Lecture 10 - Learning Motif Models with Gibbs sampling

Colin Dewey

February 20, 2008

EM Theory

- Estimate parameters for models with latent (hidden) states
- Model: X (observed), Z (latent), Θ (params)
- Want to maximize log P(X| Θ)
- Much easier to maximize log P(X,Z| Θ) but don’t know Z
- Instead, maximize expected value of log P(X,Z| Θ)
- Alternate expectation (Z) and maximization (Θ) computations
- Theorem: this also maximizes (locally) log P(X | Θ)

Gibbs Sampling: An Alternative to EM

- a general procedure for sampling from the joint distribution of a set of random variables $\Pr(X_1, \ldots, X_n)$
- Iteratively sample from

  $$\Pr(X_j | X_1, \ldots, X_{j-1}, X_{j+1} \ldots X_n)$$

  for each j
- application to motif finding: Lawrence et al. 1993
- can view it as a stochastic analog of EM for this task
- less susceptible to local minima than EM

Gibbs Sampling Approach

- in the EM approach we maintained a distribution Z_i over the possible motif starting points for each sequence
- in the Gibbs sampling approach, we'll maintain a specific starting point a_i for each sequence but we'll keep randomly resampling these

Gibbs Sampling Approach

```
given: length parameter $W$, training set of sequences
    choose random positions for $a$
    do
        pick a sequence $X_i$
         estimate $p$ given current motif positions $a$  (update step)
              (using all sequences but $X_i$ )
          sample a new motif position $a_i$ for $X_i$  (sampling step)
    until convergence
return: $p$, $a$
```

Sampling New Motif Positions

- for each possible starting position, $a_i = j$, compute a weight

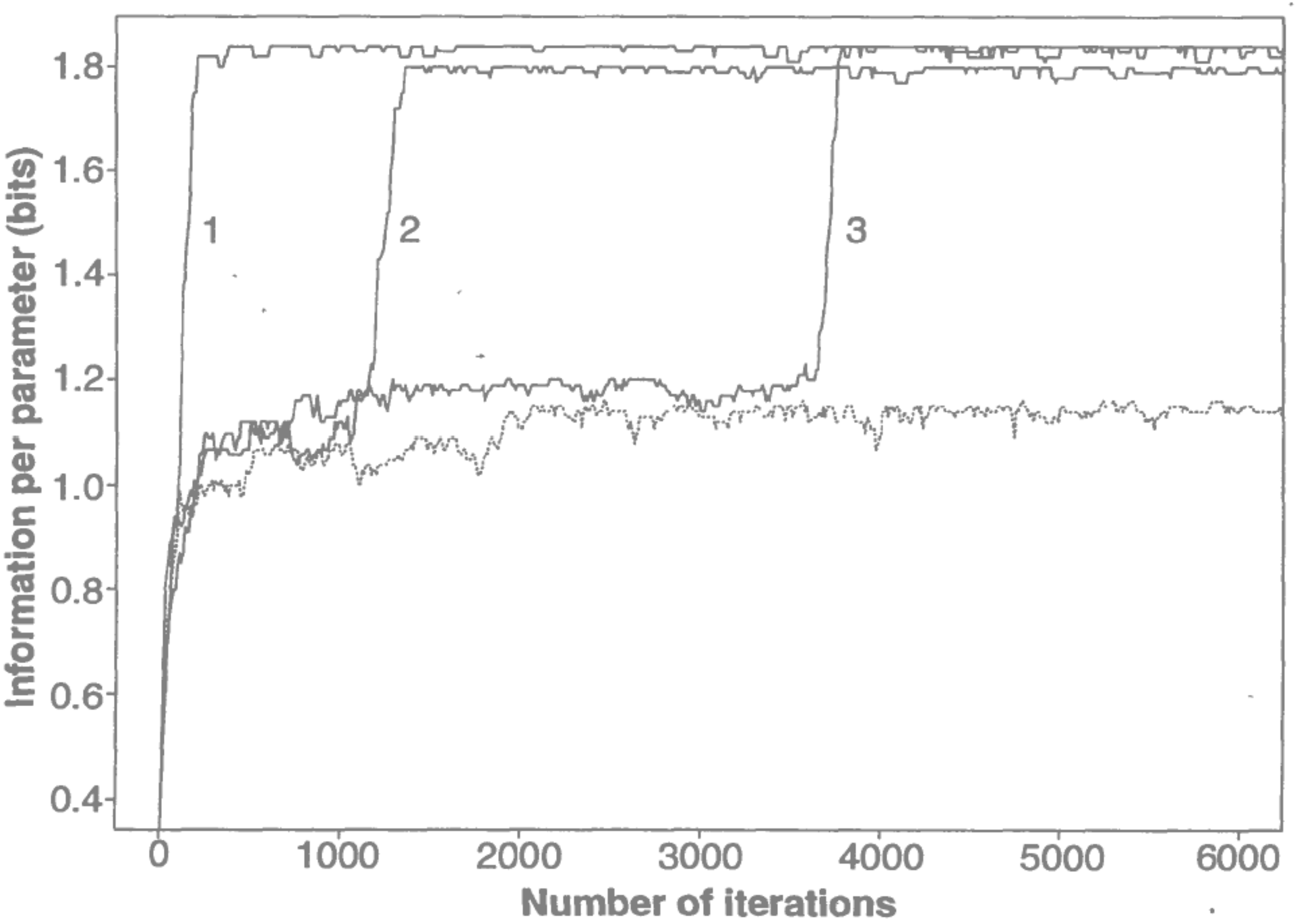
$$A_j = \frac{\prod_{k=j}^{j+W-1} p_{c_k, k-j+1}}{\prod_{k=j}^{j+W-1} p_{c_k, 0}}$$

