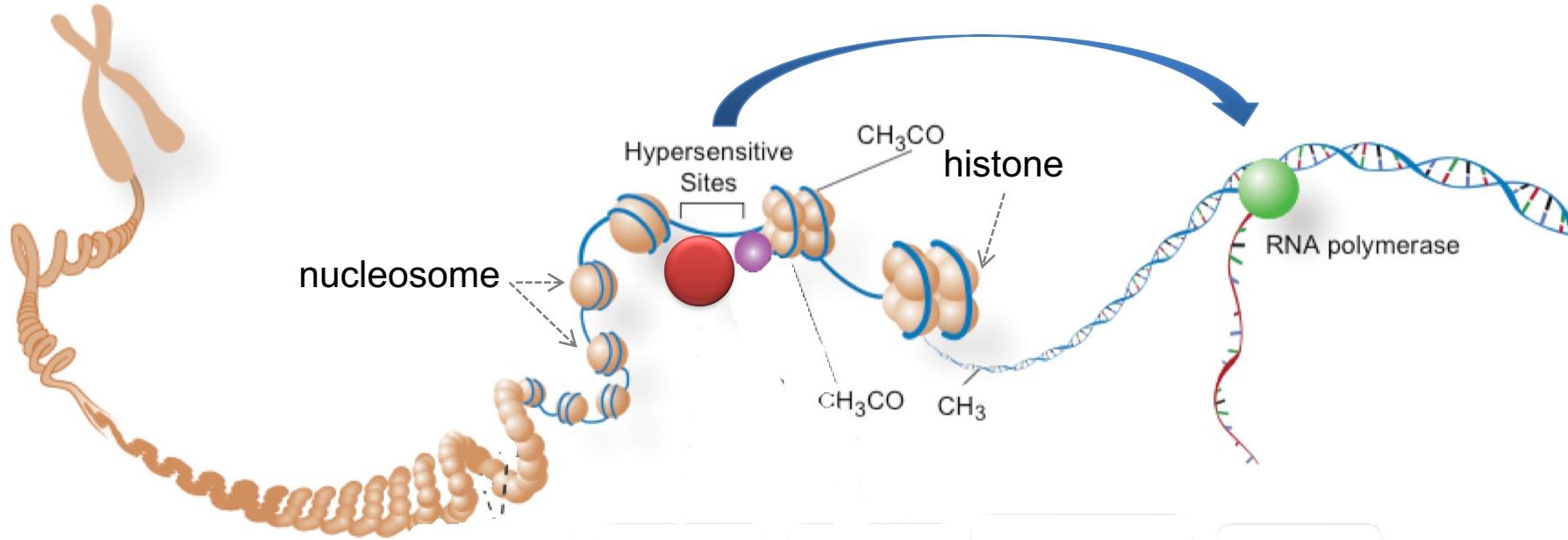


UNIT 2 EPIGENOMICS

Epigenome



The *epigenome* is a multitude of chemical compounds that can tell the *genome* what to do. The epigenome is made up of chemical compounds and proteins that can attach to DNA and direct such actions as turning genes on or off, controlling the production of proteins in particular cells.

-- from genome.gov

Epigenomic marks

	Chemical compounds	Proteins	Other molecules	Other information
DNA-associated	DNA methylation	Histones; DNA-binding proteins (Transcription factors*)	RNA (e.g., R loops)	
Chromatin-associated	Histone modifications: methylations, acetylations, ...	Histone variants; Chromatin regulators: Histone modifying enzymes: writer, readers, erasers; Chromatin remodeling complexes	Non-coding RNAs	<ul style="list-style-type: none">• Nucleosome positioning;• chromatin accessibility;• 3D genome organization;...

Unit 2: Epigenomics

- 1. Transcriptional regulation, DNA sequence motif finding
- 2. ChIP-seq: measuring chromatin epigenome, signal detection (TF binding sites)
- 3. ChIP-seq: signal detection continued (histone modification domains, etc.)
- 4-6

