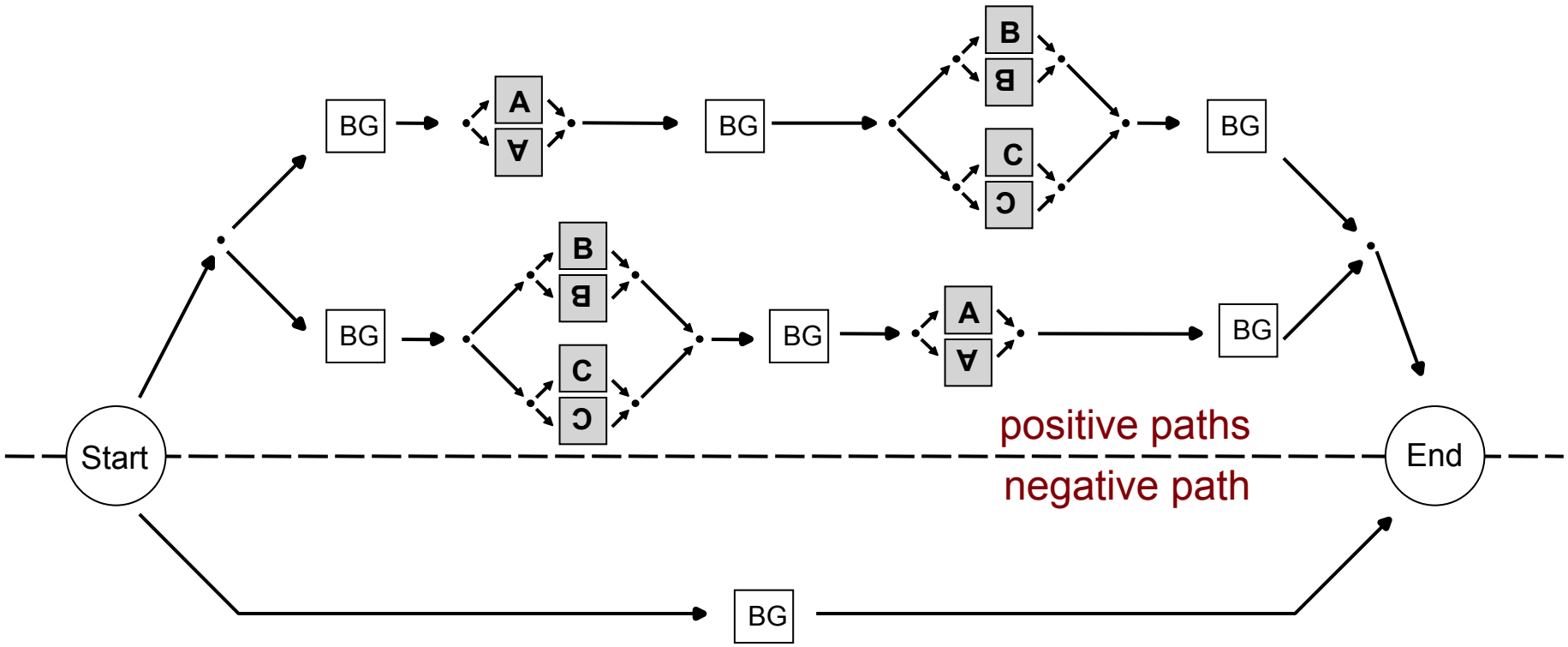


HMM with a *Negated* Binding Site



The Classification Task

- given a test sequence, we can calculate the most probable path (Viterbi algorithm) or the summed probability for all positive paths (Forward algorithm)



