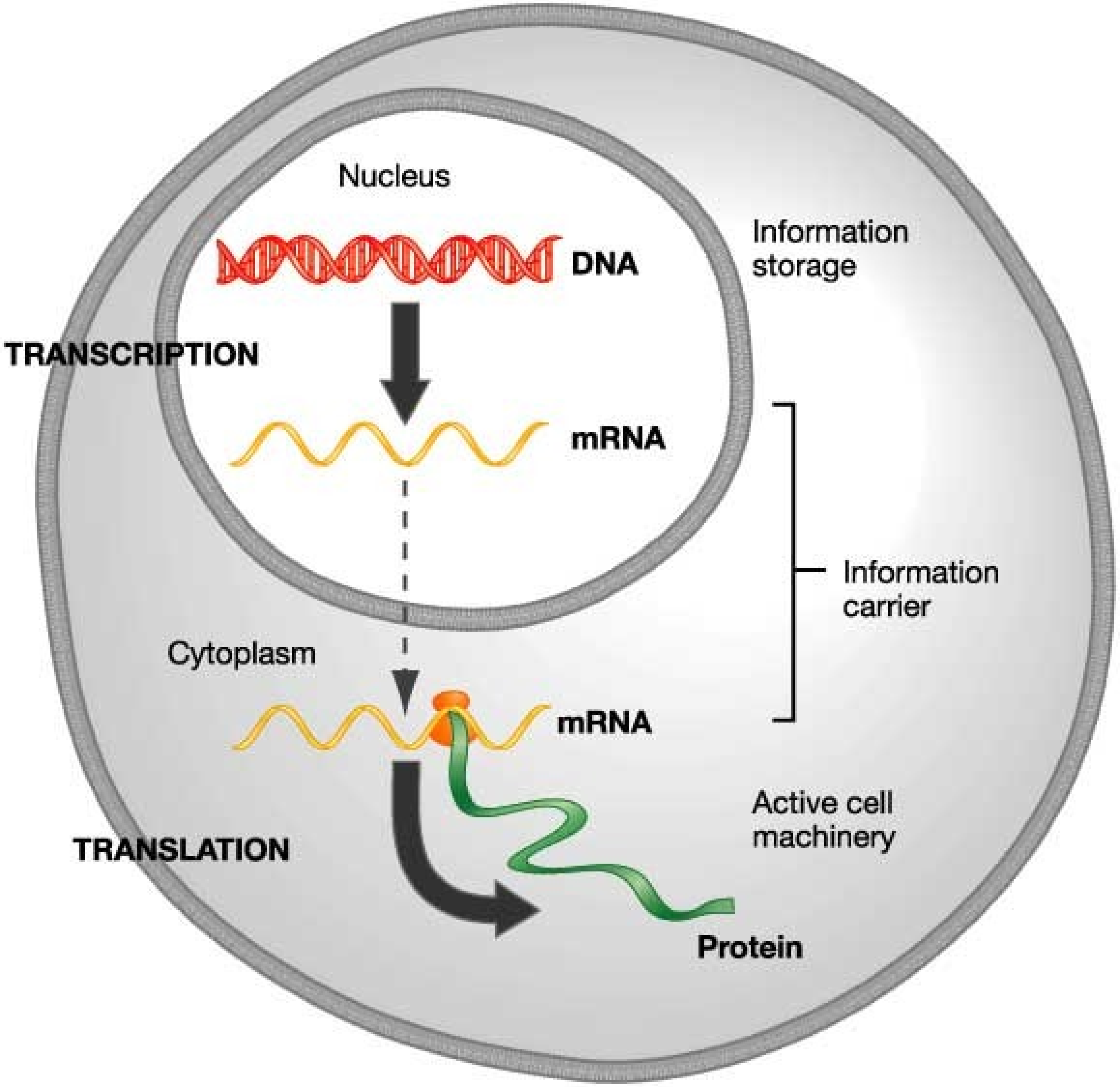
Gene regulation

- DNA is merely the blueprint
- Shared spatially (among all tissues) and temporally
- But cells manage to differentiate
  - Especially but not only during developmental stage
- And cells respond to external conditions and/or messages from other cells
- Much of this dynamic response is attained through protein or gene regulation:
  - how much and which variant of the gene is present