Transcriptional Regulation

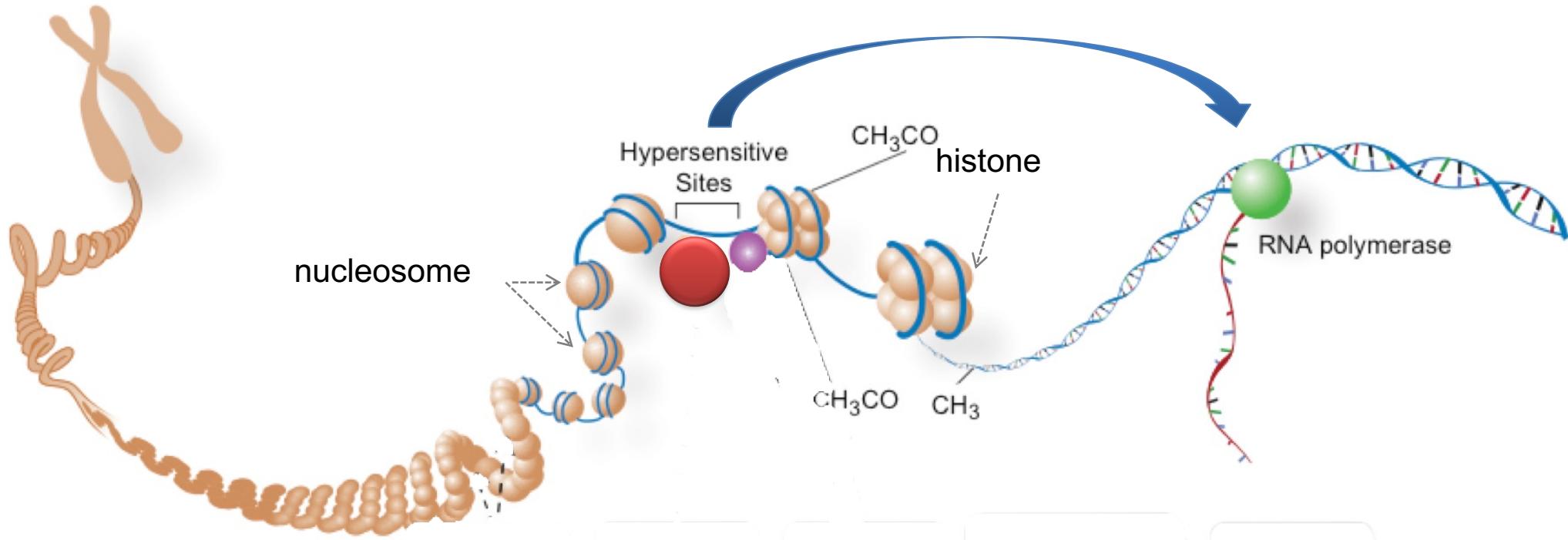
DNA Sequence Motif Finding

March 24, 2022

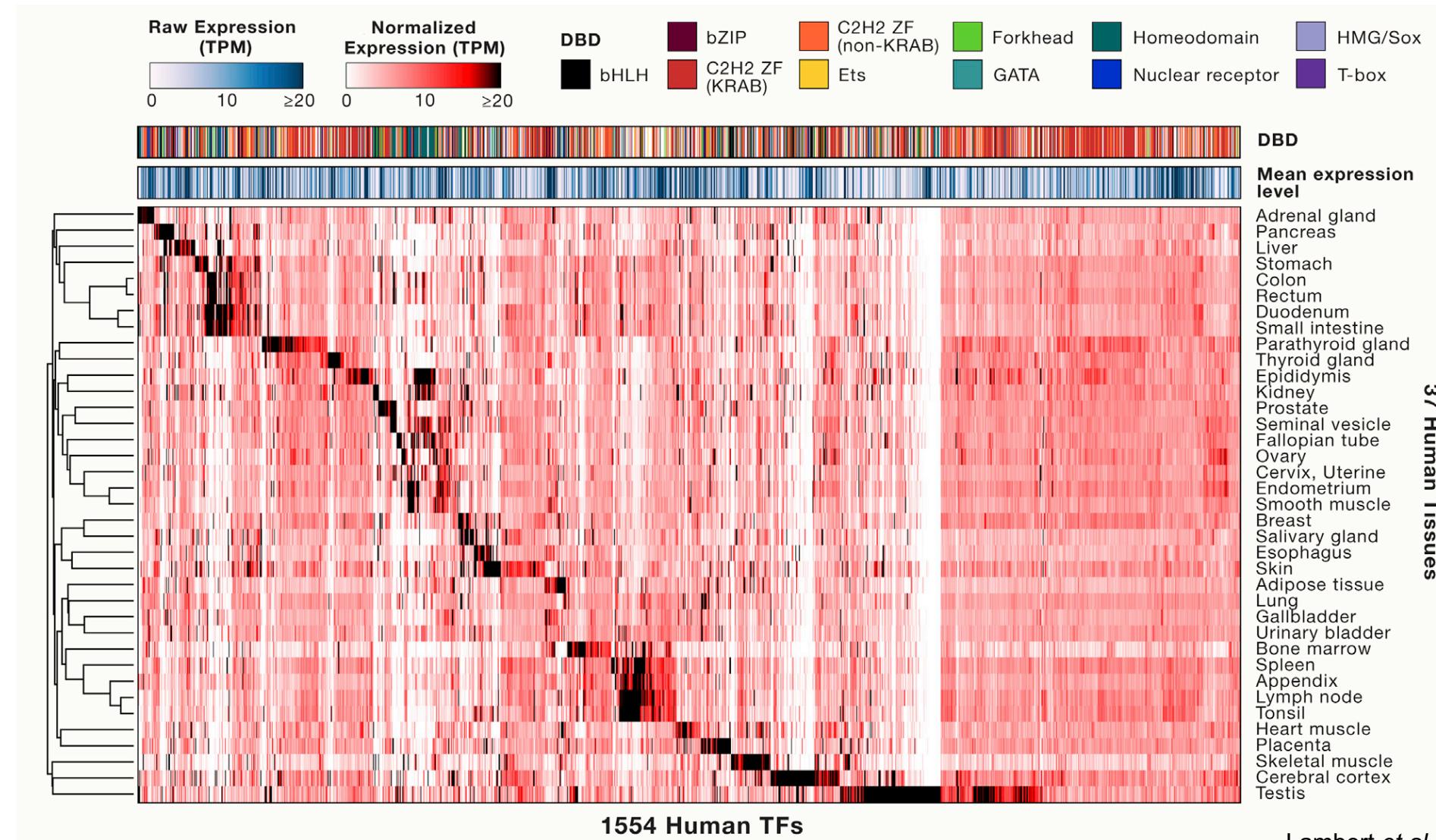
Outline

- Transcriptional regulation
- Sequence motif
- Motif representation: PWM
- Motif finding:
 - Deterministic approach: Regular expression enumeration
 - Probabilistic approach: Expectation-Maximization (E-M)
 - Probabilistic approach: Gibbs sampling

