Motif Finding

CMSC 423

Motivation

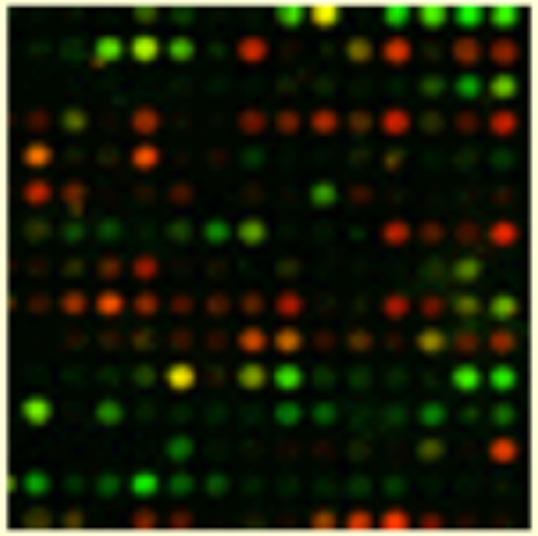
MicroArray analysis of whole genome gene expression

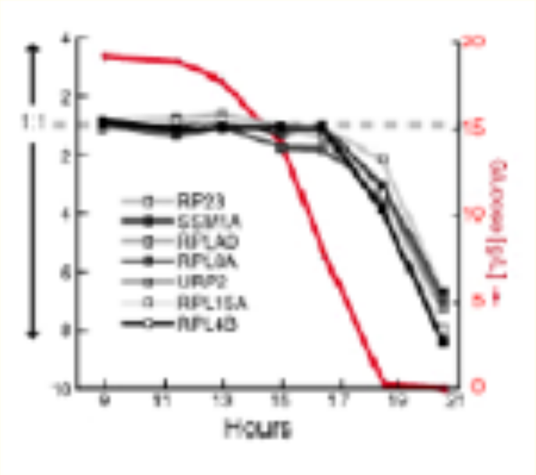
↓

Clustering of genes based on their expression pattern

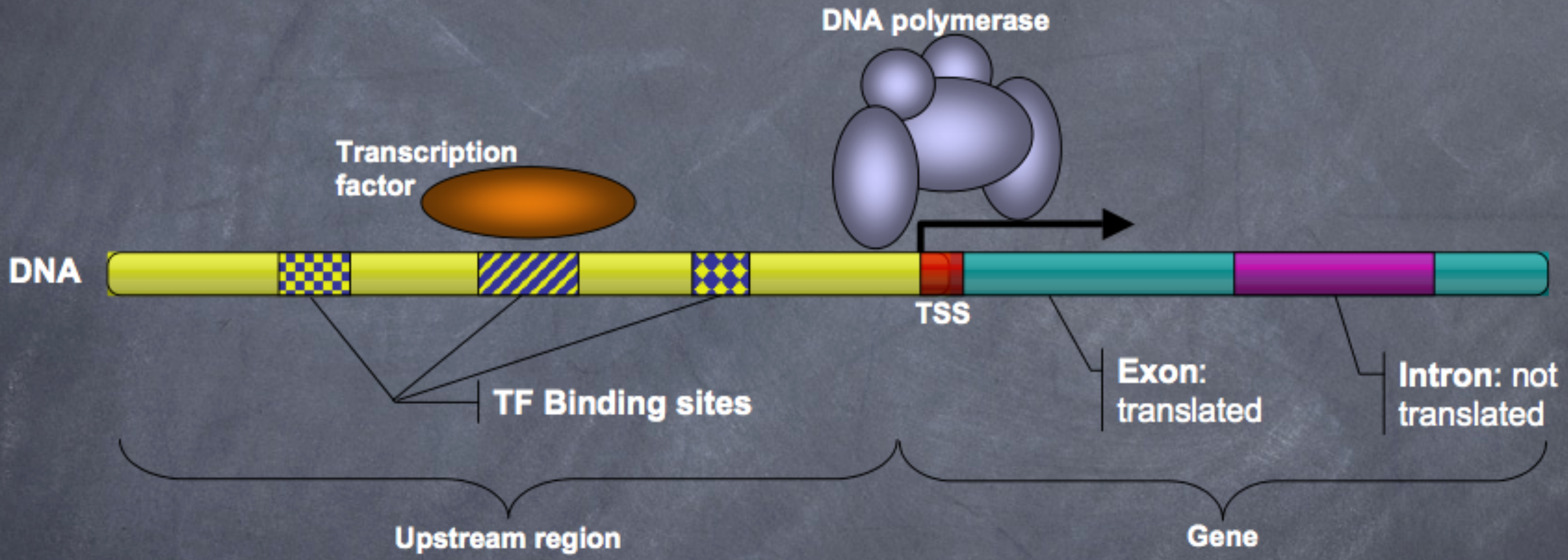
↓

Searching for conserved sequence motifs regulating the expression





DNA -> mRNA -> Protein

