COM SCI 122 Week 7

Aidan Jan

February 28, 2025

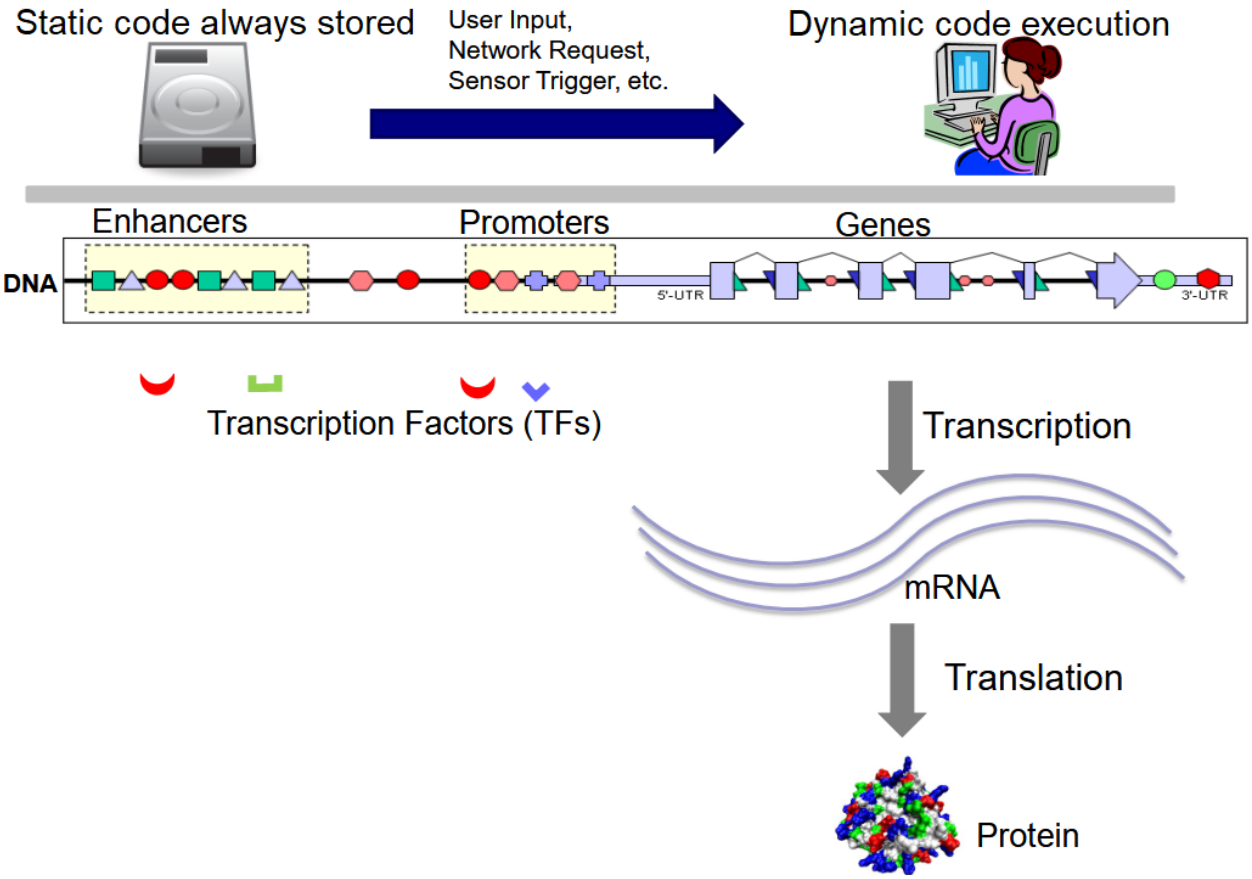
Motifs

The goal: be able to "read" the non-coding DNA sequences. This would facilitate predicting the impact of non-coding mutations

Central Dogma

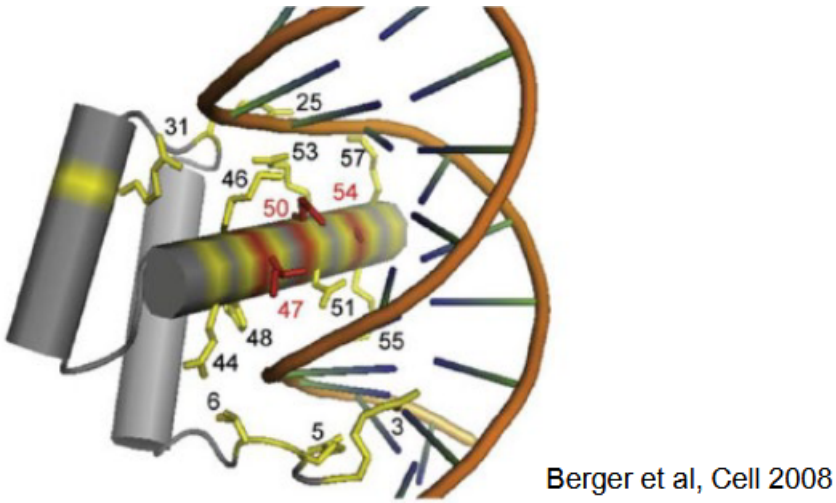
- Recall: DNA $\rightarrow$ mRNA $\rightarrow$ protein $\rightarrow$ phenotype
- The cell needs to regulate the process fo going from DNA to mRNA.
- Transcription factors (TFs) binding DNA can play a major role in controlling the process by leading to the activation or repression of gene expression.
- Thousands of TFs in human.
- How does a transcirption factor know to only bind specific locations in the genome?