- randomly select a new starting position according to these weights a_i

The Phase Shift Problem

- Gibbs sampler can get stuck in a local maxima that corresponds to the correct solution shifted by a few bases
- Solution: add a special step to shift the a values by the same amount for all sequences. Try different shift amounts and pick one in proportion to its probability score.

Convergence of Gibbs



Markov Chain Monte Carlo

- Technique for sampling from probability distribution
- Construct Markov chain with stationary distribution equal to distribution of interest
- Transition probability: $\tau(y|x) \;\; x \longrightarrow y$
- Detailed balance: $\mathbb{P}(x)\tau(y|x) = \mathbb{P}(y)\tau(x|y)$
- If detailed balance, then: $\frac{1}{N} \lim_{N\to\infty} C(y_i = x) = \mathbb{P}(x)$

MCMC with Gibbs sampling

- Markov chain transitions by changing one variable at a time
- Transition probability is conditional distribution of the variable given all others
- Show that this obeys detailed balance

$$\mathbb{P}(X_1, X_2, \ldots, X_N)$$

$$\tau(X_i^{t+1}|X_i^t) = \mathbb{P}(X_i^{t+1}|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_N)$$

EM and Gibbs

- these methods are computing a *local*, *multiple* alignment
- both methods try to optimize the likelihood of the sequences
- EM converges to a local maximum
- Gibbs will converge to a global maximum, *in the limit*
- MEME can take advantage of background knowledge by
  - tying parameters
  - Dirichlet priors

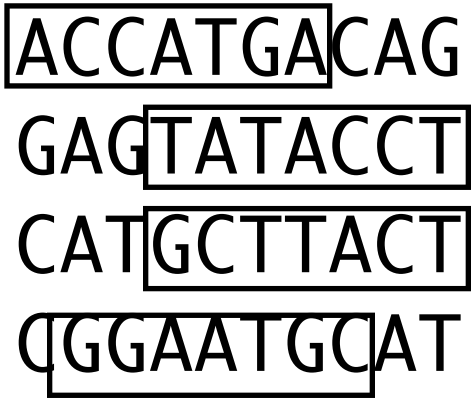
Example: The Data

- Hidden motif of width 7 in 4 sequences of length 10
- Each motif occurrence differs from consensus (GATTACA) in two positions

```
ACCATGACAG
GAGTATACCT
CATGCTTACT
CGGAATGCAT
```