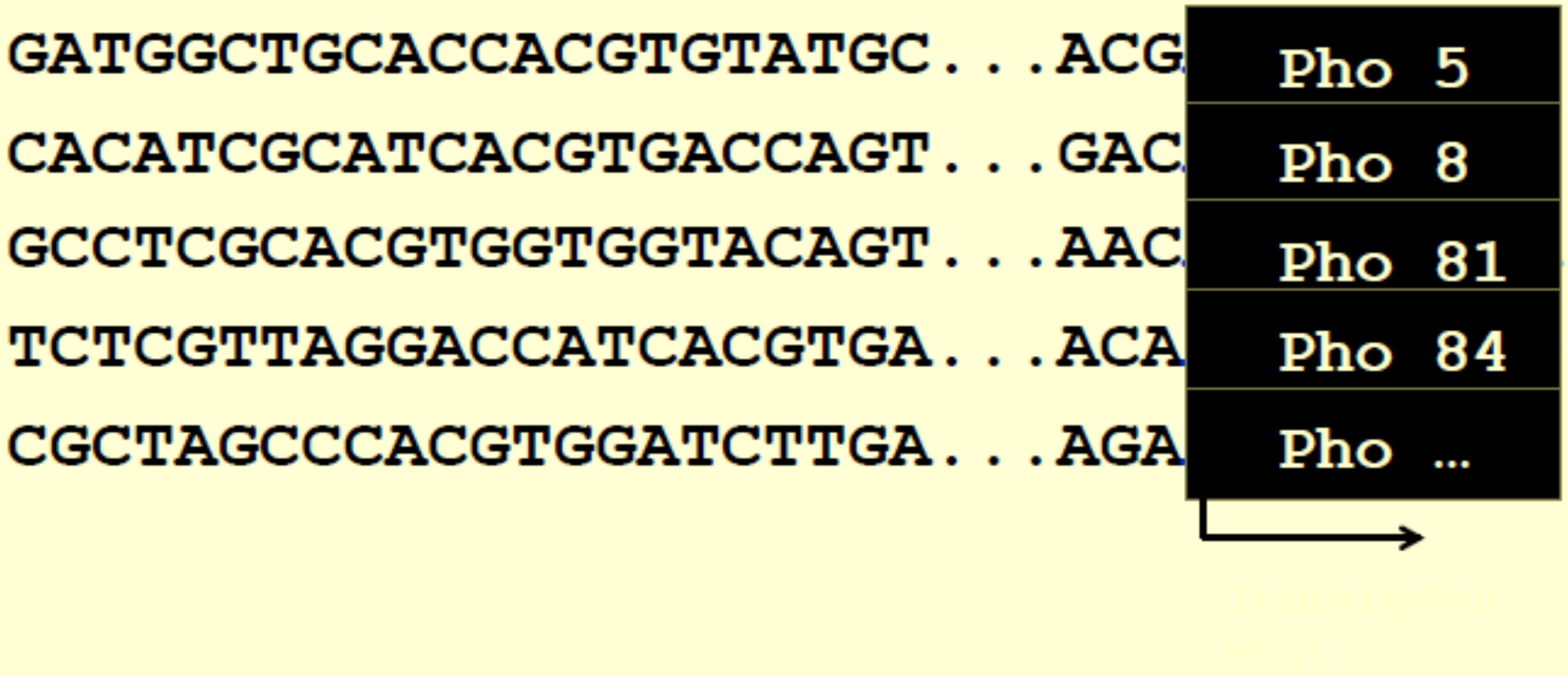


- Finding transcription factor binding sites can tell us about the cell's regulatory network.

Finding Transcription Factor Binding Sites

Upstream Regions

Co-expressed Genes



GATGGCTGCACCACGTGTATGC...ACG

CACATCGCATCACGTGACCAGT...GAC

GCCTCGCACGTGGTGGTACAGT...AAC

TCTCGTTAGGACCATCACGTGA...ACA

CGCTAGCCCACGTGGATCTTGA...AGA

Pho 5

Pho 8

Pho 81

Pho 84

Pho ...