Gene Expression and its Regulation



Transcription factor binding to DNA

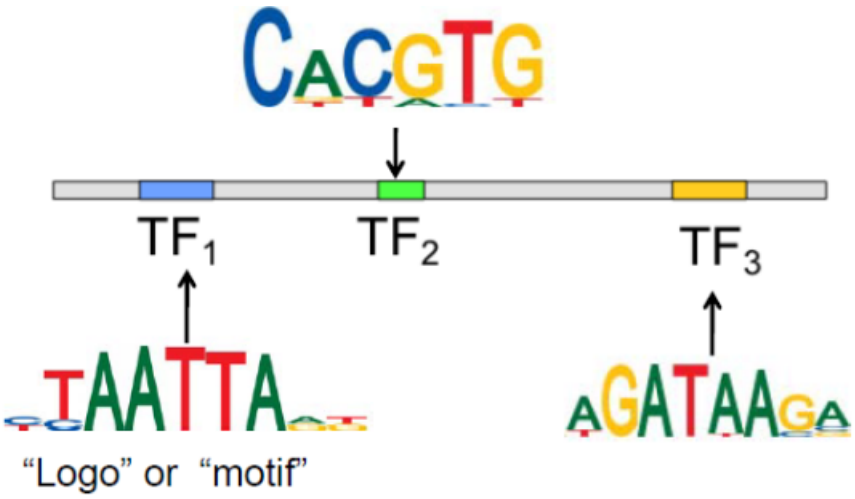
- Binding domain of transcription factors will preferentially recognize specific short DNA sequences based on biophysical constraints
- Preferences will differ between transcription factors



Berger et al, Cell 2008

DNA-binding domain of *Engrailed*

Transcription factors recognize sequence motifs in genome

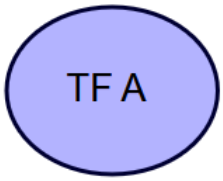


- Binding can activate or repress production of mRNA
- Important for understanding non-coding variants associated with disease

Motif Representation

Motif Instances

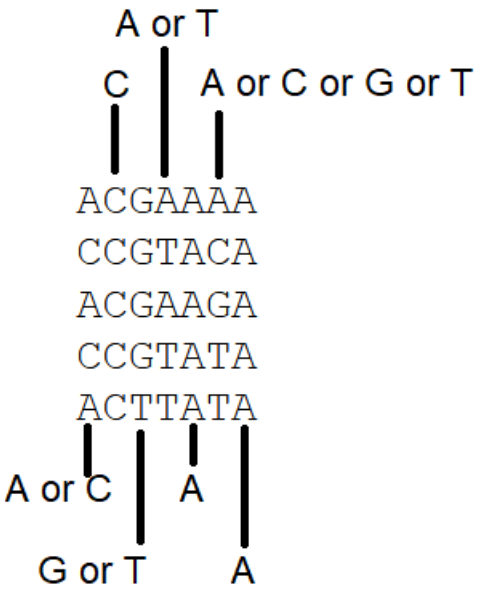
- Given set of known motif instances for a transcription factor (TF), how should we represent such a set as a motif?



ACGAAAA
CCGTACA
ACGAAGA
CCGTATA
ACTTATA

Possible ideas:

- Kmers:
  - Define a motif to be a single consensus k-mer: e.g., ACGTATA
  - Question: What are potential advantages or disadvantages of this approach?
- Kmer neighborhood
  - (k, d) motifs - a kmer and all kmers with at most d mismatches from
  - Same question as above: potential advantages or disadvantages?
- Degenerate sequence codes:



  - Question: How many non-empty combinations of four nucleotides are possible at one position? ($2^4 - 1 = 15$)
  - For this example, we will follow this chart:

Base set	IUPAC nucleotide code
A	A
C	C
G	G
T	T
A or G	R
C or T	Y
G or C	S
A or T	W

Base set	IUPAC nucleotide code
G or T	K
A or C	M
C or G or T	B
A or G or T	D
A or C or T	H
A or C or G	V
A or C or G or T	N

  - We get `MCKWANA` to describe that example sequence

- Positional Weight Matrix (PWM; profile matrix)

1234567
ACGAAAA
CCGTACA
ACGAAGA
CCGTATA
ACTTATA

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

- Question: What assumptions are being made when representing an alignment as a positional weight matrix?
  - Assuming independence between positions. Also fixed spacing.

Motif Scanning

Scoring with a Positional Weight Matrix

For the example above:

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

Scoring agreement of a sequence with the PWM

CCGTATA

$$\frac{2}{5} \times 1 \times \frac{4}{5} \times \frac{3}{5} \times 1 \times \frac{2}{5} \times 1 = \frac{48}{625}$$

- Question: What additional assumption are we implicitly making in the scoring how likely a sequence has a motif?
  - Each nucleotide is a priori equally likely

Background Models

- Probability can be evaluated relative to background model

$$\log \frac{P(\text{sequence} \mid \text{PWM})}{P(\text{sequence} \mid \text{Background})}$$

If we assume a uniform backgroun distribution over nucleotides, then each is assumed to occur with probability 0.25.

CCGTATA $\log(\frac{\frac{48}{625}}{0.25^7}) = 7.14$

If we assume G or C ccur with probability 0.2 and As and Ts with probability 0.3,

$$\log\left(\frac{\frac{48}{625}}{0.2^3 \times 0.3^4}\right) = 7.08$$

For simplicity we will assume a uniform background which will make the denominator the same for all sequences of a fixed length

- We can now use this to score any sequence. Using the PWM from above, consider `ACTTATA`.

$$\frac{3}{5} \times 1 \times \frac{1}{5} \times \frac{3}{5} \times 1 \times \frac{2}{5} \times 1 = \frac{18}{625}$$

- Question: How can we run into problems using this PWM for scoring?
  - Any sequence that has a nucleotide not previously observed in a position will always get a score of 0. Consider `ACTTATC`.

$$\frac{3}{5} \times 1 \times \frac{1}{5} \times \frac{3}{5} \times 1 \times \frac{2}{5} \times 0 = 0$$

PWM based on pseudo-counts

To address the problem, we add one observation for each nucleotide at each position. (We can even add fractional observations or more than one observation).