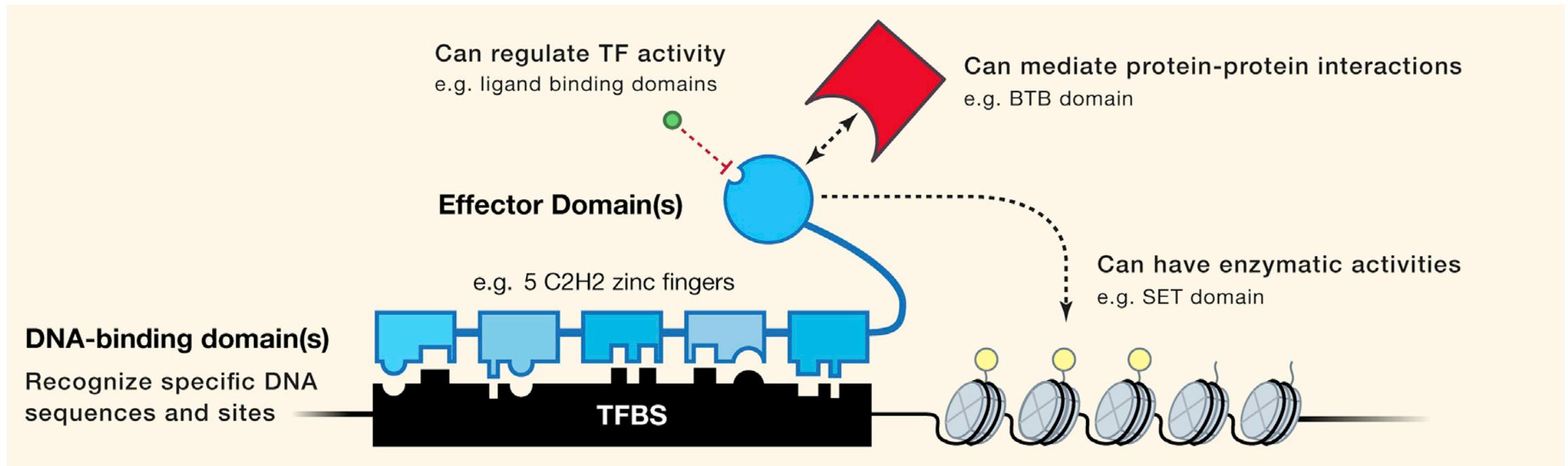
Transcription factors



Many TFs exhibit tissue/cell-type-specific expression patterns



Transcription factors

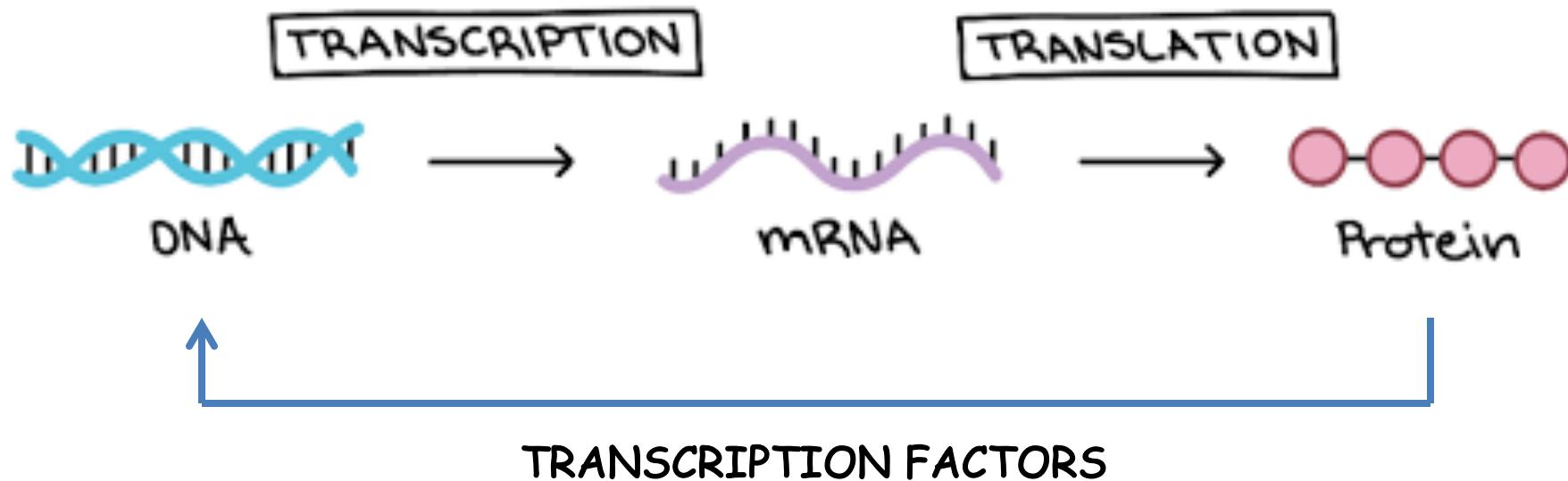


Lambert et al. Cell 2018

Transcription factors

- Structure: Effector domain and DNA binding domain(s)
- Functional studies:
 - Cell-type specific expression
 - Binding DNA sequence motif
 - Genome-wide binding sites
 - Target genes
 - Co-factors, etc.

Central Dogma of Molecular Biology



Entropy