Initialization

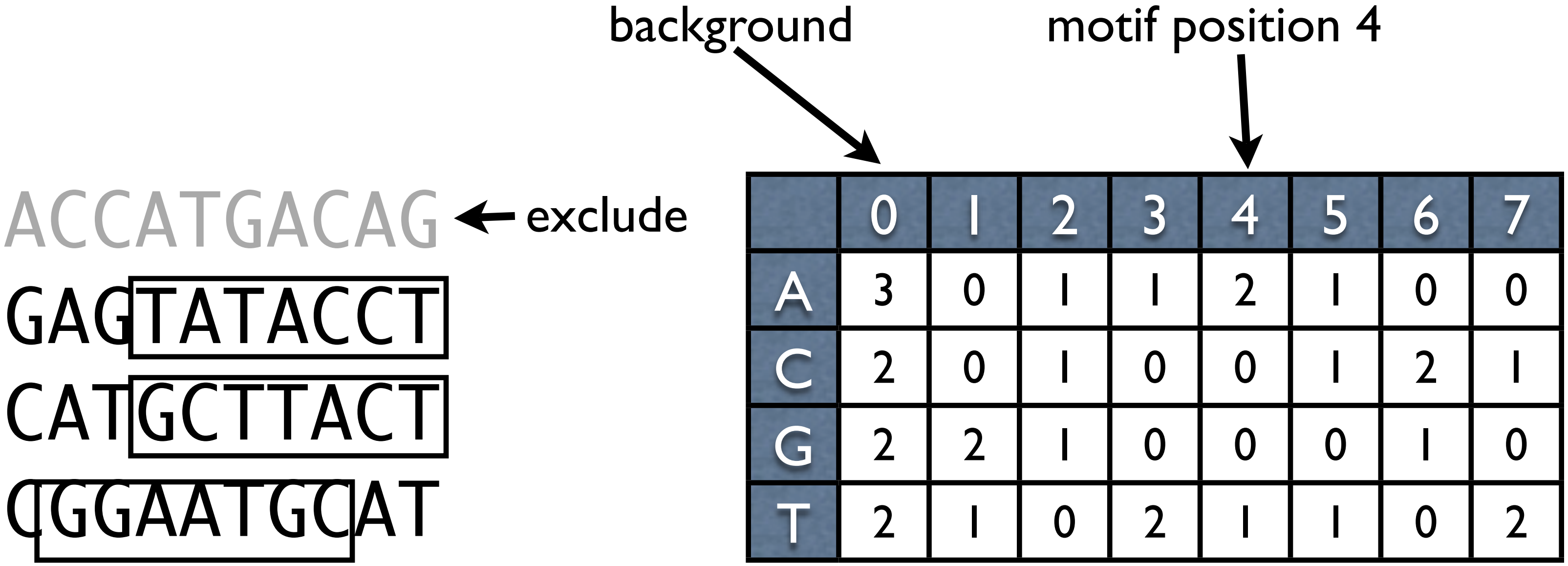
- Choose initial positions of motif at random



ACCATGACAG
GAGTATACCT
CATGCTTACT
CGGAATGCAT

Predictive update step

- Update profile matrix based on motif and background frequencies and pseudocounts



ACCATGACAG

GAGTATACCT

CATGCTTACT

CGGAATGCAT

	0	1	2	3	4	5	6	7
A	3	0	1	1	2	1	0	0
C	2	0	1	0	0	1	2	1
G	2	2	1	0	0	0	1	0
T	2	1	0	2	1	1	0	2

Predictive update step

- Calculate profile matrix from frequencies and pseudocounts

	0	1	2	3	4	5	6	7
A	3	0	1	1	2	1	0	0
C	2	0	1	0	0	1	2	1
G	2	2	1	0	0	0	1	0
T	2	1	0	2	1	1	0	2

$$p_{1,\texttt{A}} = \frac{c_{1,\texttt{A}} + b_{\texttt{A}}}{N-1+B} = \frac{0+0.5}{4-1+2} = 0.1$$

$$p_{1,\texttt{C}} = \frac{c_{1,\texttt{C}} + b_{\texttt{C}}}{N-1+B} = \frac{0+0.5}{4-1+2} = 0.1$$