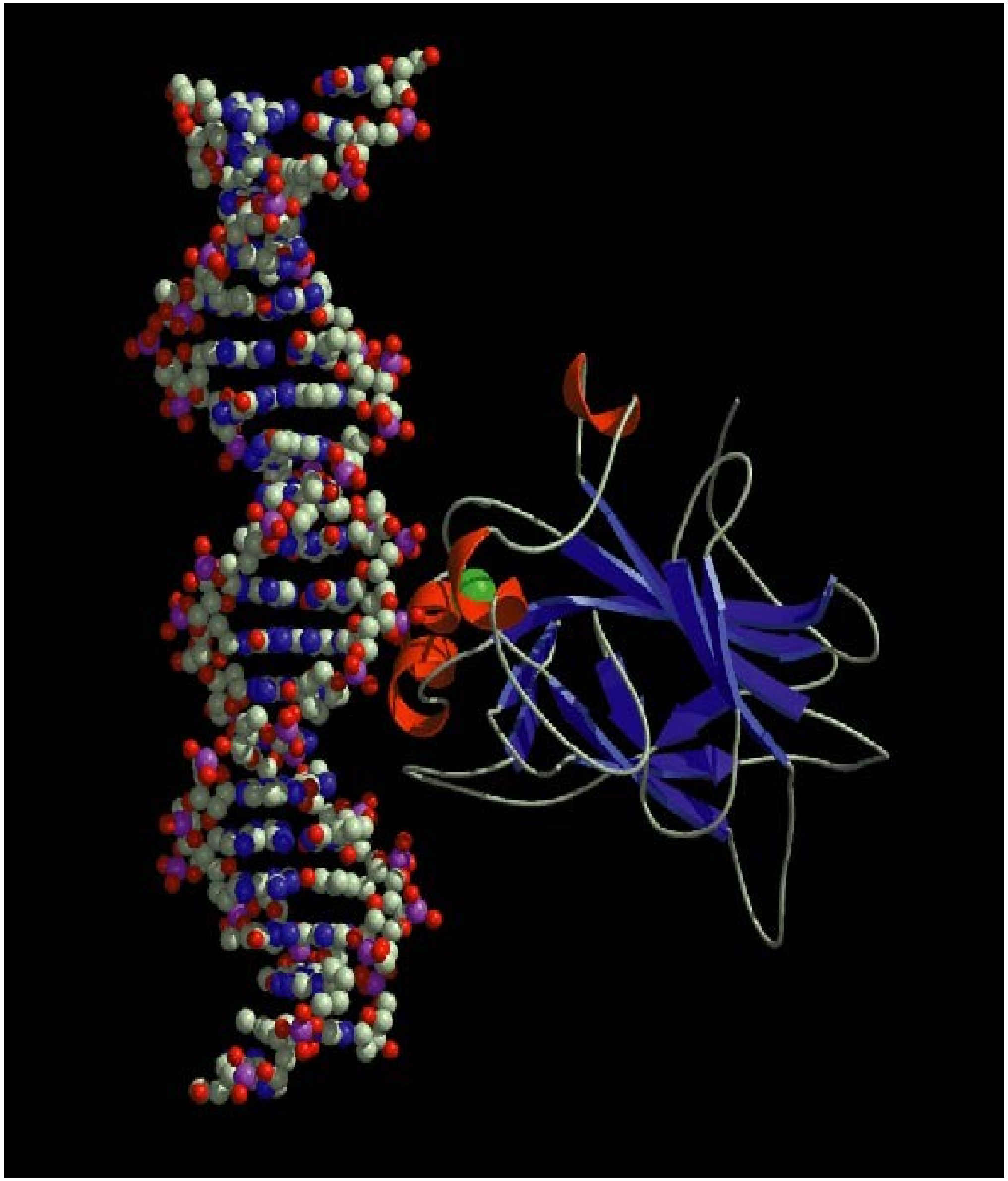
The central dogma



Mechanisms of gene regulation

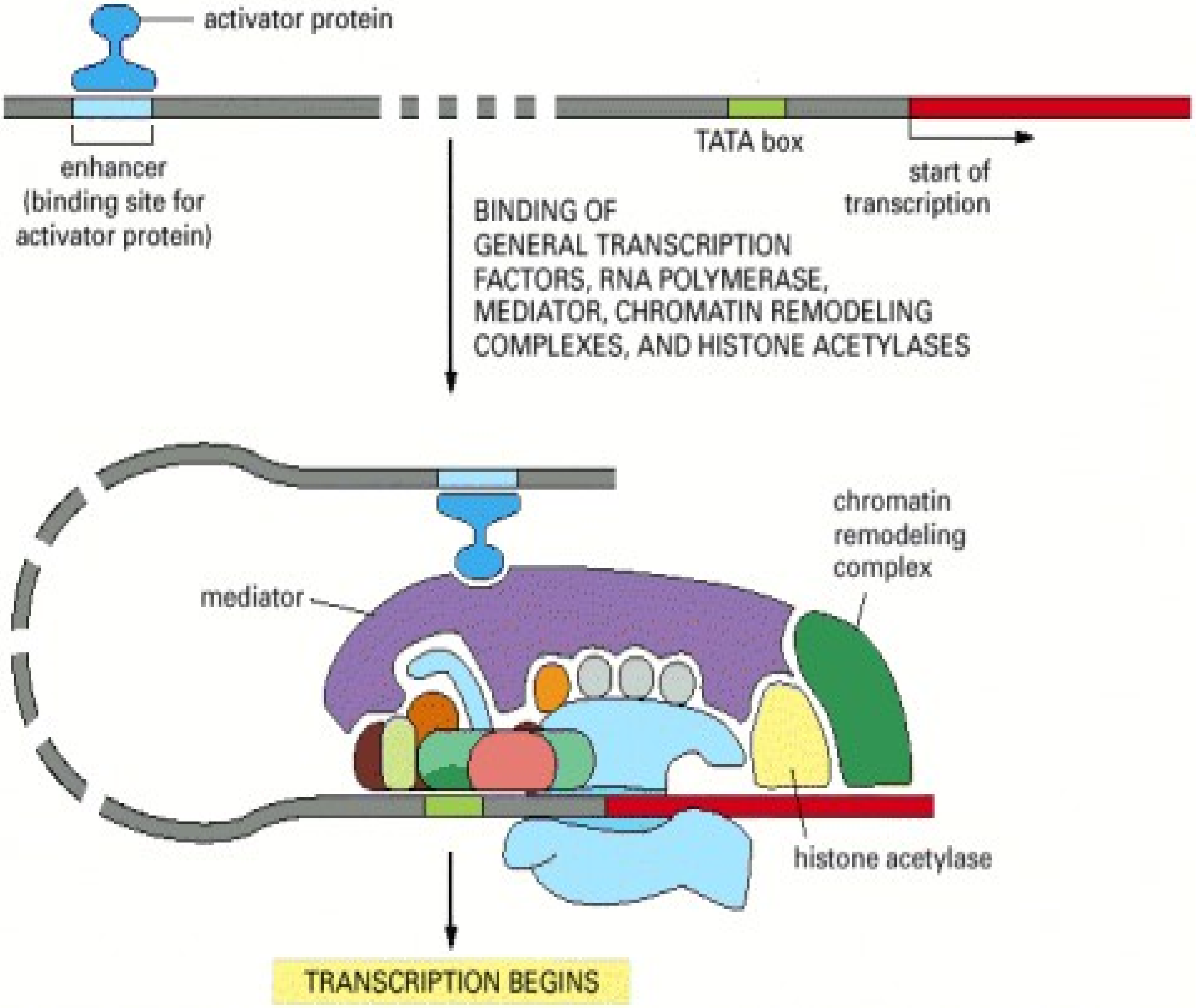
- Pre-transcription: accessibility of the gene
  - the chromatin structure which packs the DNA is dynamic
- Transcription: rate
- Post-transcription: mRNA degradation rate
- Translation: rate
- Post-translation:
  - Modifications
  - Rate of degradation