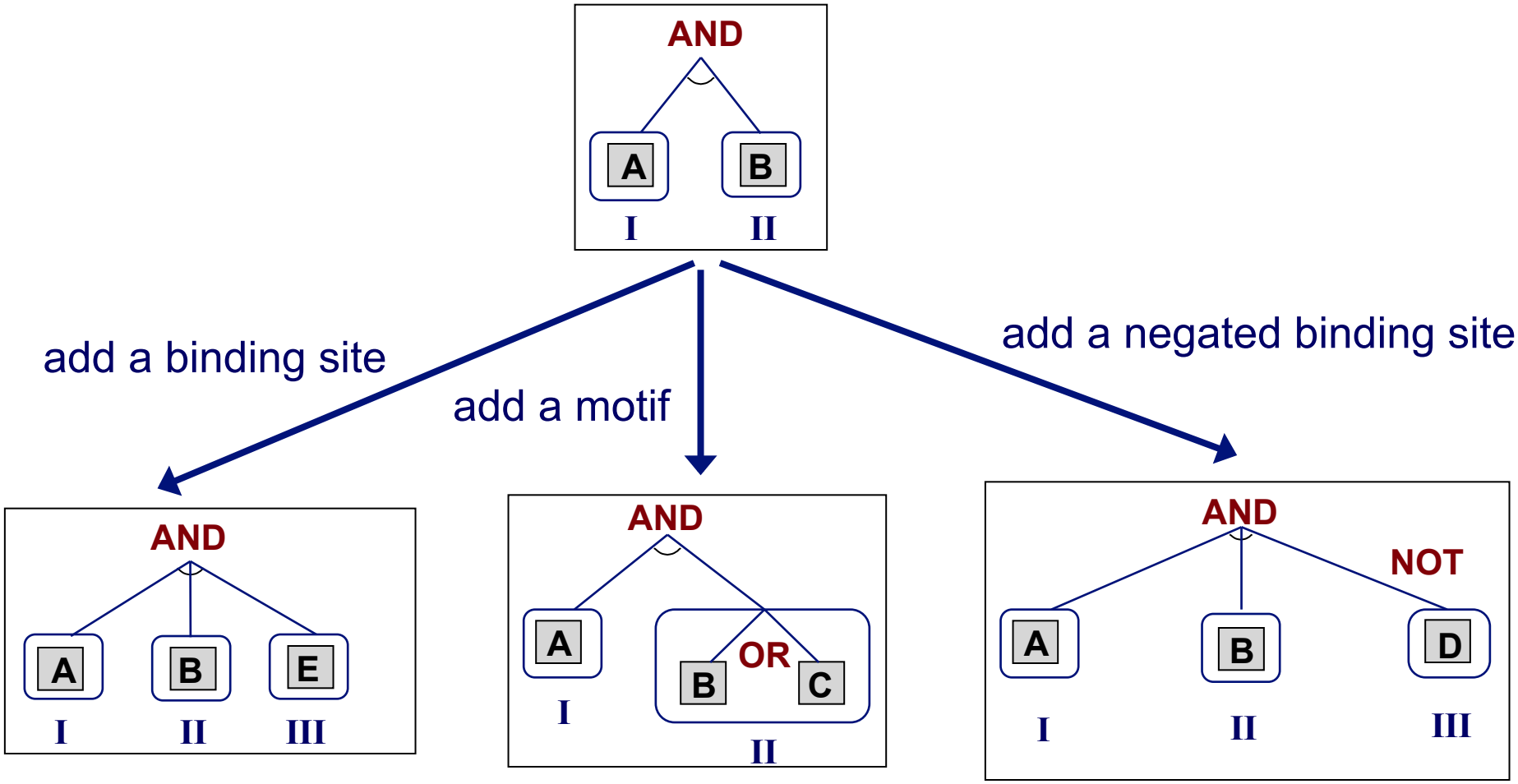
The Learning Tasks

1. *parameter learning* : estimating the probability parameters of the HMM
   - use Baum-Welch (EM) since we don't know the correct parse for training sequences

2. *structure learning*: determining the topology of the HMM
   - we'll view this is as a heuristic search problem

Learning The HMM Structure

- the structure search is carried out on the *compact* representation
- three operators



Beam Search for Structure Learning

given: initial structure I, search operators
$beam \leftarrow \{ I \}$
$best\_model \leftarrow I$
repeat
    $H \leftarrow$ pop highest scoring structure in $beam$
    foreach possible application of operator O
        $S \leftarrow O$ applied to H
        if score(S) > score($best\_model$)
            $best\_model \leftarrow S$
        push S onto $beam$
    limit size of $beam$ to k best scoring structures
until stopping criteria met
return: $best\_model$

k = “beam width”

Beam Search in Structure Learning

"scoring" a structure entails

1. retraining parameters of HMM using something like Baum-Welch (unsupervised learning)
   - thus learning a model for the new motif
   - for background states, only length parameters are retrained

2. estimating accuracy of model using a held-aside tuning set (supervised learning)

Empirical Evaluation of CRM Learning Method

- 25 yeast data sets
  - each with 15 – 214 promoter sequences
  - 100 negative promoter examples in each set
- each has evidence of two binding proteins
  (Lee *et al.*, *Science* 2002)

- cross-validation methodology

Accuracy Metrics

predicted class \ actual class	positive	negative
positive	true positives (TP)	false positives (FP)
negative	false negatives (FN)	true negatives (TN)

$$\text{sensitivity (recall)} = \frac{TP}{\text{actual pos}} = \frac{TP}{TP + FN}$$

(true positive rate)

$$\text{specificity} = \frac{TN}{\text{actual neg}} = \frac{TN}{TN + FP}$$

(1 - false positive rate)

$$\text{precision} = \frac{TP}{\text{predicted pos}} = \frac{TP}{TP + FP}$$

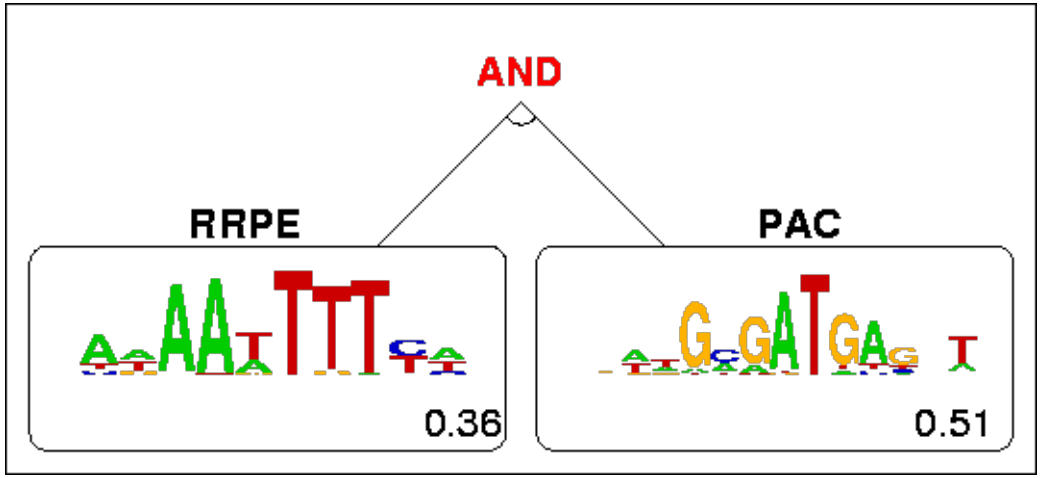
Empirical Evaluation of CRM Learning Method