$$p_{1,\texttt{G}} = \frac{c_{1,\texttt{G}} + b_{\texttt{G}}}{N-1+B} = \frac{2+0.5}{4-1+2} = 0.5$$

$$p_{1,\texttt{T}} = \frac{c_{1,\texttt{T}} + b_{\texttt{T}}}{N-1+B} = \frac{1+0.5}{4-1+2} = 0.3$$

$$p_{0,\texttt{A}} = \frac{c_{0,\texttt{A}} + b_{\texttt{A}}}{\sum_{i \neq 1}(\ell_i - W) + B} = \frac{3+0.5}{3(3)+2} = \frac{7}{22}$$

$$p_{0,\texttt{C}} = \frac{c_{0,\texttt{C}} + b_{\texttt{C}}}{\sum_{i \neq 1}(\ell_i - W) + B} = \frac{2+0.5}{3(3)+2} = \frac{5}{22}$$

$$p_{0,\texttt{G}} = \frac{c_{0,\texttt{G}} + b_{\texttt{G}}}{\sum_{i \neq 1}(\ell_i - W) + B} = \frac{2+0.5}{3(3)+2} = \frac{5}{22}$$

$$p_{0,\texttt{T}} = \frac{c_{0,\texttt{T}} + b_{\texttt{T}}}{\sum_{i \neq 1}(\ell_i - W) + B} = \frac{2+0.5}{3(3)+2} = \frac{5}{22}$$

	0	1	2	3	4	5	6	7
A	0.3	0.1	0.3	0.3	0.5	0.3	0.1	0.1
C	0.2	0.1	0.3	0.1	0.1	0.3	0.5	0.3
G	0.2	0.5	0.3	0.1	0.1	0.1	0.3	0.1
T	0.2	0.3	0.1	0.5	0.3	0.3	0.1	0.5

Sampling step

- For each possible motif start position, calculate ratio of likelihood of next W positions from motif vs. background

	1	2	3	4
A_i	0.16	0.13	0.26	0.017

$$A_1 = \frac{p_{1,\texttt{A}} \cdot p_{2,\texttt{C}} \cdot p_{3,\texttt{C}} \cdot p_{4,\texttt{A}} \cdot p_{5,\texttt{T}} \cdot p_{6,\texttt{G}} \cdot p_{7,\texttt{A}}}{p_{0,\texttt{A}} \cdot p_{0,\texttt{C}} \cdot p_{0,\texttt{C}} \cdot p_{0,\texttt{A}} \cdot p_{0,\texttt{T}} \cdot p_{0,\texttt{G}} \cdot p_{0,\texttt{A}}} \approx \frac{0.1 \cdot 0.3 \cdot 0.1 \cdot 0.5 \cdot 0.3 \cdot 0.3 \cdot 0.1}{0.31 \cdot 0.23 \cdot 0.23 \cdot 0.31 \cdot 0.23 \cdot 0.23 \cdot 0.31} \approx 0.16$$

ACCATGACAG

	0	1	2	3	4	5	6	7
A	0.3	0.1	0.3	0.3	0.5	0.3	0.1	0.1
C	0.2	0.1	0.3	0.1	0.1	0.3	0.5	0.3
G	0.2	0.5	0.3	0.1	0.1	0.1	0.3	0.1
T	0.2	0.3	0.1	0.5	0.3	0.3	0.1	0.5