Transcription Start

Finding Transcription Factor Binding Sites

Upstream Regions **Co-expressed Genes**

GATGGCTGCACCACGTGTATGC...ACGATGTCTCGC

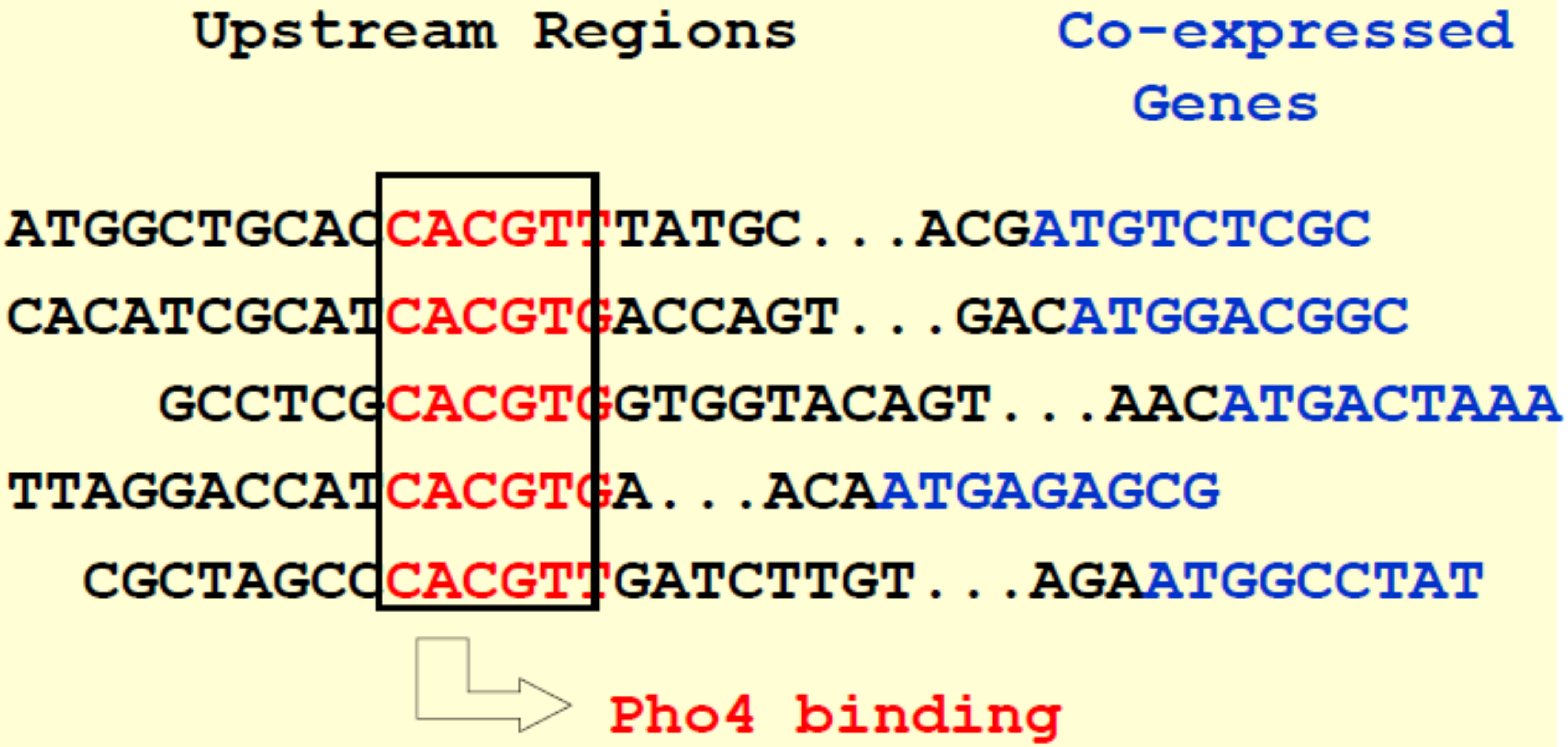
CACATCGCATCACGTGACCAGT...GACATGGACGGC

GCCTCGCACGTGGTGGTACAGT...AACATGACTAAA

TCTCGTTAGGACCATCACGTGA...ACAATGAGAGCG

CGCTAGCCCACGTGGATCTTGT...AGAATGGCCTAT

Finding Transcription Factor Binding Sites



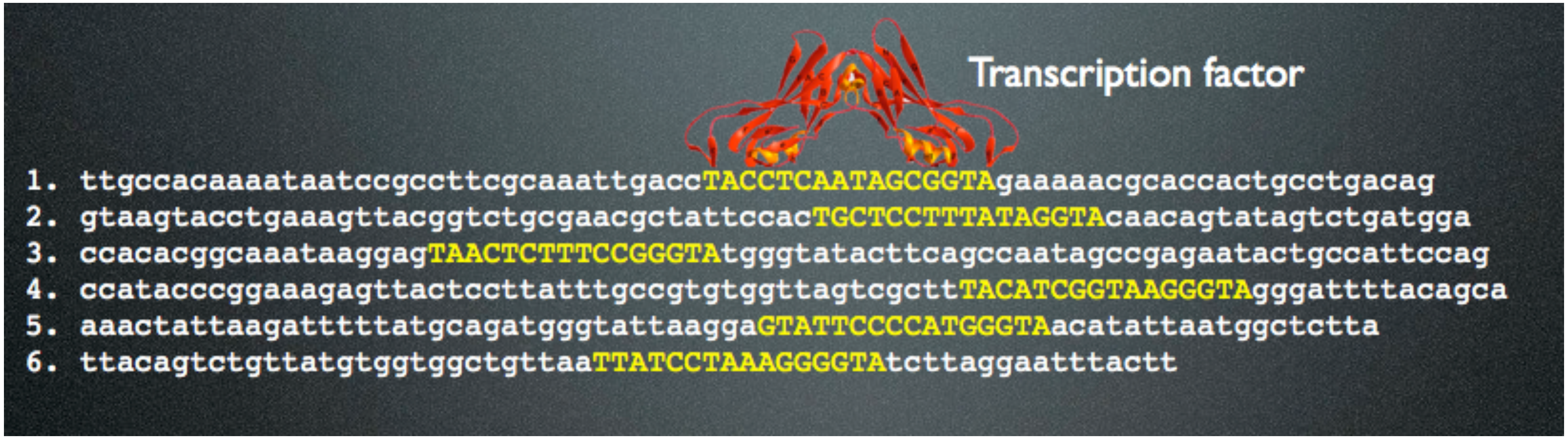
Upstream Regions

Co-expressed Genes

```
ATGGCTGCACCACGTTTATGC...ACGATGTCTCGC
CACATCGCATCACGTGACCAGT...GACATGGACGGC
    GCCTCGCACGTGGTGGTACAGT...AACATGACTAAA
TTAGGACCATCACGTGA...ACAATGAGAGCG
  CGCTAGCCCACGTTGATCTTGT...AGAATGGCCTAT
```

Pho4 binding

Motif Finding



```
1. ttgccacaaaataatccgccttcgcaaattgaccTACCTCAATAGCGGTAgaaaaacgcaccactgcctgacag
2. gtaagtacctgaaagttacggtctgcgaacgctattccacTGCTCCTTTATAGGTAcaacagtatagtctgatgga
3. ccacacggcaaataaggagTAACTCTTTCCGGGTAtgggtatacttcagccaatagccgagaatactgccattccag
4. ccatacccggaaagagttactccttatttgccgtgtggttagtcgcttTACATCGGTAAGGGTAgggattttacagca
5. aaactattaagatttttatgcagatgggtattaaggaGTATTCCCCATGGGTAacatattaatggctctta
6. ttacagtctgttatgtggtggctgttaaTTATCCTAAAGGGGTAtcttaggaatttactt
```

Given t sequences of length n, find most mutually similar set of k-mers (one from each)

Problem: We don't know what the correct motif is!