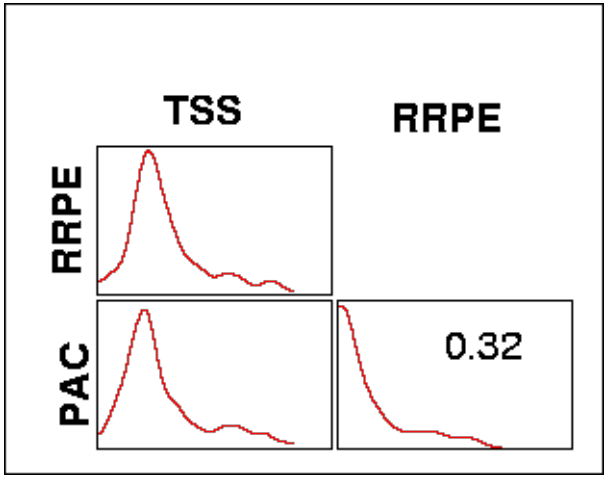
- assess models by calculating how likely they would have test-set accuracy as good by chance

- Segal and Sharan's method (*RECOMB* 2004) induced models with (statistically significant) accuracy better than chance on 12 out of 25 data sets
- Noto and Craven's method learns models that are better than chance on 21 out of 25

Analyzing Yeast Stress Data Sets

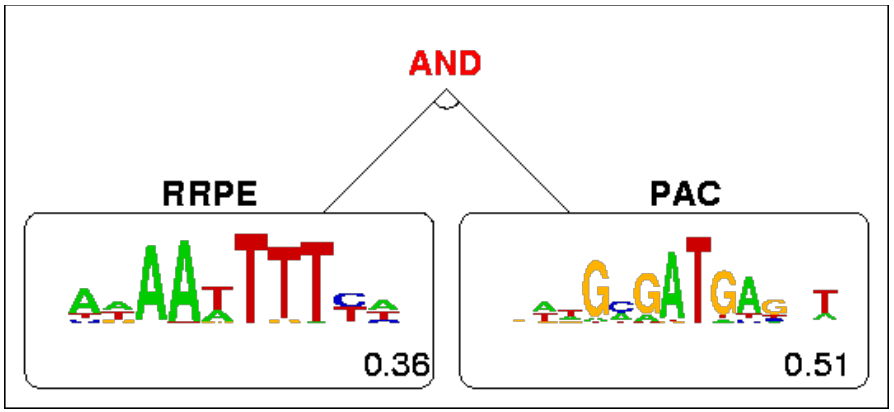
- three data sets, each associated with environmental stress response (ESR)
- Noto and Craven method attains statistically significant accuracy on all three
- one data set includes known promoter elements, which are recovered

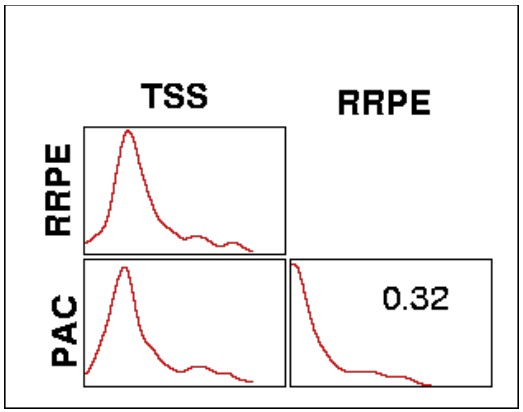


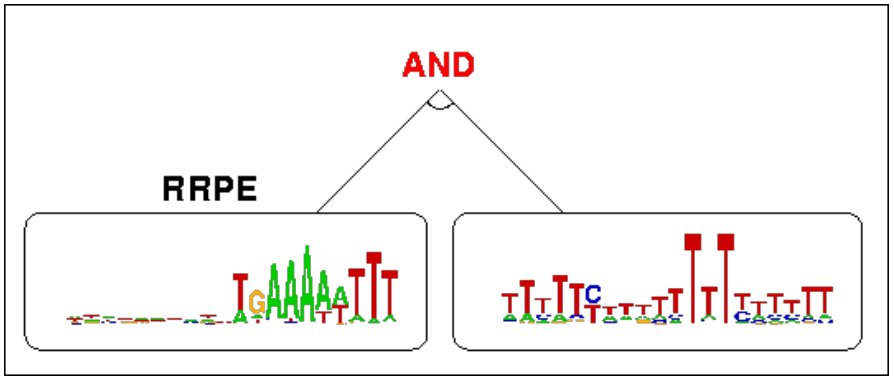


Does the Rich Representation Provide Value?

- consider a variant of the learner that cannot represent logical and spatial aspects (“bag of motifs” representation)