1234567
ACGAAAA
CCGTACA
ACGAAGA
CCGTATA
ACTTATA
AAAAAAA
CCCCCCC
GGGGGGG
TTTTTTT

	1	2	3	4	5	6	7
A	4/9	1/9	1/9	3/9	6/9	2/9	6/9
C	3/9	6/9	1/9	1/9	1/9	2/9	1/9
G	1/9	1/9	5/9	1/9	1/9	2/9	1/9
T	1/9	1/9	2/9	4/9	1/9	3/9	1/9

Now, the sequence we had before, `ACTTATC` will get a low score instead of zero:

$$\frac{4}{9} \times \frac{6}{9} \times \frac{2}{9} \times \frac{4}{9} \times \frac{6}{9} \times \frac{3}{9} \times \frac{1}{9} = 0.000723$$

What if we had a sequence that is longer?

- We will score each sub-sequence that is length of the PWM and record subsequence above some threshold that depends on the PWM or in some cases the best match.

For example, `ACTTATCGA`

ACTTATCGA
ACTTATCGA = 0.000723
ACTTATCGA = 0.00000753
ACTTATCGA = 0.0000100

	1	2	3	4	5	6	7
A	4/9	1/9	1/9	3/9	6/9	2/9	6/9
C	3/9	6/9	1/9	1/9	1/9	2/9	1/9
G	1/9	1/9	5/9	1/9	1/9	2/9	1/9
T	1/9	1/9	2/9	4/9	1/9	3/9	1/9

Using PWMs for Variant Effect Prediction

Strategy to predict variant effect with PWM

- Score reference and mutated sequence with PWM
- Check if at least one meets a score threshold
- Score change between the two sequences

Suppose the reference sequence is CAT, under the PWM model below how would you rank the three mutations in terms of greatest predicted impact?

	1	2	3
A	1/7	4/7	1/7
C	4/7	1/7	2/7
G	1/7	1/7	2/7
T	1/7	1/7	2/7

CAT: 4/7*4/7*2/7=32/343

CTT: 4/7*1/7*2/7=8/343

CAA: 4/7*4/7*1/7=16/343

CAG: 4/7*4/7*2/7=32/343

Libraries of hundreds of PWMs exist and can be used to compute motif enrichments

- PWM libraries derived from aligned sets of short-curated sequences from small-scale experiments or discovered de novo from high-throughput experiments
- Can be used to compute motif enrichments for a set of sequences of interest and implicate specific transcription factors

De Novo Motif Discovery

Problem: Give a collection of sequences identify motifs *de novo*.

```
Sequence 1   AATCAGTTATCTGTTGTATACCCGGAGTCC
Sequence 2   AGGTCGAATGCAAAACGGTTCTTGCACGTA
Sequence 3   GAGATAACCGCTTGATATGACTCATTGCCA
Sequence 4   ATATTCCGGACGCTGTGACGATCCGGTTGT
Sequence 5   GAACGCAACCAGTTCAGTGCTTATCATGAA
```

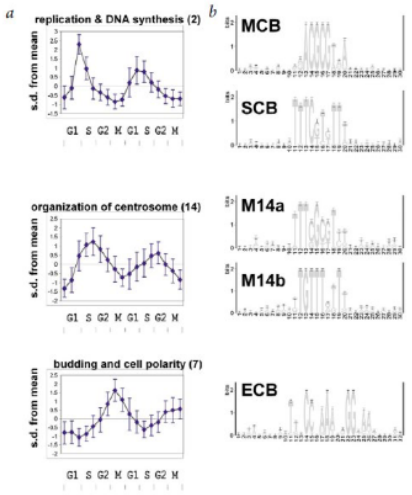
Do you see any shared pattern in the above set of sequences?

```
Sequence 1   AATCAGTTATCTGTTGTATACCCGGAGTCC
Sequence 2   AGGTCGAATGCAAAACGGTTCTTGCACGTA
Sequence 3   GAGATAACCGCTTGATATGACTCATTGCCA
Sequence 4   ATATTCCGGACGCTGTGACGATCCGGTTGT
Sequence 5   GAACGCAACCAGTTCAGTGCTTATCATGAA
```

To look for instances in motifs, we can scan the genome one kmer at a time using the sliding window technique, until we find a kmer that is close enough (use the DP algorithm!).

Examples of sets of sequences for motif discovery

- Promoter regions of co-expressed genes



letter

Systematic determination of genetic network architecture

Saeed Tavazoie[1], Jason D. Hughes[1,2], Michael J. Campbell[3], Raymond J. Cho[4] & George M. Church[1]