Example:
10-mer with at most 4 mismatches

Correct 10-mer is AAAAAGGGGG

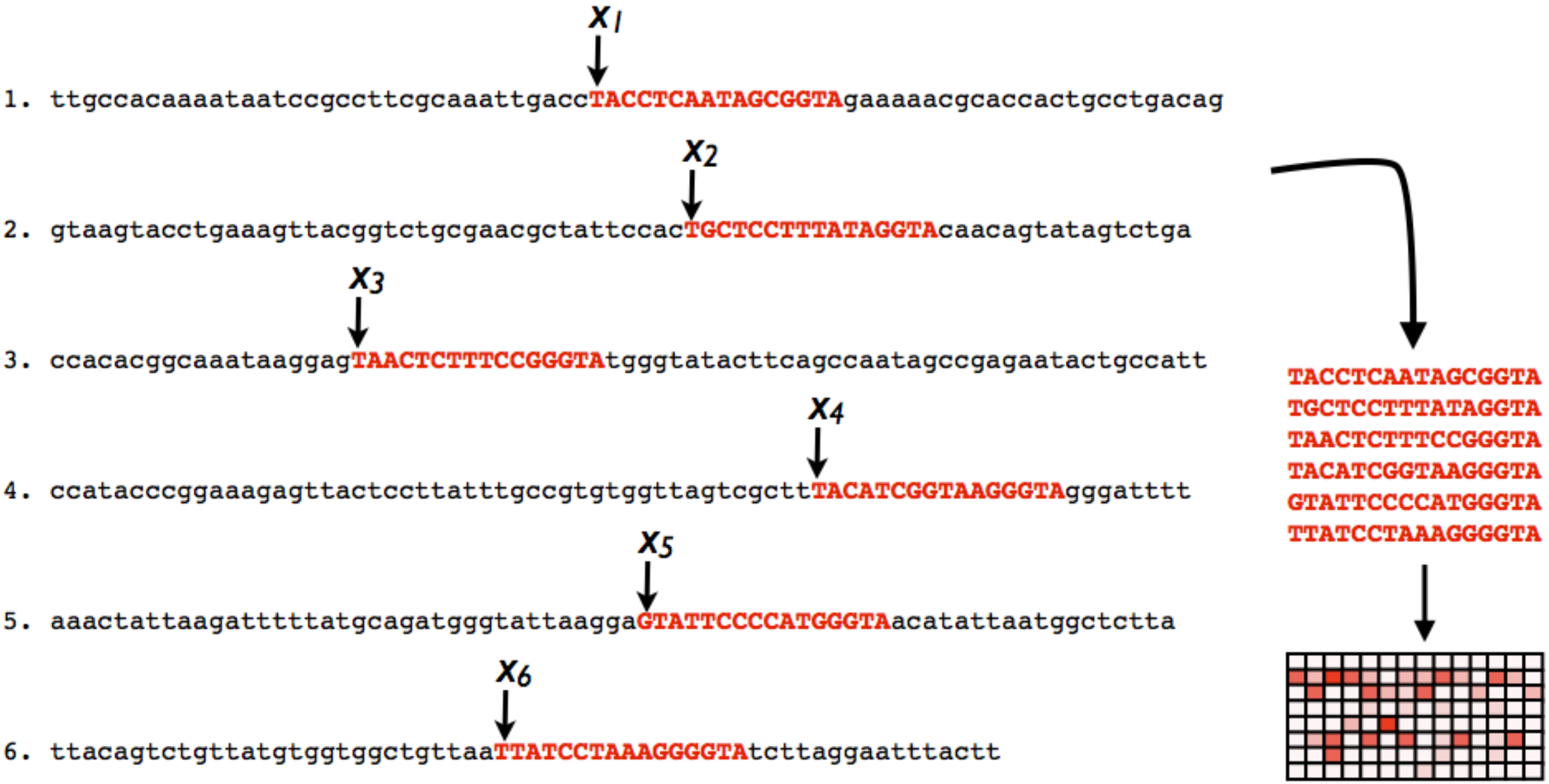
AgAAtGaGGc
cAAtAGGatG

10-mers have 8 mismatches from each other although they have only 4 mismatches between the correct motif

Solution: Scoring Motifs

- Scoring the individual instance of motifs depending on how similar they are to the "ideal" motif.
- BUT WE DON'T KNOW THE IDEAL MOTIF!
- **Solution:** select motifs from each string and score them depending on how similar they are to each other.

If we knew the starting point of the motif in each sequence, we could construct a Sequence Profile (PSSM) for the motif:



1. ttgccacaaaataatccgccttcgcaaattgaccTACCTCAATAGCGGTAgaaaaacgcaccactgcctgacag

2. gtaagtacctgaaagttacggtctgcgaacgctattccacTGCTCCTTTATAGGTAcaacagtatagtctga

3. ccacacggcaaataaggagTAACTCTTTCCGGGTAtgggtatacttcagccaatagccgagaatactgccatt

4. ccatacccggaaagagttactccttatttgccgtgtggttagtcgcttTACATCGGTAAGGGTAgggatttt

5. aaactattaagatttttatgcagatgggtattaaggaGTATTCCCCATGGGTAacatattaatggctctta

6. ttacagtctgttatgtggtggctgttaaTTATCCTAAAGGGGTAtcttaggaatttactt

TACCTCAATAGCGGTA
TGCTCCTTTATAGGTA
TAACTCTTTCCGGGTA
TACATCGGTAAGGGTA
GTATTCCCCATGGGTA
TTATCCTAAAGGGGTA

Motifs

T	C	G	G	G	G	a	T	T	T	t	t
c	C	G	G	t	G	A	c	T	T	a	C
a	C	G	G	G	G	A	T	T	T	t	C
T	t	G	G	G	G	A	c	T	T	t	t
a	a	G	G	G	G	A	c	T	T	C	C
T	t	G	G	G	G	A	c	T	T	C	C
T	C	G	G	G	G	A	T	T	c	a	t
T	C	G	G	G	G	A	T	T	c	C	t
T	a	G	G	G	G	A	a	c	T	a	C
T	C	G	G	G	t	A	T	a	a	C	C