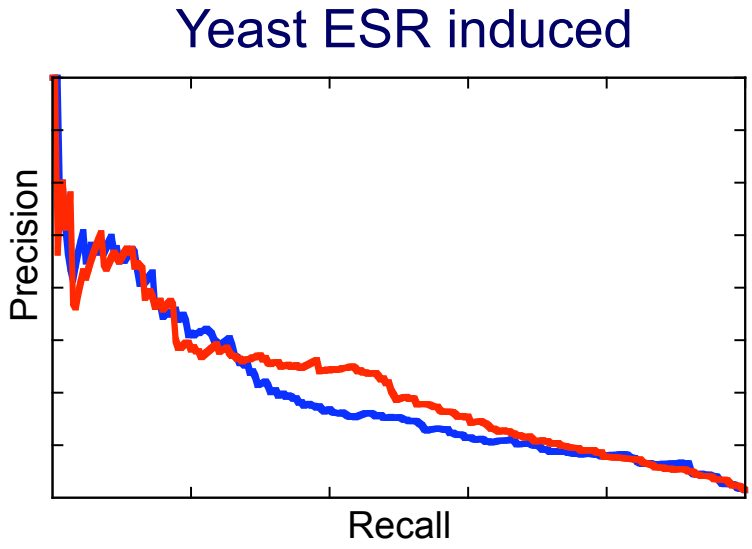


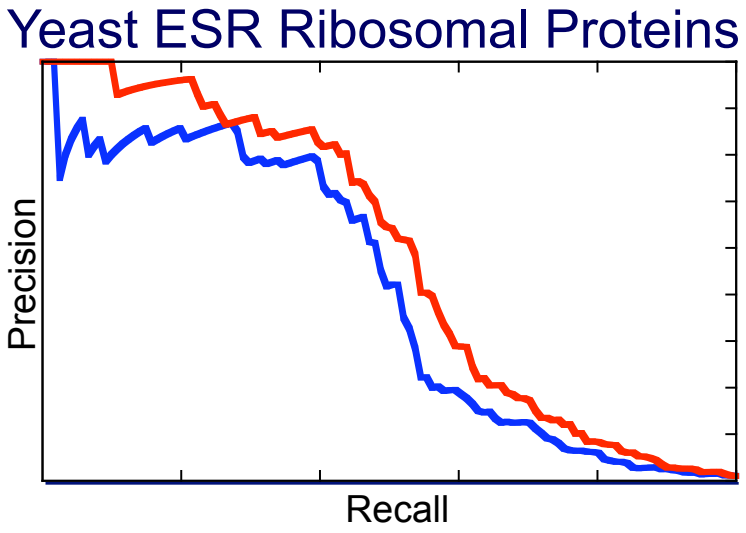


bag of motifs

- PAC element not recovered
- predictive accuracy worse

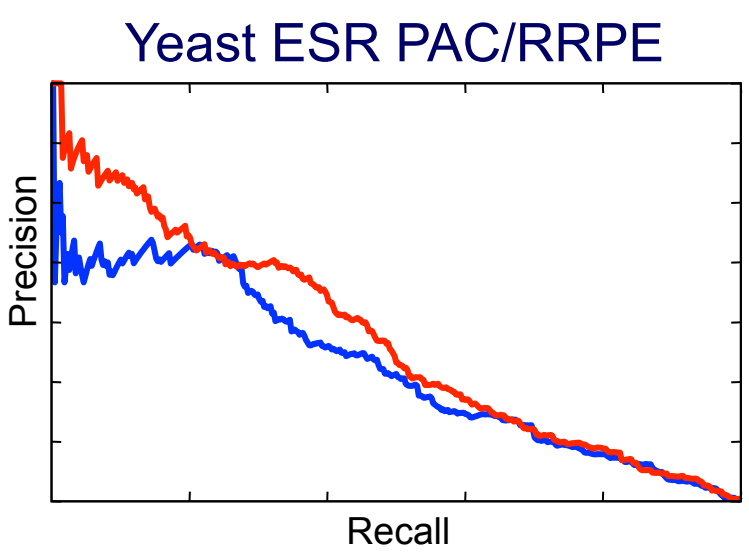
Does the Rich Representation Provide Value?