nature genetics • volume 22 • july 1999

Applied motif discovery on 600bp upstream of genes in the same k-means clusters

- Locations of TF binding across the genome from a mapping experiment

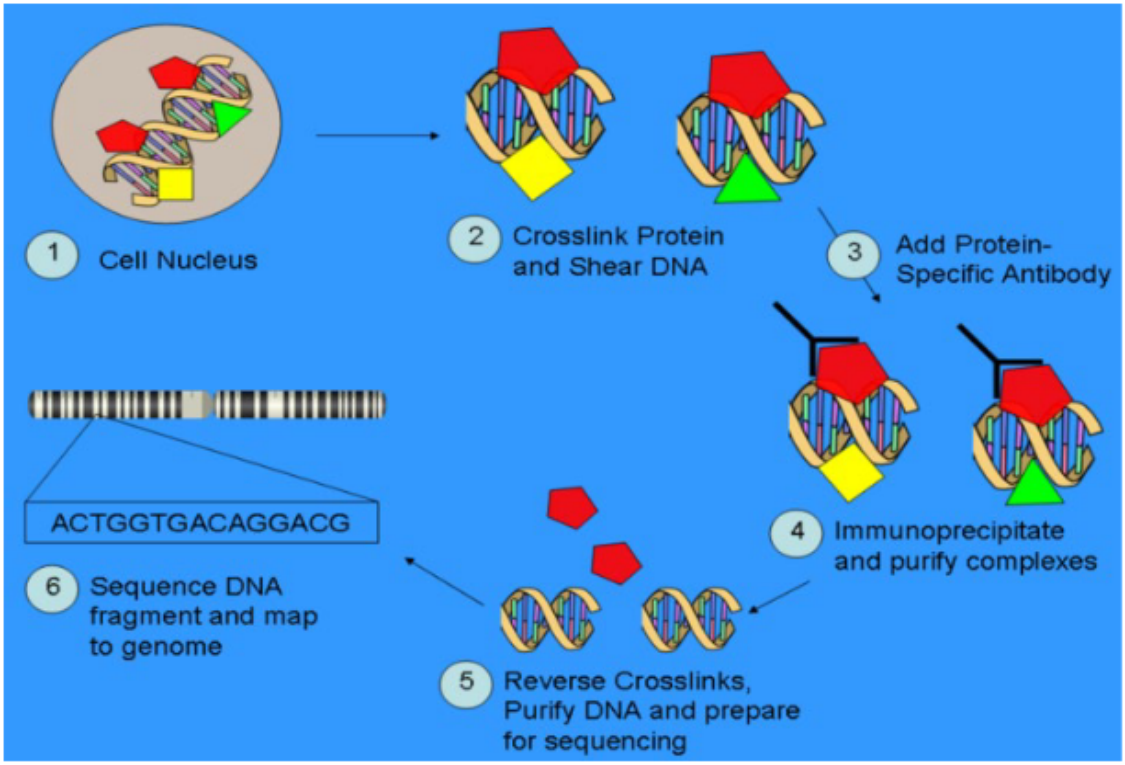
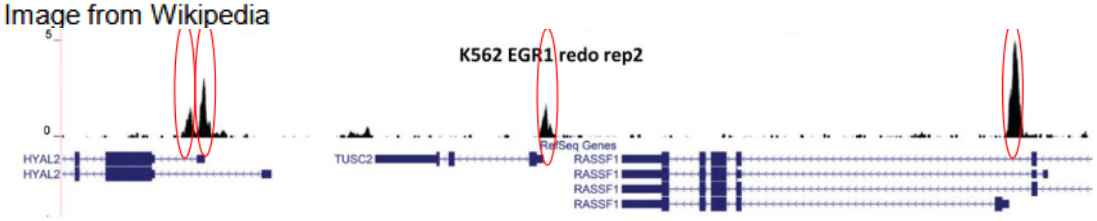
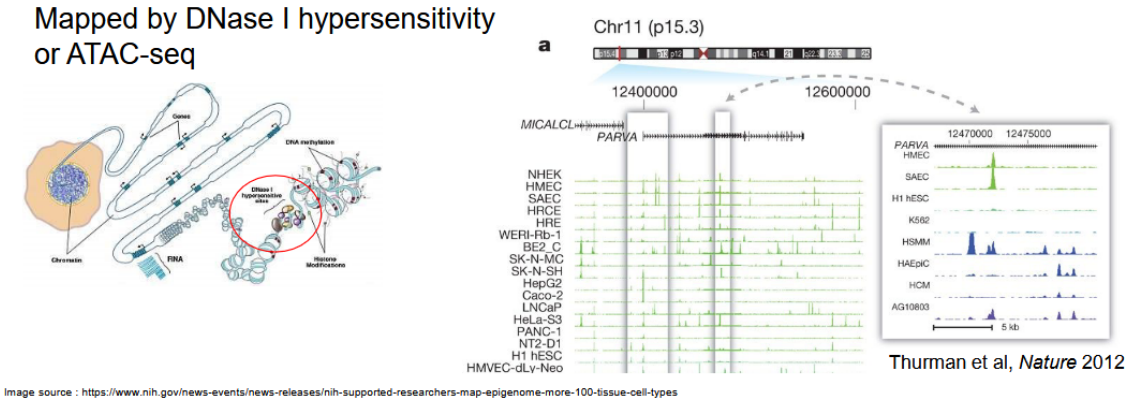


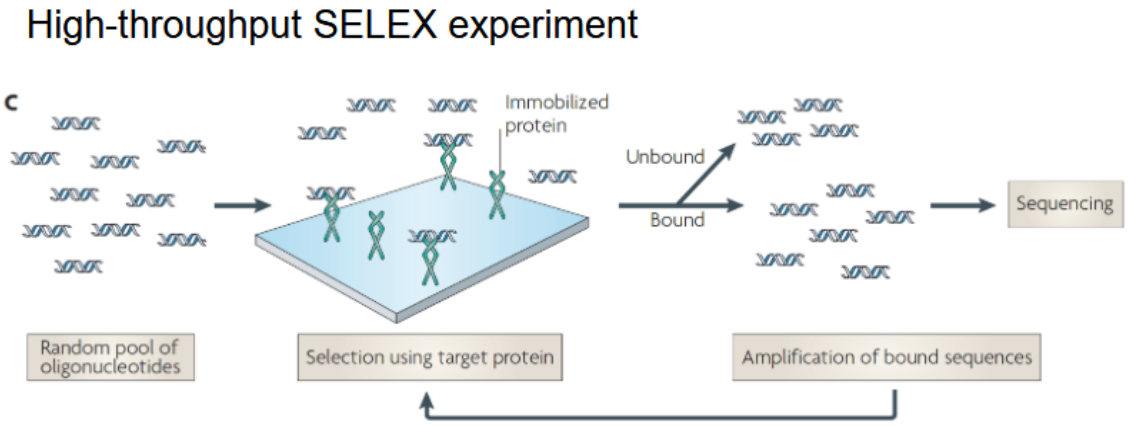
Image from Wikipedia



- Regions across the genome where the DNA is accessible in a cell type from a mapping experiment