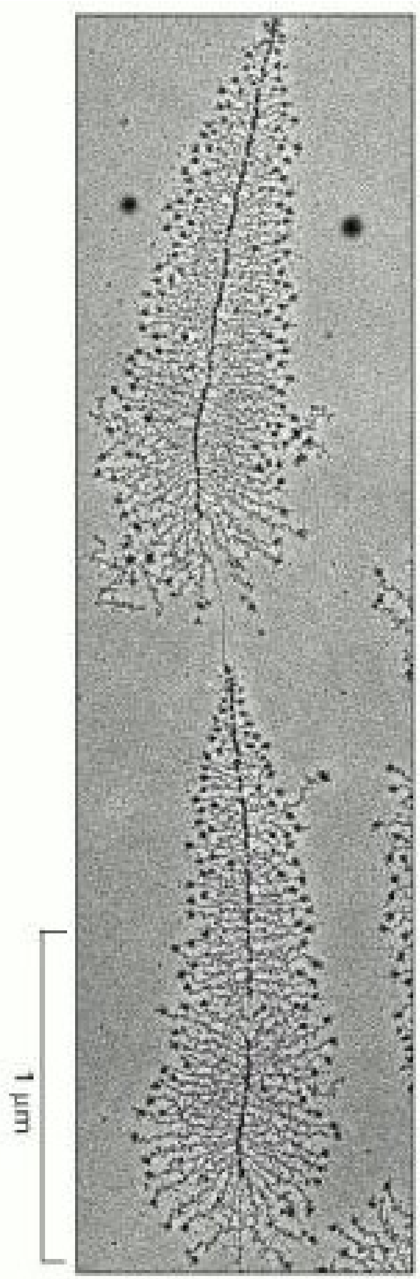
Transcription factors



- Bind to specific DNA sites: Transcription Factor Binding Sites
- Typically downstream effect on mRNA transcription rate

Transcription rate





Motif finding

- Motif finding is the computational problem of identifying TFBSs
- Implicit assumption: different TFBSs of the *same* TF should be similar to another
    - Hence the name motif
- Two related tasks:
    - Given a specific model of TF motif compiled from a known list of TFBSs find additional sites (scanning)
    - Identify the unknown motif given only the DNA sequences

Modelling motifs

- Discovered sites:

TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATATT

- How do we model the motif?
  - important for finding additional sites

- Consensus pattern: TATAAT
  - generalizes to regular expressions

- Positional profile:

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

Generative models

- Consensus pattern: each instance is a randomly mutated version of the consensus
  - substitution only: the same TF binds to the various sites, so indels are unlikely to occur as the DNA-TF contact region remains the same
- Profile: instances are drawn according to the probability implied by the positional profile assuming each position is drawn independently
  - Pseudocounts are typically added to avoid excluding unseen letters

Counts to frequencies profile

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

	1	2	3	4	5	6
A	0.1	0.7	0.1	0.5	0.5	0.1
C	0.1	0.1	0.2	0.1	0.2	0.1
G	0.2	0.1	0.1	0.3	0.1	0.1
T	0.6	0.1	0.6	0.1	0.2	0.7

What is the pseudocount in this example?

The fitness of a TFBS

- How well does a putative TFBS w fits the model?
- For a consensus model we typically use $s_C(w) = d_H(C, w)$, the Hamming distance to the consensus pattern C.
    - It is convenient to work with but more appropriate for uniform nucleotide sample
- For a profile parametrized by $M = (f_{ik})_{i=1:l,k=1:4}$, it is natural to use the likelihood score: $s_M(w) = P_M(w) = \prod_{i=1}^{l} f_{iw_i}$
- Better: use the LLR (loglikelihood ratio) score

$$s_M(w) = \log \frac{P_M(w)}{P_B(w)} = \sum_{i=1}^{l} \log \frac{f_{iw_i}}{b_{w_i}},$$

where B specifies an iid background model with nucleotide frequency $(b_k)_1^4$, typically taken from the organism or the scanned sample

Scanning for TFBS

- Given a parametrized motif model and an associated fitness function looking for additional sites is algorithmically trivial

- However, setting a cutoff score typically requires carefully analyzing the FP rates

- These FP rates are set using a model of random sequences
  - Markov chains
  - shuffling
  - using random chunks of DNA

Motif finding

- Do these sequences share a common TFBS?

- tagcttcatcgttgacttctgcagaaagcaagctcctgagtagctggccaagcgagc
  tgcttgtgcccggctgcggcggttgtatcctgaatacgccatgcgccctgcagctgc
  tagaccctgcagccagctgcgcctgatgaaggcgcaacacgaaggaaagacgggacc
  agggcgacgtcctattaaaagataatcccccgaacttcatagtgtaatctgcagctg
  ctcccctacaggtgcaggcacttttcggatgctgcagcggccgtccggggtcagttg
  cagcagtgttacgcgaggttctgcagtgctggctagctcgacccggattttgacgga
  ctgcagccgattgatggaccattctattcgtgacacccgacgagaggcgtcccccg
  gcaccaggccgttcctgcagggccaccctttgagttaggtgacatcattcctatgt
  acatgcctcaaagagatctagtctaaatactacctgcagaacttatggatctgaggg
  agaggggtactctgaaaagcgggaacctcgtgtttatctgcagtgtccaaatcctat