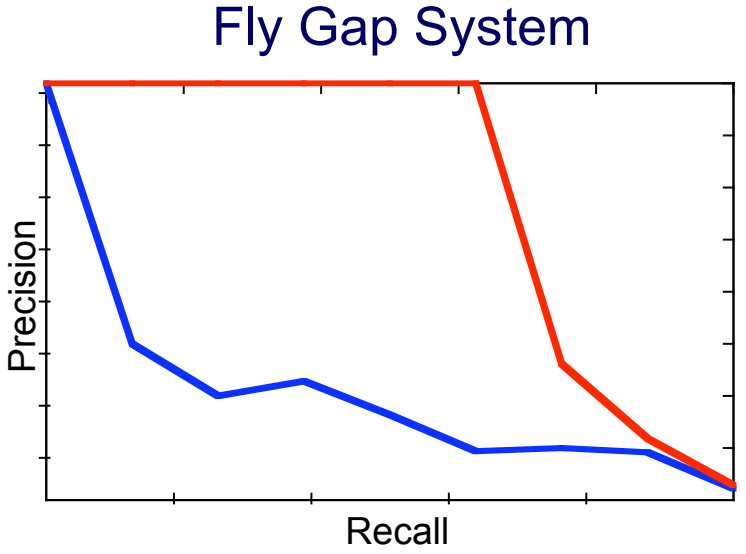




with logical and spatial aspects

bag of motifs

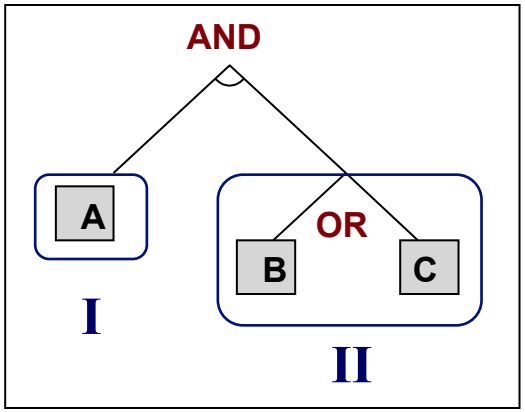


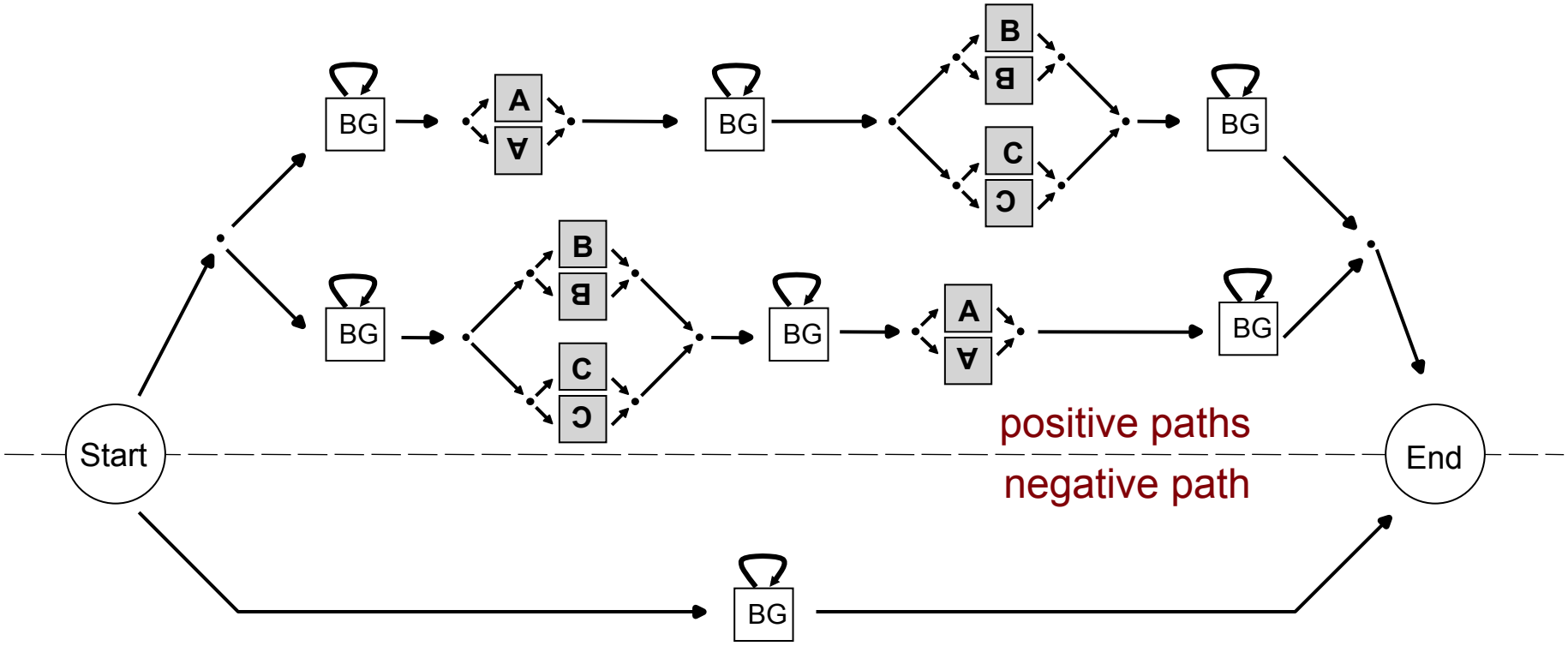


$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

The CRM representation viewed as an HMM



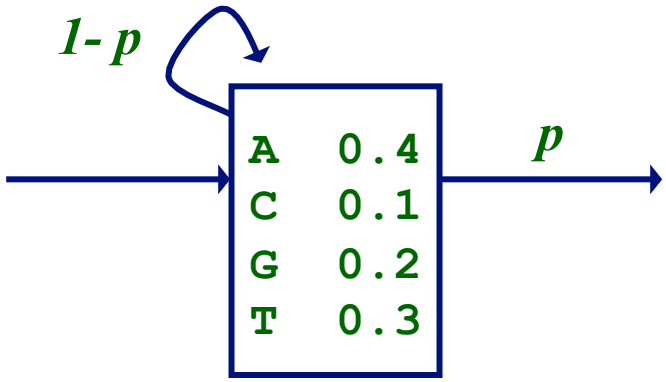


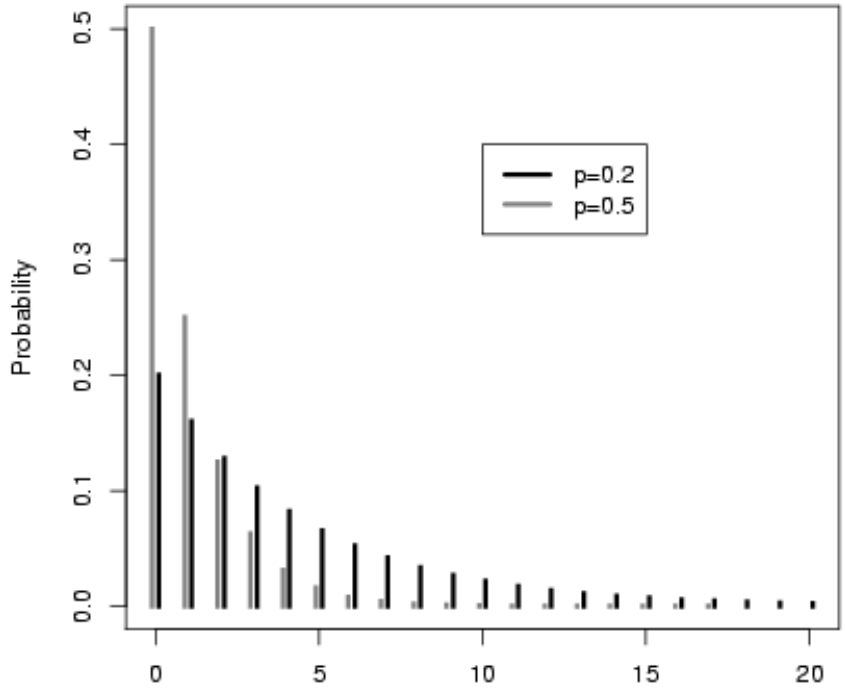
Semi-Markov HMMs (a.k.a. Generalized HMMs)

- to encode distance preferences, models use semi-Markov states for the background
- key idea: decouple length from composition