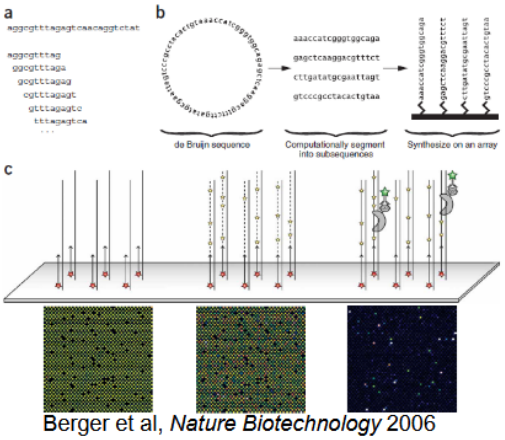
- Experiments designed to measure TF binding specificity



Protein binding microarray

Design properties:

- 44,000 sequences of 35bp
- Sequences designed such all 10mers appear once
- All 8-mers appear 16 times



Formulating the de novo motif discovery problem

- Give an input motif of length k and set of t sequences
- Output a motif instance for each input sequence and a corresponding motif that optimizes some objective function

- Assumption: each input sequence has one instance of the motif
  - This will depend on the motif representation and the scoring function.
  - Also will need a way to optimize the score.

Scoring a set of motif instances

- Will depend on motif representation
- Need to score the selected instance from each sequence and then combine the scores
- Question: If our motif representation was a k-mer string, how could we score motif instances?
  - Hamming distance - number of mismatches. (e.g., d(CAT, TAT) = 1)
- Question: What could our overall optimization function be?
  - Minimize sum across all instances. "Median string problem."
- Question: If our motif representation was a PWM, how could we score motif instances?
  - Can use probabilities derived earlier or log of them
- Question: What could our overall optimization function be?
  - Maximize sum of the log probabilities for selected motif instances

Optimization Problem

- We want to find a motif instance from each sequence and corresponding motif that optimizes our objective.
- E.g., for PWM assuming uniform background:

$$\max_{PWM,w} \sum_{i=1}^{t} \log \left(\sum_{j=1}^{n-k+1} w_{i,j} P(S_{i,j\,:\,(j+k-1)} | PWM) \right)$$

  - t sequences
  - n nucleotides per sequence
  - k is length of motif
  - $S_{i,j\,:\,(j+k-1)}$ is the subsequence in sequence i starting at position j of length k
  - $w_{i,j}$ is 1 if subsequence starting at position j of sequence i contains motif instance, and 0 otherwise
  - Will assume for now $w_{i,j}$ is 1 for exactly one j for each sequence i.

De novo motif discovery - brute force strategies

Idea 1:

Try every possible combination of positions in each sequence.

- Suppose we have t sequences
- n nucleotides per sequence
- k is length of motif
- What is the complexity of this assuming $k << n$? ($O((n-k+1)^t)$)

Idea 2:

Brute force search over the motif representation

- Suppose our motif representation is a kmer sequence with t sequences
- n nucleotides per sequence
- k is length of motif
- What is the complexity of trying and evaluating all k-mers?
  - 4^k possible k-mers. O(nkt) time to evaluate each k-mer.
  - Would require O($4^k nkt$) times
- This compares favorably to the previous approach for large n, because it can independently score each sequence

Now, what if we used a PWM for motif representation?

- t sequences
- n nucleotides per sequence
- k is length of motif

How cam we try to (approximately) optimize this if applying brute force on the PWM representation?

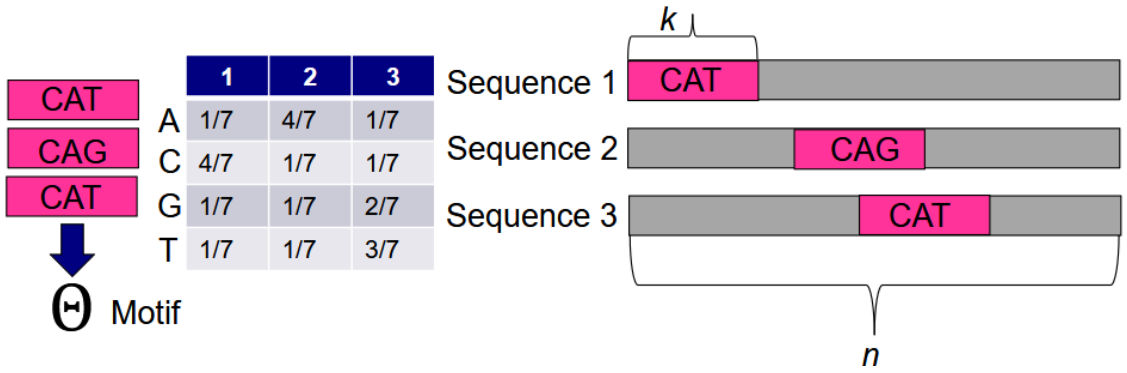
- Discretize entries into d possible values for each entry of PWM. Could cover space of PWMs with $d = t + 1$.
- The complexity of this is now O($d^{3k} nkt$)