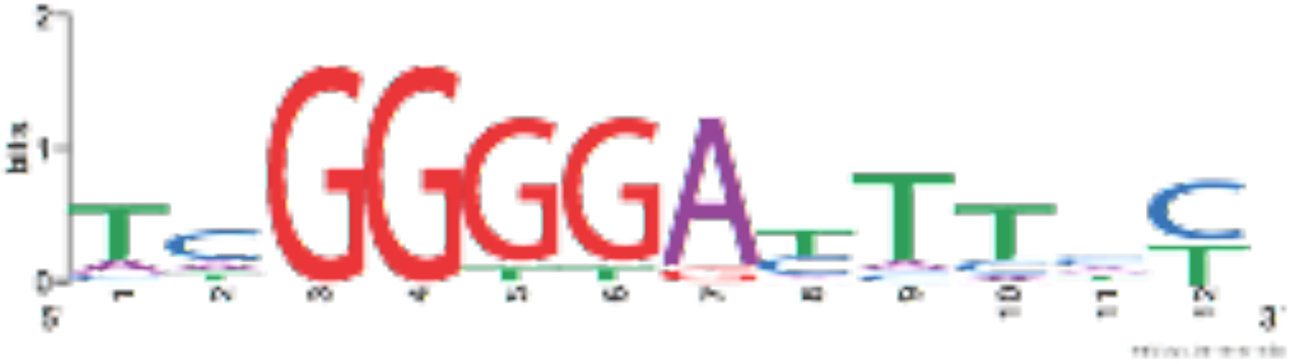
Motifs	T	C	G	G	G	G	a	T	T	T	t	t
	c	C	G	G	t	G	A	c	T	T	a	C
	a	C	G	G	G	G	A	T	T	T	t	C
	T	t	G	G	G	G	A	c	T	T	t	t
	a	a	G	G	G	G	A	c	T	T	C	C
	T	t	G	G	G	G	A	c	T	T	C	C
	T	C	G	G	G	G	A	T	T	c	a	t
	T	C	G	G	G	G	A	T	T	c	C	t
	T	a	G	G	G	G	A	a	c	T	a	C
	T	C	G	G	G	t	A	T	a	a	C	C

Score(*Motifs*) 3 + 4 + 0 + 0 + 1 + 1 + 1 + 5 + 2 + 3 + 6 + 4 = 30

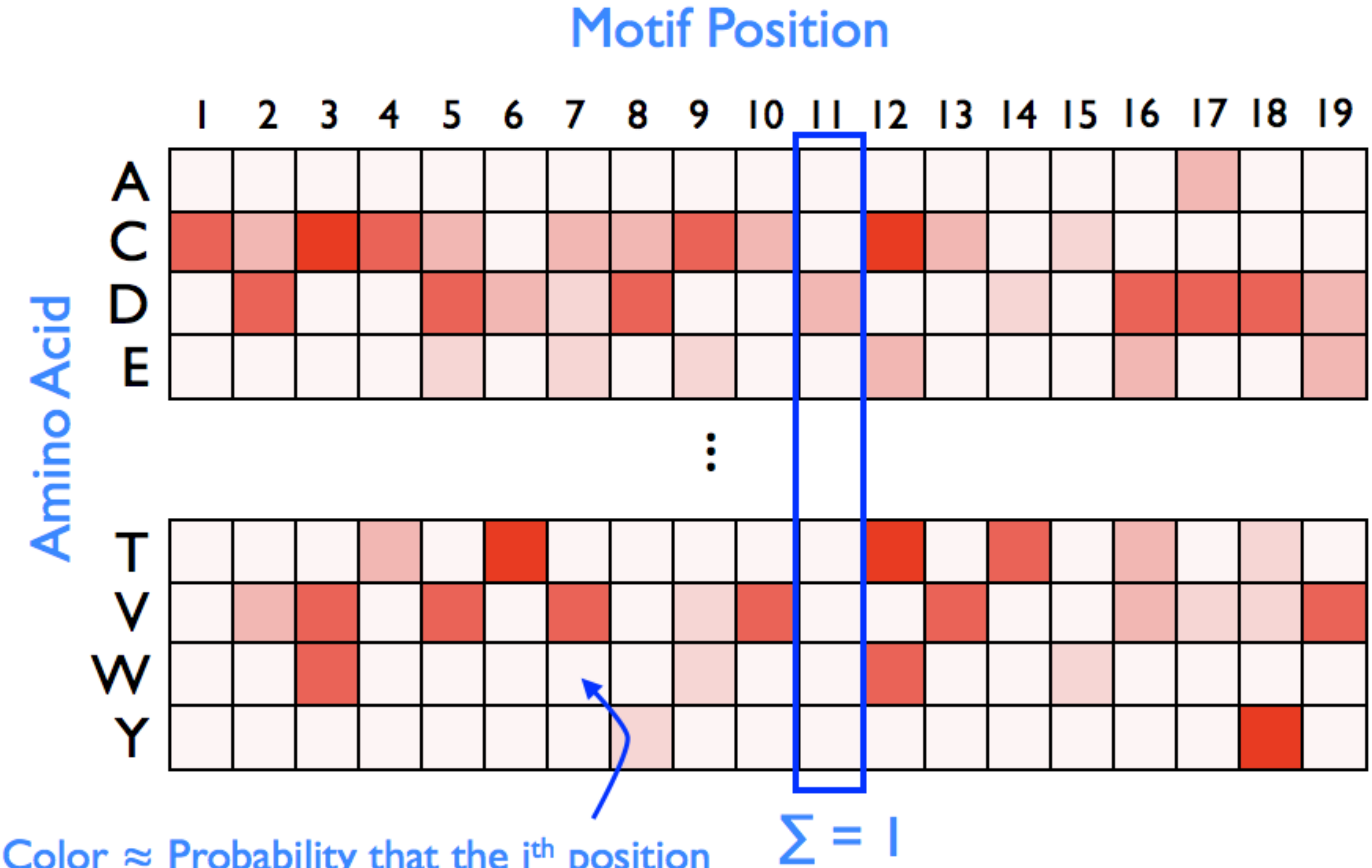
Count(*Motifs*)	A:	2	2	0	0	0	0	9	1	1	1	3	0
	C:	1	6	0	0	0	0	0	4	1	2	4	6
	G:	0	0	10	10	9	9	1	0	0	0	0	0
	T:	7	2	0	0	1	1	0	5	8	7	3	4

Profile(*Motifs*)	A:	.2	.2	0	0	0	0	.9	.1	.1	.1	.3	0
	C:	.1	.6	0	0	0	0	0	.4	.1	.2	.4	.6
	G:	0	0	1	1	.9	.9	.1	0	0	0	0	0
	T:	.7	.2	0	0	.1	.1	0	.5	.8	.7	.3	.4