Duration Modeling in HMMs

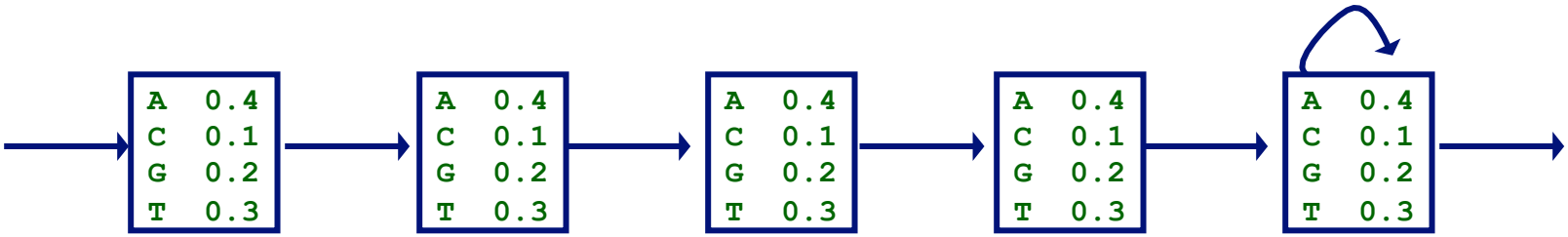
- suppose we have a type of sequence for which the base distribution is the same regardless of length
- the simplest way to model it:



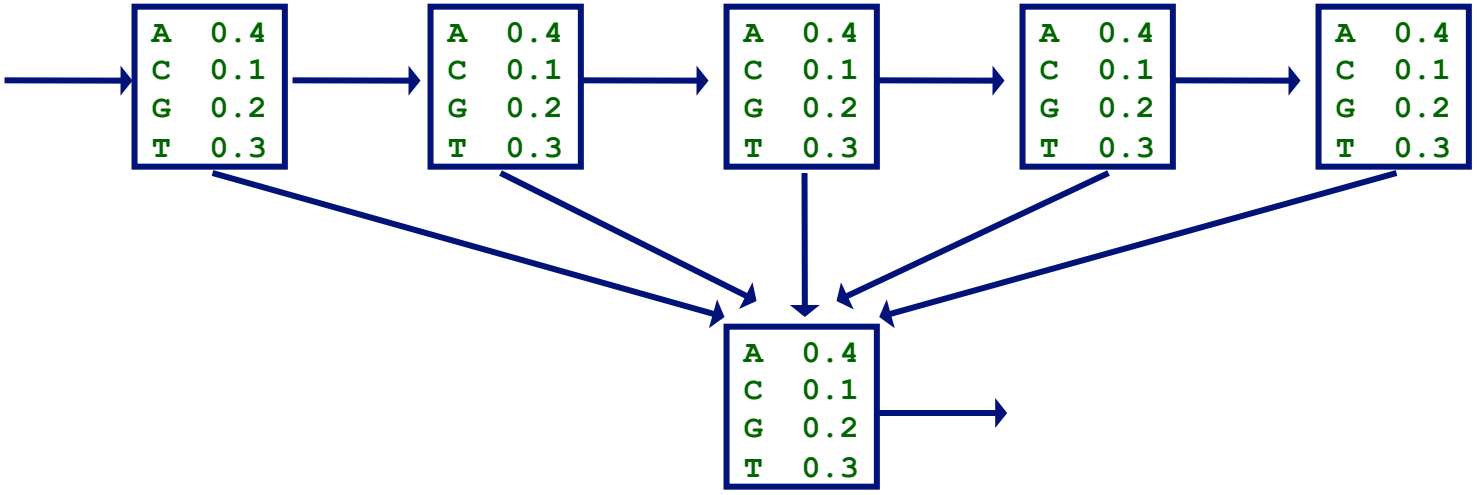


- this encodes a *geometric* distribution (shifted by 1) on the length of sequences