De novo motif discovery - Non-brute force strategies

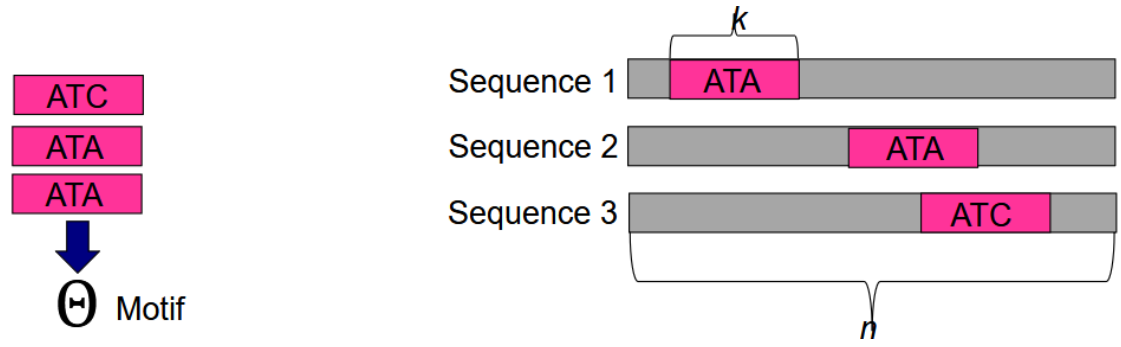
- Greedy Motif Search
- Random Initialization + Iterative Batch Greedy Updates
- Gibbs Sampling
- Expectation-Maximization (EM)

Greedy Approach to Motif Discovery

1. Start by placing a motif instance at first position in first sequence.
2. Build motif based off of it
3. Identify highest scoring motif instance in the next sequence
4. Update motif
5. Repeat for remaining sequence(s), go to step 3.



6. Repeat at the next starting position of sequence 1.



Example: Greedy Approach to Motif Discovery

Sequence 1 CATAG
Sequence 2 ACATT
Sequence 3 GCCAT

	1	2	3
A	1/5	2/5	1/5
C	2/5	1/5	1/5
G	1/5	1/5	1/5
T	1/5	1/5	2/5

Sequence 1 CATAG
Sequence 2 ACATT
Sequence 3 GCCAT

	1	2	3
A	1/5	2/5	1/5
C	2/5	1/5	1/5
G	1/5	1/5	1/5
T	1/5	1/5	2/5

ACA:1/5*1/5*1/5=1/125
CAT:2/5*2/5*2/5=8/125
ATT:1/5*1/5*2/5=2/125

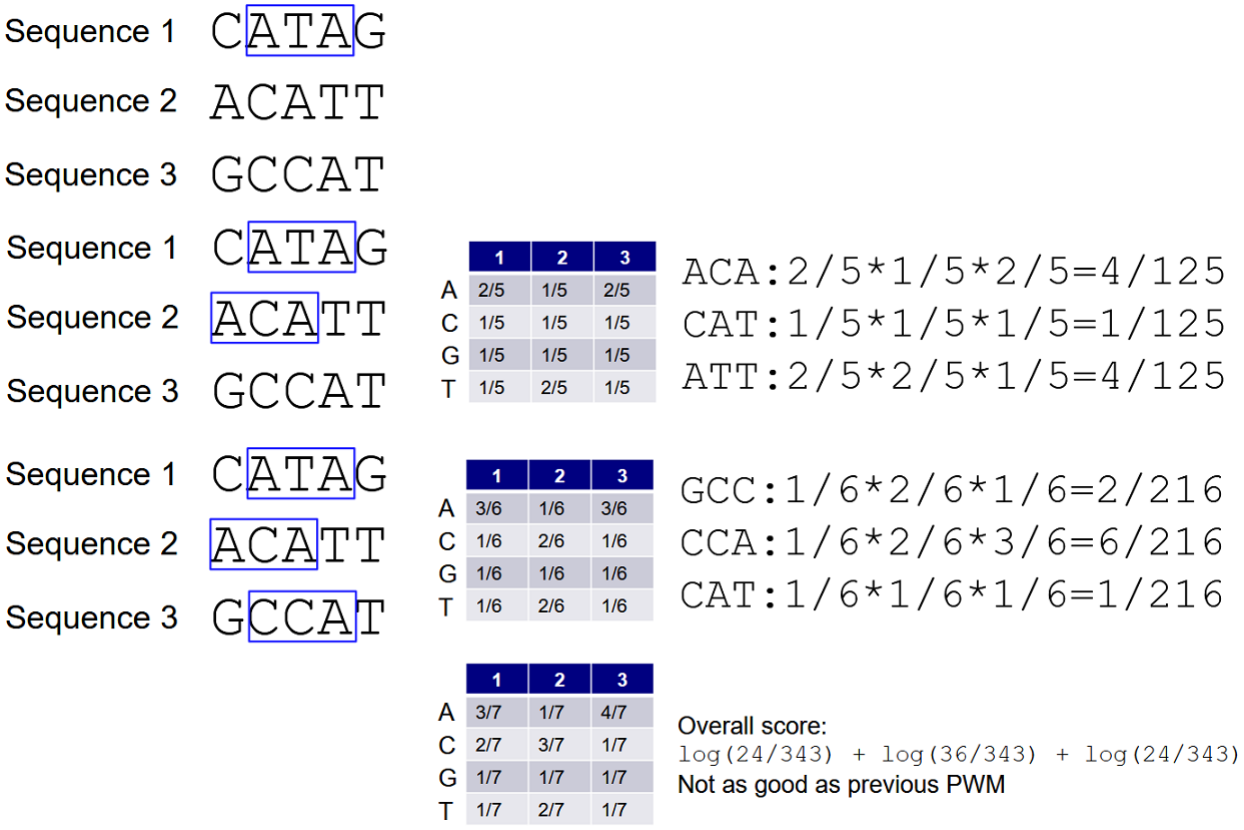
Sequence 1 CATAG
Sequence 2 ACATT
Sequence 3 GCCAT