If only life could be that simple

- The binding sites are almost never exactly the same
- A more likely sample is:

```
tagcttcatcgttgactttTGaAGaaagcaagctcctgagtagctggccaagcgagc
tgcttgtgcccggctgcggcggttgtatcctgaatacgccatgcgccCTGgAGctgc
tagaccCTGCAGccagctgcgcctgatgaaggcgcaacacgaaggaaagacgggacc
agggcgacgtcctattaaaagataatccccgaacttcatagtgtaatCTGCAGctg
ctcccctacaggtgcaggcacttttcggatgCTGCttcggccgtccggggtcagttg
cagcagtgttacgcgaggttCTaCAGtgctggctagctcgacccggattttgacgga
CTGCAGccgattgatggaccattctattcgtgacacccgacgagaggcgtcccccg
gcaccaggccgttcCTaCAGgggccaccctttgagttaggtgacatcattcctatgt
acatgcctcaaagagatctagtctaaatactacCTaCAGaacttatggatctgaggg
agaggggtactctgaaaagcgggaacctcgtgtttattTGCAttgtccaaatcctat
```

Searching for motifs

- Simultaneously looking for a motif model and sites that will optimize a scoring function is significantly more difficult
- Assume for simplicity the OOPS model (One Occurrence Per Sequence model): $w^m \in S^m$ for $m = 1 : n$
- A natural way to score a putative combination of a motif M and sites $(w^m)_1^n$ is by summing the fitness scores of all sites:

$$s(M; w^1, \ldots, w^n) := \sum_{m=1}^{n} s_M(w^m)$$

- Thus, our goal is to search the joint space of motifs, M (consensus or profile), and alignments, $w^m \in S^m$, so as to optimize this score
- Fortunately, for both models this can be done sequentially so we do not have to optimize simultaneously over the alignment and the motif

Optimizing the motif or the alignment

- Once we choose the alignment, $w^m \in S^m$ for $m = 1 : n$, the optimal motif for that alignment is trivial

- For the consensus model it is a consensus word as it clearly minimizes the total distance to the words in the alignment

- For the profile model we find with a little more effort that the best model is the one which coincides with how we define a profile: $f_{ik} = \frac{n_{ik}}{n}$, where n_{ik} is the number of occurrence of the letter k at position i.

- Conversely, if we know the model we can find the optimal sites for the putative motif by linearly scanning the sequences

- Often a motif finder will combine both the motif's and the alignment's optimizations and indeed they are in some sense equivalent

Heuristic vs. guaranteed optimizations

- Assume for now l is known (we can enumerate over possible ls) and let N_m be the length of S^m
- By considering all, roughly, $\prod_{m=1}^{n} N_m$ gapless alignments made of $w^m \in S^m$ we are guaranteed to find the optimal alignment under both possible motif models
    - Unfortunately, this number is prohibitively expensive for all but a few cases

Finding an optimal pattern

- Consistent with our previous discussion under the OOPS model the score of a consensus word C is often the *total distance*:

$$TD(C) := \sum_{m=1}^{n} d_H(C, S^m) = \sum_{m=1}^{n} \min_{w' \in S^m} d_H(C, w')$$

- Problem: find a word C that minimizes the total distance
- Naive solution: enumerate all 4^l possible consensus words
  - Complexity: $O(4^l D)$
  - While this approach is feasible for a larger set of parameters than the one available for alignment enumeration it is still often too expensive

Heuristic approaches: Sample Driven

- Most of the 4^l patterns we explore in the exhaustive enumeration have little to do with our sample
- Sample driven approach: compute $TD(w)$ only for words w in the sample
- Complexity: $O(D^2)$ where $D = \sum_{m=1}^{n} N_m$ is the size of the sample
- Analysis:
    - fast
    - but can miss the optimal pattern if it is missing from the sample
- More sophisticated methods were developed based on the sample driven approach

CONSENSUS - greedy profile search (Hertz & Stormo '99)

- Assume the OOPS model and that l is given
  - There is a version that does not assume l is given (WCONSENSUS)
- CONSENSUS Follows a greedy strategy looking first for the best alignment of just two sites:
  - For each $i \neq j$, and $w \in S^i$, $w' \in S^j$ compute the information content of the alignment made of w and w':

$$I = \sum_{i=1}^{l} \sum_{k=1}^{4} n_{ik} \log \frac{n_{ik}/2}{b_k}$$

  - Keep the top q_2 alignments (matrices)

- It then greedily adds one word at a time from the sequences that are not already represented in the alignment