Duration Modeling in HMMs



- min length = 5; geometric distribution over longer sequences



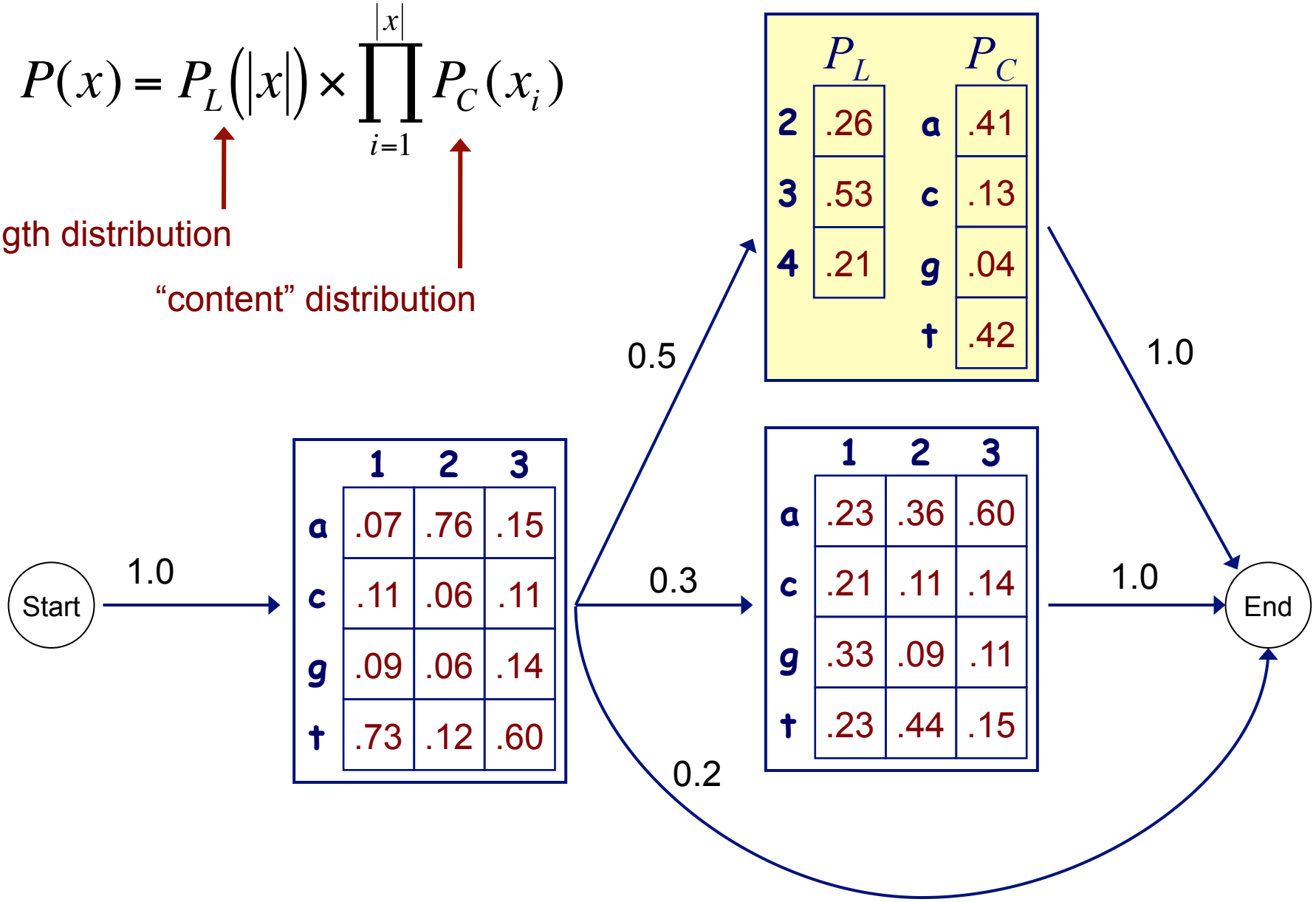
- any distribution over length 2 to 6

A Simple Semi-Markov Model

$$P(x) = P_L(|x|) \times \prod_{i=1}^{|x|} P_C(x_i)$$

length distribution

"content" distribution



Semi-Markov Models

- represent a parse π, as a sequence of states and associated lengths (durations)

$$\vec{q} = \{q_1, q_2, \ldots, q_n\} \qquad \vec{d} = \{d_1, d_2, \ldots, d_n\}$$

- the joint probability of generating parse π and sequence x

$$P(x,\pi) = a_{start,q_1} P(d_1 \mid q_1) P(x_1 \mid q_1, d_1) \times \prod_{k=2}^{n} a_{q_{k-1},q_k} P(d_k \mid q_k) P(x_k \mid q_k, d_k)$$

transition probabilities

the k^{th} segment of the sequence

DP with Semi-Markov Models

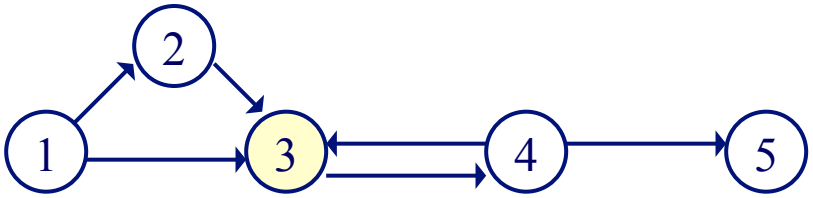
- review: Forward algorithm recurrence for HMMs

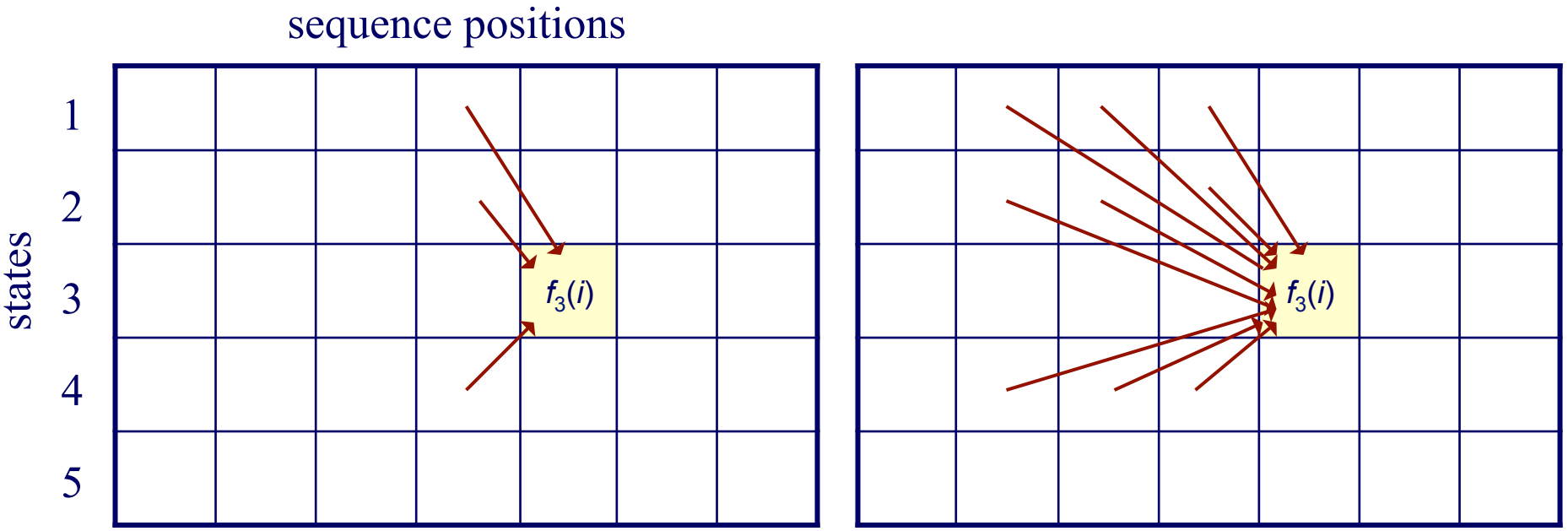
$$f_l(i) = \sum_k f_k(i-1)\ \underbrace{a_{kl}}_{\text{transition from } k \text{ to } l}\ \underbrace{P(x_i \mid q_l)}_{\text{prob. of emitting } x_i \text{ from } l}$$