	1	2	3
A	1/6	3/6	1/6
C	3/6	1/6	1/6
G	1/6	1/6	1/6
T	1/6	1/6	3/6

GCC:1/6*1/6*1/6=1/216
CCA:3/6*1/6*1/6=8/216
CAT:3/6*3/6*3/6=27/216

	1	2	3
A	1/7	4/7	1/7
C	4/7	1/7	1/7
G	1/7	1/7	1/7
T	1/7	1/7	4/7

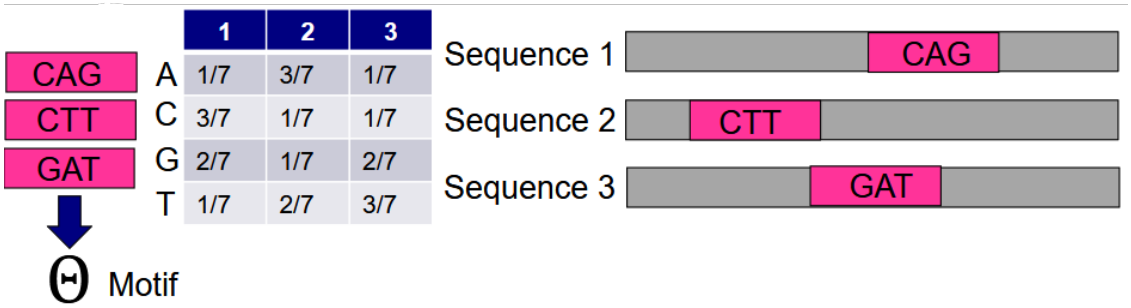
Overall score:
log(64/343) + log(64/343) + log(64/343)



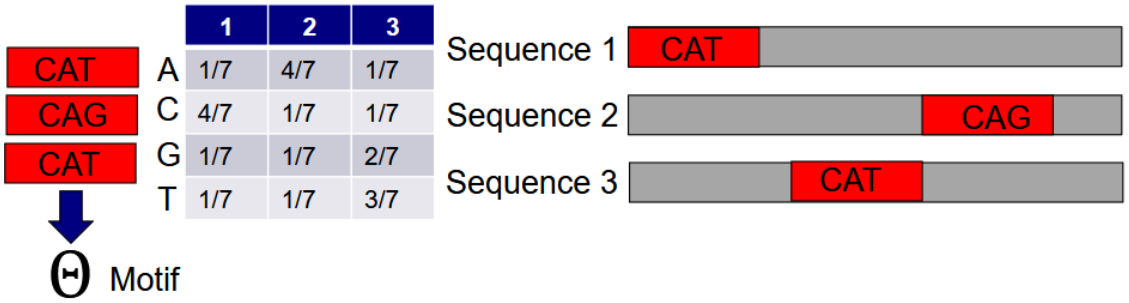
- With this example, starting with the `TAG` in sequence 1 would not lead to better results either.
- In this case, the Greedy Motif Search found the optimal motif. However, is it possible that it *can* miss a better motif?
  - Yes! Consider the case of the first sequence is GGGGG. In this case, there is only one possible solution the greedy algorithm will output.
- What problems are there with the Greedy method?
  - Highly sensitive to initial sequence
  - May only explore a small portion of the search space and easily miss the optimal solution.

Random Initialization and Iterative Batch Greedy Updates

1. Start by placing a motif instance randomly for each sequence
2. Create a motif matrix



3. Update motif instances to be highest score based on the current motif
4. Update motif based on current motif instances



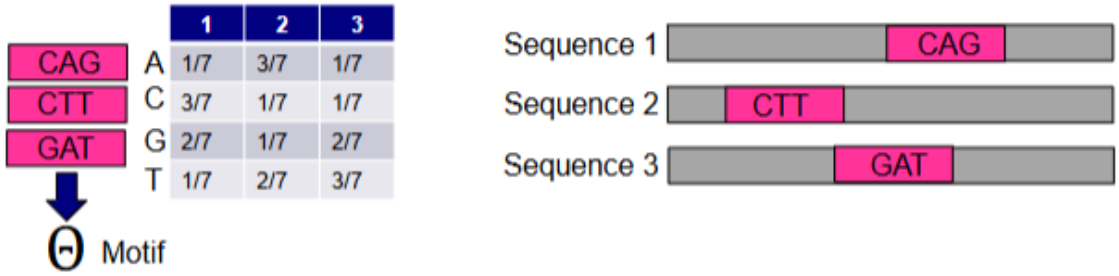
5. Iterate until convergence
6. Repeat for multiple different initializations.

This gives better results than the greedy algorithm, but still is not perfect.

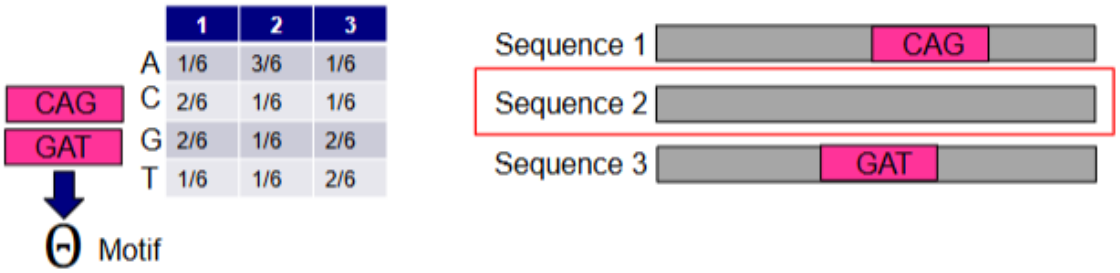
- The algorithm may converge at a local minimum instead of the global minimum.