Consensus(*Motifs*)	T	C	G	G	G	G	A	T	T	T	C	C



Sequence Profiles (PSSM)

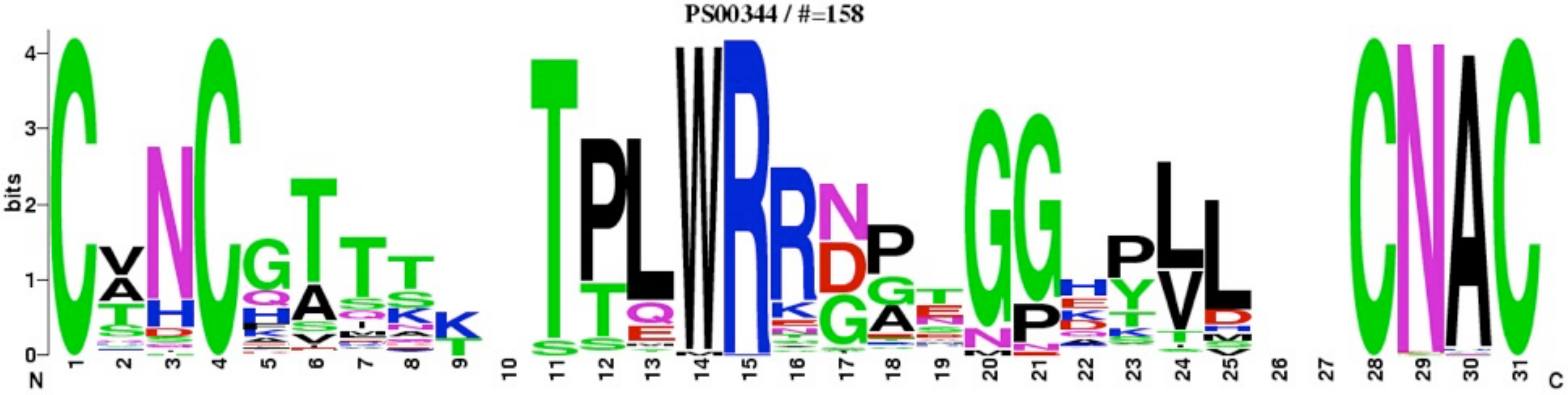


Color ≈ Probability that the i^{th} position has the given amino acid = $e_i(x)$.

Sequence Logos

Height of letter ≈ fraction of time that letter is observed at that position.

(Height of all the letters in a column ≈ to how conserved the column is)



Motif Position

Motif Finding Problem:

Given a collection of string, find the set of k-mers, one from each string that minimizes score of the resulting motif.

INPUT: A collection of strings Dna and integer k.
OUTPUT: A collection motifs of k-mers, one from each string in the Dna, minimizing SCORE(Motifs) among all possible choices of k-mers

What is most simple solution this problem?

Brute Force algorithm

BruteForceMotifSearch(Dna,k)

consider each possible choice of k-mers Motifs from Dna and return the collection Motifs having minimum score.

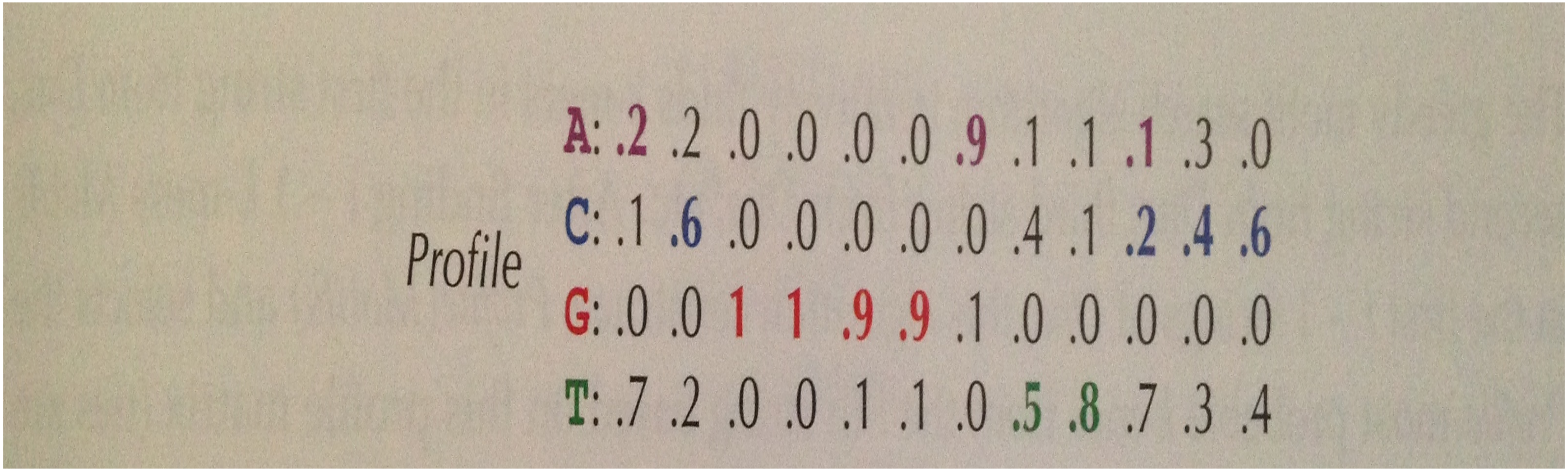
What is the complexity of this algorithm?

Same Motif Finding Problem:

Given a collection of string, find the set of k-mers, one from each string that minimizes score of the resulting motif.

INPUT: A collection of strings Dna and integer k.