- for semi-Markov models: each Forward value assumes we're ending a segment in the given state

$$f_l(i) = \sum_k \sum_{d=1}^{D} \left[f_k(i-d)\ a_{kl}\ \underbrace{P(d \mid q_l)}_{\text{prob. of length } d \text{ segment from } l}\ \underbrace{\prod_{j=i-d+1}^{i} P(x_j \mid q_l)}_{\text{prob. of emitting } x_{i-d+1} \ldots x_i \text{ from } l} \right]$$

DP with Semi-Markov Models





complexity of Viterbi/Forward/Backward in standard HMMs is $O(S^2L)$ where S = number of states, L = sequence length

complexity in semi-Markov HMMs is $O(S^2LD)$ where D = maximum length of a segment

Speeding up this DP

- key idea: don't fill in elements in the DP matrix that are likely to have small values

- operationalize this idea by:
  - scan motif models across sequence to see where they have strong matches
  - prune DP matrix elements that don't line up with these matches

Speeding up this DP

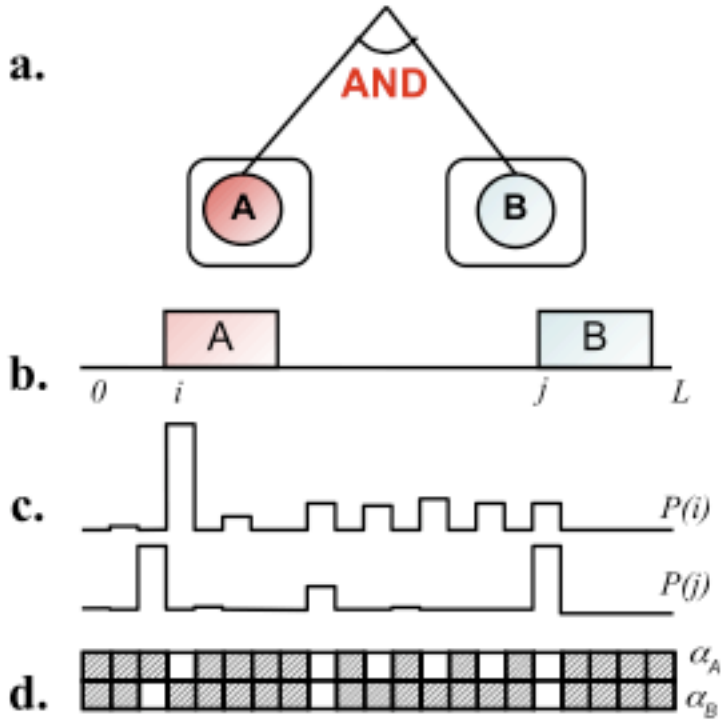
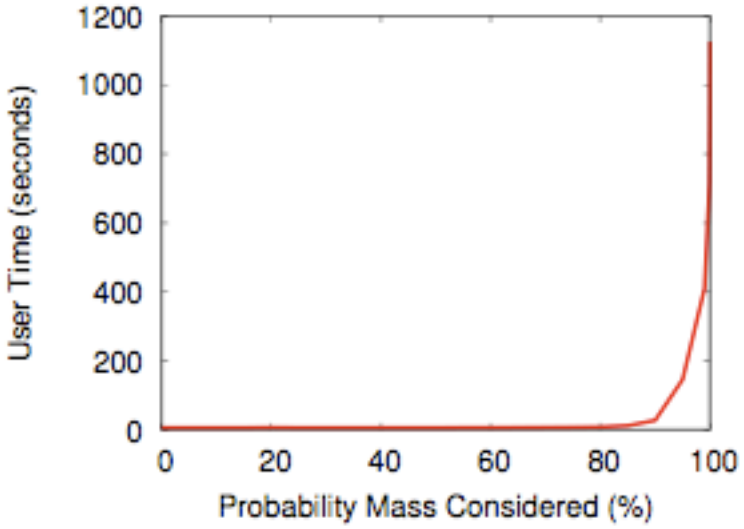


Figure 4.8 Illustration of efficient dynamic programming in SCRM2. **a.** A CRM logical structure. **b.** Possible binding site locations on a DNA sequence **x**. **c.** A probability distribution over the locations of binding sites A and B, respectively. These probabilities tend to be extreme (a motif is present at a location or it is not) and high probabilities are sparsely distributed. **d.** A forward dynamic programming matrix f, where $f_A(i)$ represents the likelihood of sequence **x** from location 1 to i when site A occurs at location i.

Speeding up this DP



Probability Mass Considered (%)	User Time (min:sec)
100	18:45
99.99	15:22
99.9	11:52
99	6:50
95	2:26
90	0:27.2
75	0:7.19
50	0:5.78
10	0:6.00
1	0:5.79

can consider parses comprising 90% of probability mass in 2.4% of the time required to consider all parses

Figure 4.9 Time to train one two-binding site CRM model on 500 bp yeast sequences from [Lee *et al.*, 2002] as a function of the motif location probability mass examined during SCRM2's dynamic programming calculations.

Learning The HMM Structure