Gibbs Sampling

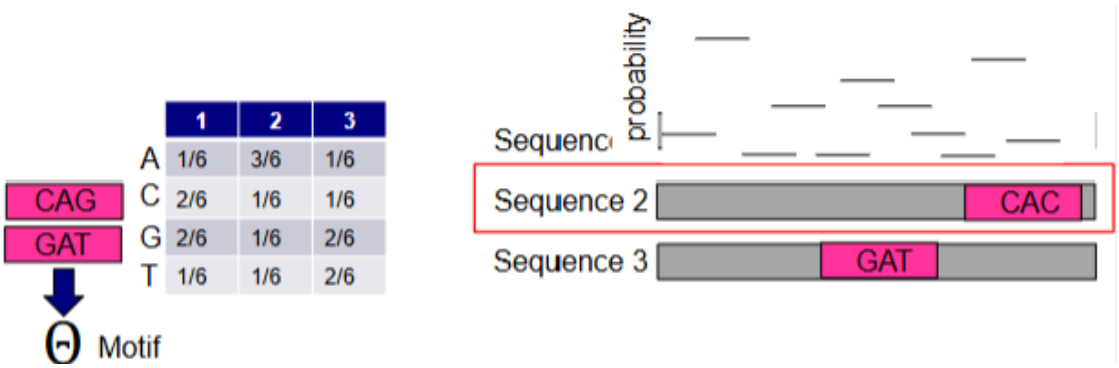
1. Start by placing a motif instance randomly for each sequence and create the initial motif



2. Select a sequence at random to exclude
3. Update the motif

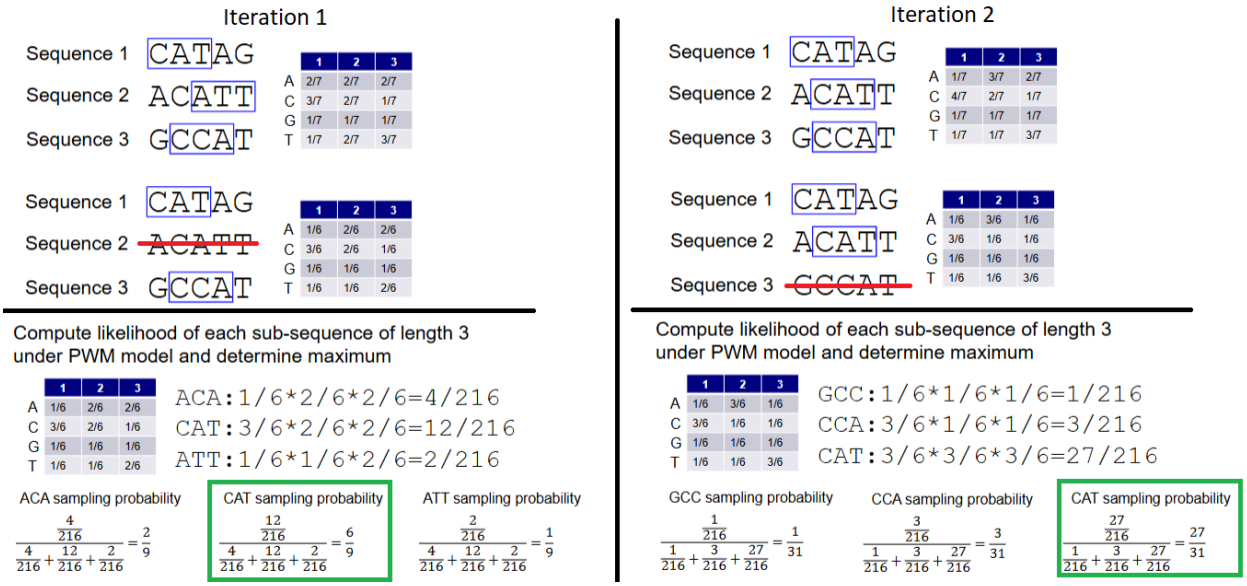


4. Compute a probability of each position in the selected sequence containing a motif instance based on the motif
5. Sample a new motif instance based on the probabilities and update motif



6. Repeat (from step 2) until termination.

   - Termination is usually a specified number of iterations
   - Other strategies include e.g., limited change in motif parameters or limited change in objective function over sufficiently large number of iterations.
   - This can be repeated from multiple random initializations



Gibbs Sampling in General

- Often difficult to sample from a distribution $\mathrm{P}(U_1, \ldots, U_n)$ directly but feasible to sample $\mathrm{P}(U_j | U_1, \ldots, U_{j-1}, U_{j+1}, \ldots U_n$ for each j
  - In the motif example, easier to compute distribution of the motif instance in one sequence if we know where the instances are in all the other sequences than a joint distributions
- By iteratively sampling the distribution of variables conditioned on the others can eventually approximate joint distribution,

EM Algorithm Optimization Problem

- We want to find a motif instance from each sequence and corresponding motif that optimizes our objective. e.g. for PWM assuming uniform background,

$$\max_{PWM,w} \sum_{i=1}^{t} \log \left(\sum_{j=1}^{n-k+1} w_{i,j} P(S_{i,j\,:\,(j+k-1)|PWM}) \right)$$

- t sequences
- n nucleotides per sequence
- k is length of motif
- $S_{i,j\,:\,(j+k-1)}$ is the subsequence i starting at position j of length k. Relax the assumption that $w_{i,j}$ need to be 0 or 1 corresponding to hard assignments and instead allow "soft-assignments" where we assume $w_{i,j}$ are between 0 and 1 and for each sequence i the sum of the $w_{i,j}$ values are 1.

1. Initialize a motif matrix. One way is by placing a motif instance randomly for each sequence and creating a motif matrix based on them.
2. E-step. Compute a probability of each position in each sequence containing a motif instance based on the motif
3. M-step. Update the motif matrix based on soft assignment probabilities by taking a weighted average.
4. Iterate until termination (number of iterations or convergence criteria, e.g., likelihood)

See earlier section (Week 6) for an example on EM.