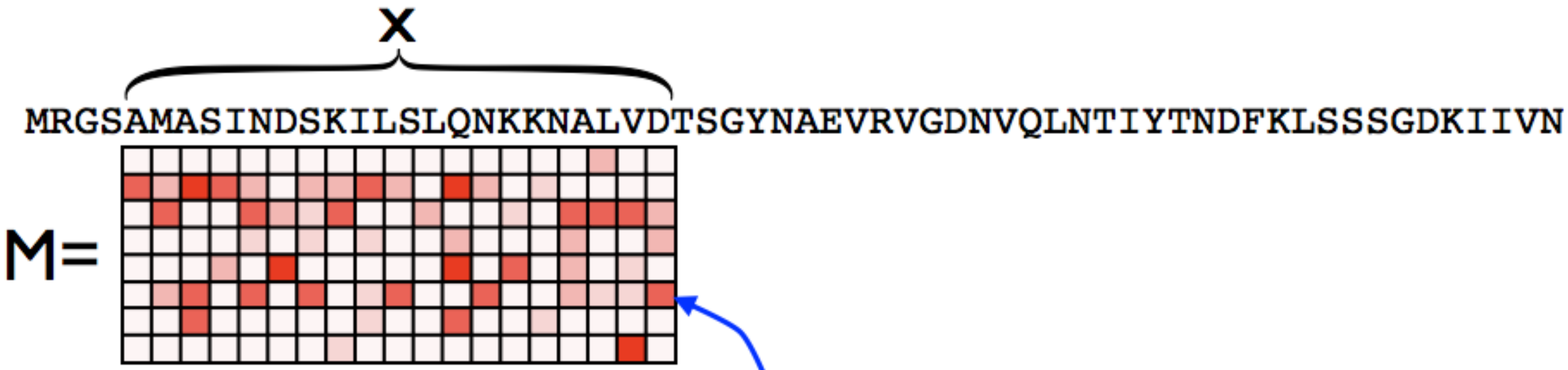
OUTPUT: The starting position vector [x1,..,xt] which minimizes SCORE([x1,..,xt]) over all possible vectors.



Profile	A:	.2	.2	.0	.0	.0	.0	.9	.1	.1	.1	.3	.0
	C:	.1	.6	.0	.0	.0	.0	.0	.4	.1	.2	.4	.6
	G:	.0	.0	1	1	.9	.9	.1	.0	.0	.0	.0	.0
	T:	.7	.2	.0	.0	.1	.1	.0	.5	.8	.7	.3	.4

What is probability Pr(ACGGGGATTACC|Profile)?

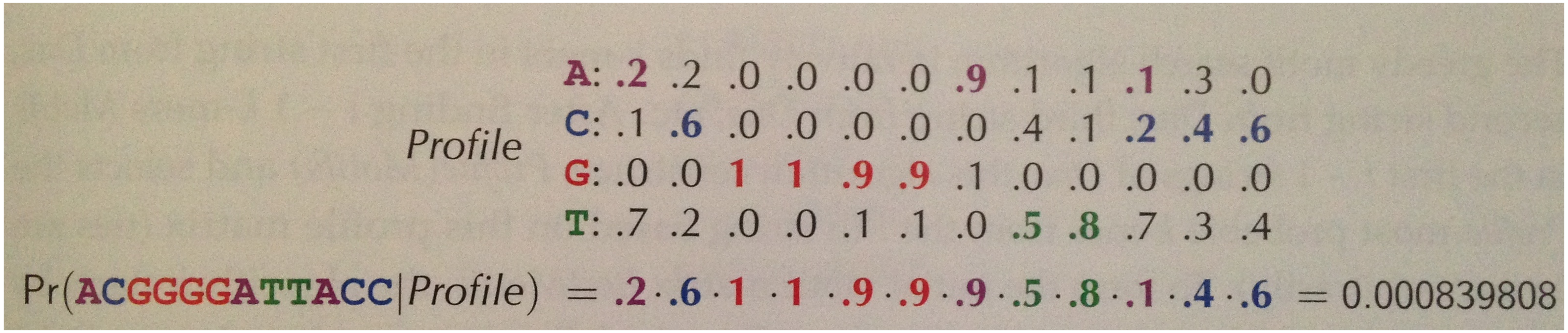
Scoring a Sequence


x
MRGSAMASINDSKILSLQNKKNALVDTSGYNAEVRVGDNVQLNTIYTNDFKLSSSGDKIIVN
M=

Color ≈ Probability that the ith position has the given amino acid = $e_i(x)$.