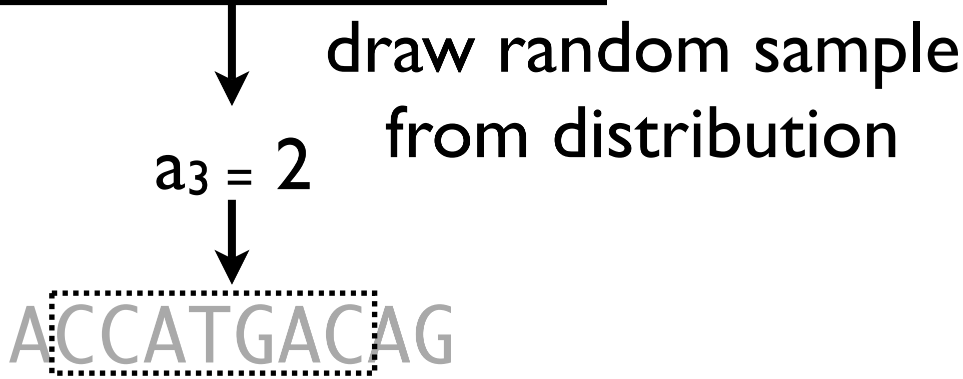
Sampling step

- Sample new position *i* in chosen sequence based on A_i

	1	2	3	4
A_i	0.16	0.13	0.26	0.017

↓ normalize

	1	2	3	4
A_i	0.28	0.23	0.46	0.03



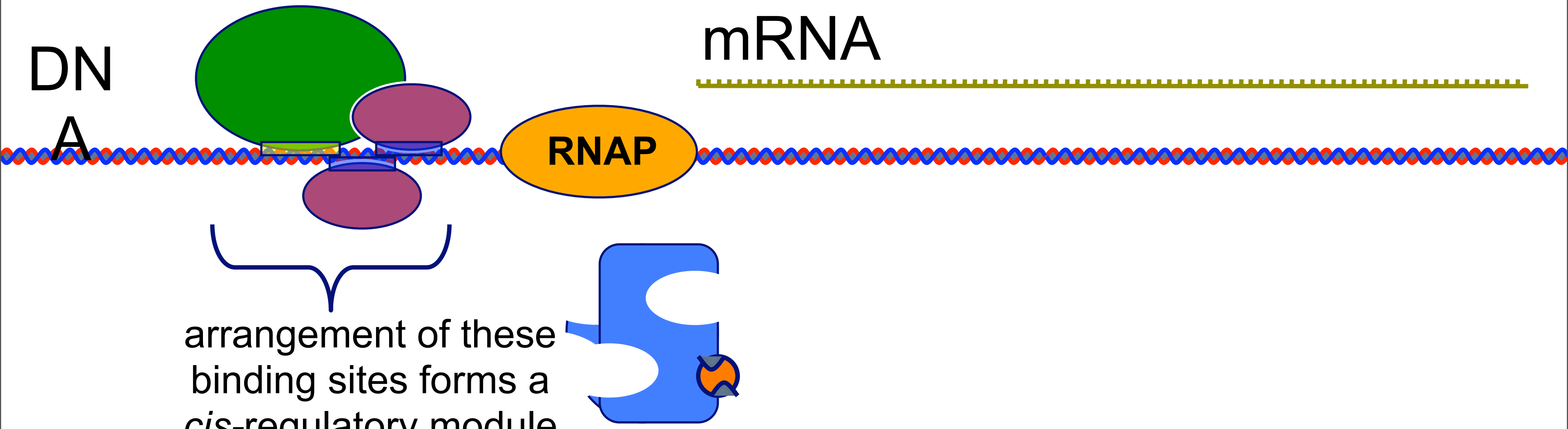
draw random sample from distribution

$a_3 = 2$

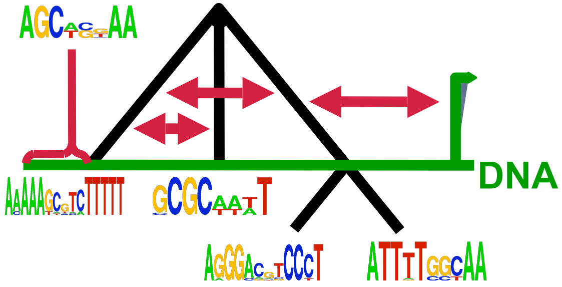
Calculate likelihood

- Calculate likelihood (or some related value) after each iteration
- Iterate:
  - choose sequence
  - predictive update
  - sample new motif position in sequence
- After many iterations, choose motif positions and corresponding profile matrix

Inferring *cis* Regulatory Modules (CRMs)