- Let $m := 3$ denote the number of sequences in the current alignments

- While $m < n$
  - for each of the top saved q_{m-1} alignments A of $m-1$ rows compute $I\left(\left[\begin{array}{c} A \\ w \end{array}\right]\right)$ for all words w which come from sequences that are not already in A
  - keep the best q_m alignments and set $m := m + 1$

MEME (Bailey & Elkan '94)

- MEME: Multiple EM for Motif Elicitation
  - the multiple part is for dealing with multiple motifs
  - probabilistic generative model, deterministic algorithm
- Recall that given the motif model we can linearly scan the sequences for instances
- Conversely, given the instances deducing the profile is trivial
- MEME alternates between the two tasks

MEME's outline

- Starting from a heuristically chosen initial profile
  - Sample driven: the profile is derived from the word in the sample that has a minimal total distance
- MEME iterates the following two steps until convergence
  - score each word according to how well it fits the current profile
  - update the profile by taking a weighted average of all the words
- The EM in MEME stands for Expectation Maximization (Dempster, Laird & Rubin '77) which MEME's two step procedure follows
  - EM is guaranteed to monotonically converge to a local maximum (intelligent choice of a starting point is crucial)

Gibbs Sampler (Lawrence et al. '93)

- Probabilistic framework, random algorithm
- Assumes the OOPS model for simplicity (many more variants)
- Suppose we selected putative instances $w_i \in S^i$, these define the profile or motif model as in MEME
- As in the EM context we compute the LR score of every word in the sample: $L_w = \frac{P_M(w)}{P_B(w)}$
- In EM we use a soft assignment of words to the list of selected sites (instances), alternatively we can use hard assignment:
  - e.g., we can choose: $w_i = \mathrm{argmax}_{w \in S^i} L_w$
  - or, we can randomly choose a site with probability proportional to its LLR score (Gibbs Sampling)
- Iterate

Gibbs' outline

- Start with a random choice of $w_i \in S_i$
- While there has been an improvement in the total LLR score (over all sites) in the last L (the plateau period) iterations do
  - For $i = 1 \ldots n$: remove word w_i from the motif model M and randomly pick a new site from sequence S^i with probability proportional to L_w (an iteration is one such loop)
- Note that there is no convergence in the naive sense
- There is an alternative formulation in terms of a Gibbs Sampler: the goal is to randomly sample alignments from a distribution where each alignment's probability is roughly proportional to its LR score. The Gibbs Sampler defines an MCMC that converges to a stationary distribution with that property thus allowing us to sample from this distribution.

Is this significant?

- Motif finders always find *something*
- A high scoring motif reported by CONSENSUS
  - The alignment

```
cCGATAAGGTaAG
TCGATAAGGaGAG
TCGATAAaGTaAG
TCGATAAGGTcAG
TCGATAAGGTGAG
```

  - The profile

A				5		5	5	1		1	2	5	
C	1	5									1		
G			5					4	5		2		5
T	4				5					4			

Assessing the significance

- How likely are we to see such alignments or better by chance?
- Clearly you need more information:
  - What is the *null* model, or how are random (chance) sequences generated?
    - ▷ typically iid (independent identically distributed or 0-th order Markov)
  - What is the size of the search space:
    - ▷ How many input sequences are there and how long are they?
    - ▷ What was CONSENSUS instructed to look for?
  - What is a better alignment, or how do we score a motif
    - ▷ information content (Stormo 88): $\sum_{i=1}^{L} \sum_{j=1}^{A} n_{ij} \log \frac{n_{ij}/n}{b_j}$

Quantifying the significance: the E-value / p-value

- The E-value of an alignment with score s is:
  - The expected number of random such alignments with score $\geq s$
  - . . . given the size of the search space
  - An E-value of 0.01 is better than 100
  - It is computed by multiplying the size of the search space by the p-value of the alignment
- The p-value of an alignment with score s is:
  - The probability that the score of a random alignment of the same width and depth is $\geq s$

Phylogenetically aware finders

- So far we looked at de novo motif finders:
    - the only input is the set of presumably co-regulated sequences
- Binding sites are functional and functional elements tend to be more conserved
- Therefore we should look for conserved words in phylogenetically related species
- This can be done when we search for a known motif (MONKEY - Moses et al. 2004)
- Or, when looking for unknown motif (PhyloGibbs - Siddharthan et al. 2005)
- Other finders might add ChIP-chip data (MDscan - Liu et al. 2002)