$$\text{Score}(x) = \Pr(x \mid M) = \prod_{i=1}^{L} e_i(x_i)$$

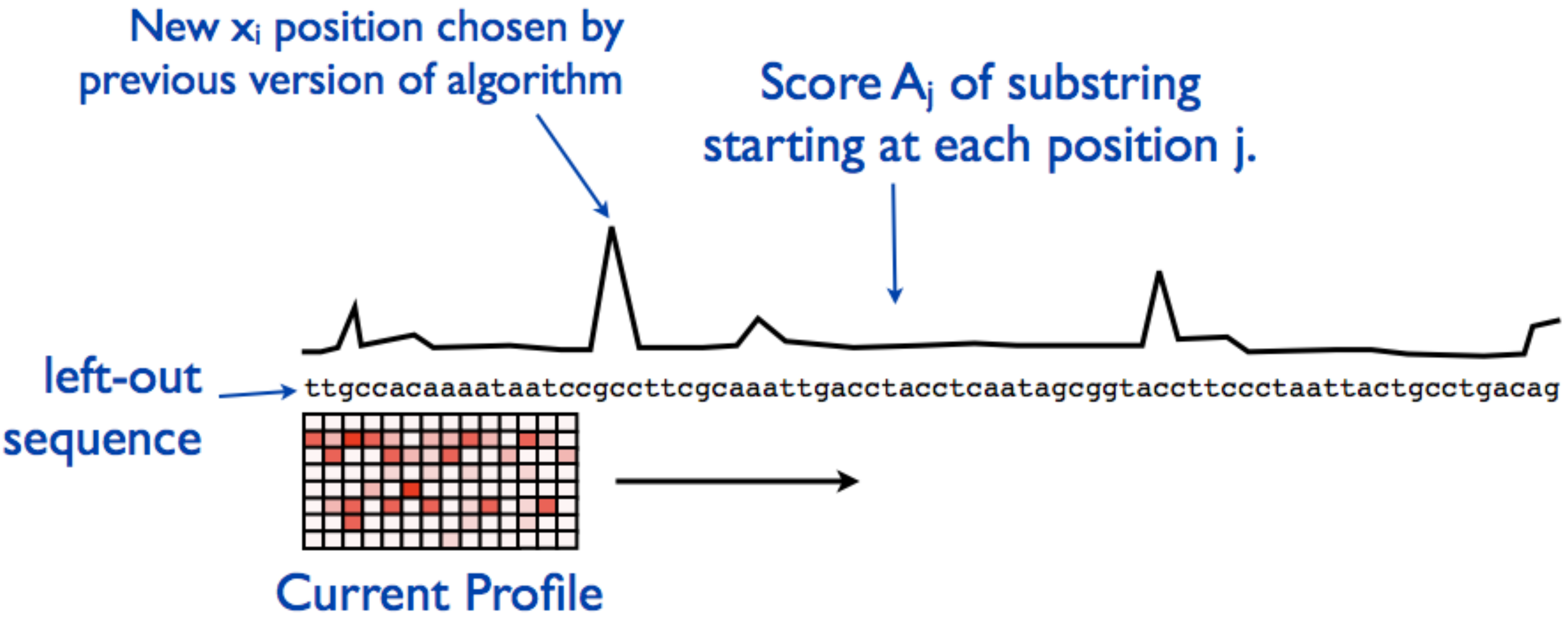
Score of a string according to profile M = Product of the probabilities you would observe the given letters.



Profile	**A:**	**.2**	.2	.0	.0	.0	.0	**.9**	.1	.1	**.1**	.3	.0
	C:	.1	**.6**	.0	.0	.0	.0	.0	.4	.1	**.2**	**.4**	**.6**
	G:	.0	.0	**1**	**1**	**.9**	**.9**	.1	.0	.0	.0	.0	.0
	T:	.7	.2	.0	.0	.1	.1	.0	**.5**	**.8**	.7	.3	.4

$$\Pr(\texttt{ACGGGGATTACC}\,|\,\textit{Profile}) = .2\cdot.6\cdot1\cdot1\cdot.9\cdot.9\cdot.9\cdot.5\cdot.8\cdot.1\cdot.4\cdot.6 = 0.000839808$$

Profile Probability Distribution



New x_i position chosen by previous version of algorithm

Score A_j of substring starting at each position j.

left-out sequence

ttgccacaaaataatccgccttcgcaaattgacctacctcaatagcggtaccttccctaattactgcctgacag

Current Profile

Instead of choosing the position with the best match, choose a position randomly such that:

$$\text{Probability of choosing position } j = \frac{A_j}{\sum_i A_i}$$

(Lawrence, et al., *Science*, 1994)

Profile-most Probable k-mer Problem:

Given a profile matrix, we can evaluate the probability of every k-mer in a string Text and find a Profile-most probable k-mer in text.

Input: A string text, an integer k and 4xk profile matrix.

Output: A Profile-most probable k-mer in text.

Proposed Greedy Algorithm

The basic idea of the greedy motif search algorithm is to find the set of motifs across a number of DNA sequences that match each other most closely.

To do this we:

- Run through each possible k-mer in our first dna string
- Identify the best matches for this initial k-mer within each of the following dna strings thus creating a set of motifs at each step
- Score each set of motifs to find and return the best scoring set.

GREEDYMOTIFSEARCH(Dna, k, t)

```
  BestMotifs ← motif matrix formed by first k-mers in each string from
                 Dna
  for each k-mer Motif in the first string from Dna
     Motif1 ← Motif
     for i = 2 to t
        form Profile from motifs Motif1, …, Motifi - 1
        Motifi ← Profile-most probable k-mer in the i-th string
                 in Dna
     Motifs ← (Motif1, …, Motift)
     if Score(Motifs) < Score(BestMotifs)
        BestMotifs ← Motifs
  output BestMotifs
```

Example

GreedyMotifSearch Example

What if we want the index of best Motifs?

```
GREEDYMOTIFSEARCH(Dna, k, t)
   BestVector ← [1,1,.......,1]
  for j= 1 ..............., n-k+1
    CurrentVector← [j]
    for i = 2 to t
      form Profile from CurrentVector
      u ← Position of Profile-most probable k-mer in the i-th string in Dna
      CurrentVector ← CurrentVector + [u]
    if Score(CurrentVector) < Score(BestVector )
      BestVector ← CurrentVector
  output BestVector
```

Problem with GREEDYMOTIFSEARCH

Motifs:

T	A	A	A
G	T	C	T
A	C	T	A
A	G	G	T

Count(Motifs):

A:	2	1	1	2
C:	0	1	1	0
G:	1	1	1	0
T:	1	1	1	2

Profile(Motifs):

A:	2/4	1/4	1/4	2/4
C:	0	1/4	1/4	0
G:	1/4	1/4	1/4	0
T:	1/4	1/4	1/4	2/4

What is probability Pr(CAGT|Profile)?
And Pr(CAGC|Profile)?

Solution: Laplace's Rule of Succession (Add 1 to each element to avoid zeros)

Motifs:

T	A	A	C
G	T	C	T
A	C	T	A
A	G	G	T

Count(Motifs):

A:	2+1	1+1	1+1	2+1
C:	0+1	1+1	1+1	0+1
G:	1+1	1+1	1+1	0+1
T:	1+1	1+1	1+1	2+1

Profile(Motifs):

A:	3/8	2/8	2/8	3/8
C:	1/8	2/8	2/8	1/8
G:	2/8	2/8	2/8	1/8
T:	2/8	2/8	2/8	3/8

What is probability Pr(CAGT|Profile)?
And Pr(CAGC|Profile)?