# CMSC423

# Chapter 2 – Defining motifs

- Recap: Looking for something common in front of multiple genes

- How do we formalize what we look for?

# Which one is a motif?

| | | | |
|---|---|---|---|
| TTACCTTAAC | TTACCTTAAC | CTACCTTAAC | CTAACTTGAC |
| GATATCTGTC | GTTATCTGTC | CTTATCTGTC | CTTATCTGTC |
| ACGGCGTTCG | ACGTCGTTAG | ATGTCGTTAG | ATATCGTTAC |
| CCCTAAAGAG | CCCTCAAGAG | CTCTCGAGAG | CTATCGTGAC |
| CGTCAGAGGT | CTTCAGAGGC | CTTCCGTGTC | CTTACGTGTC |

# Option 1: count minority bases

*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*) $3 + 4 + 0 + 0 + 1 + 1 + 1 + 5 + 2 + 3 + 6 + 4 = 30$

# Option 2: compute entropy

$$H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{N} p_i \log_2(p_i)$$

*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 |

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 |

# Entropy vs. counts

A
A
A
A
C
A
C
G
T
A

Majority base = A
SCORE(column) = 4

Profile

A – 0.6
C – 0.2
G – 0.1
T – 0.1

entropy = 1.57

A
A
A
A
C
A
C
C
C
A

Majority base = A
SCORE(column) = 4

Profile

A – 0.6
C – 0.4
G – 0
T – 0

entropy = 0.97

Next: Algorithms for finding motifs