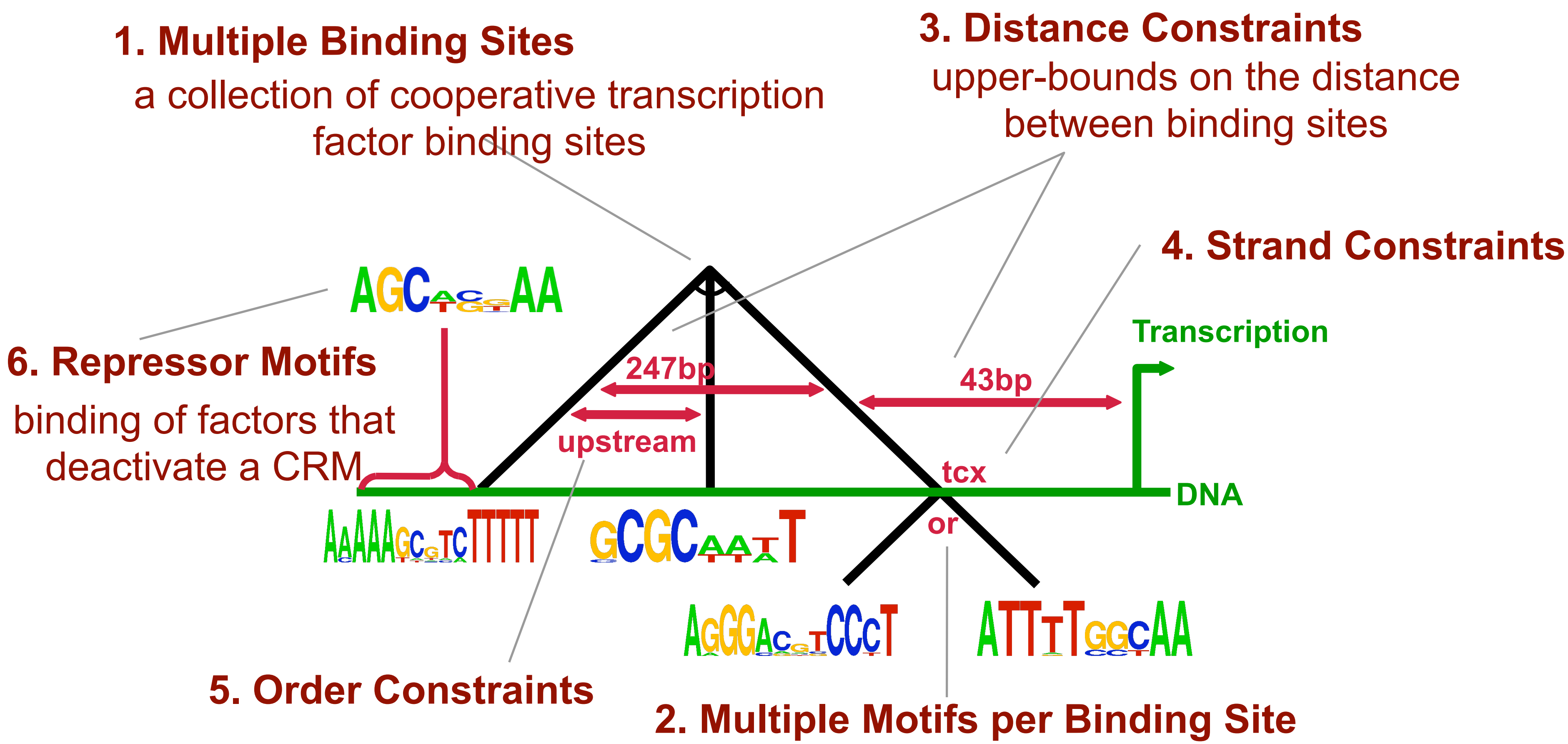


arrangement of these binding sites forms a *cis*-regulatory module



a task of growing interest: infer models of CRMs that regulate certain sets of genes

A Representation for CRMs [Noto & Craven]



1. Multiple Binding Sites

a collection of cooperative transcription factor binding sites

3. Distance Constraints

upper-bounds on the distance between binding sites

4. Strand Constraints

6. Repressor Motifs

binding of factors that deactivate a CRM

5. Order Constraints

2. Multiple Motifs per Binding Site