- From statistical physics

$$S = k_B \ln \Omega$$

$$S_B = -k_B \sum_i p_i \ln(p_i)$$

(Boltzman entropy)



Entropy

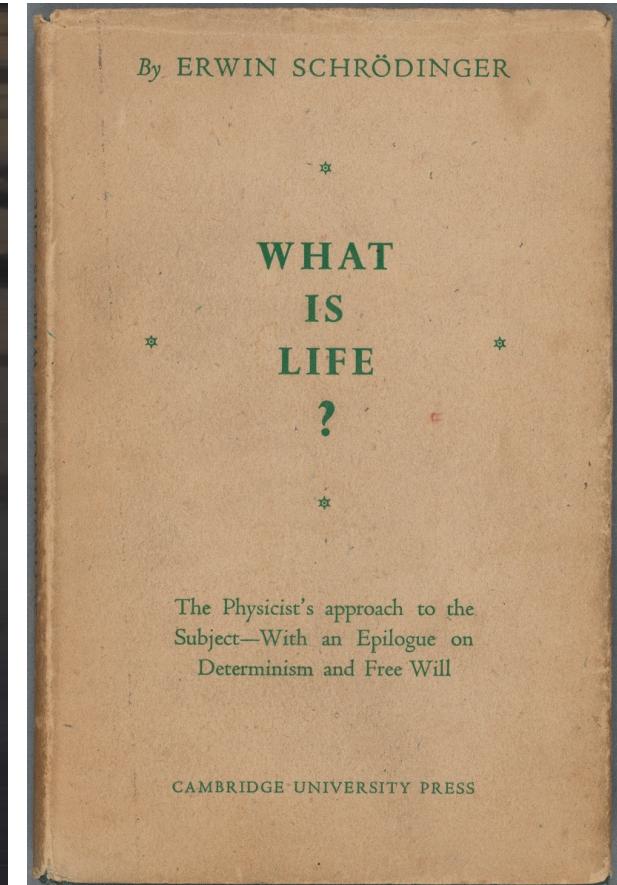
- Orderliness = negative entropy



Erwin Schrödinger
1887–1961

[A living organism] ... feeds upon negative entropy ... Thus the device by which an organism maintains itself stationary at a fairly high level of orderliness (= fairly low level of entropy) really consists in continually sucking orderliness from its environment.

Erwin Schrodinger



Entropy

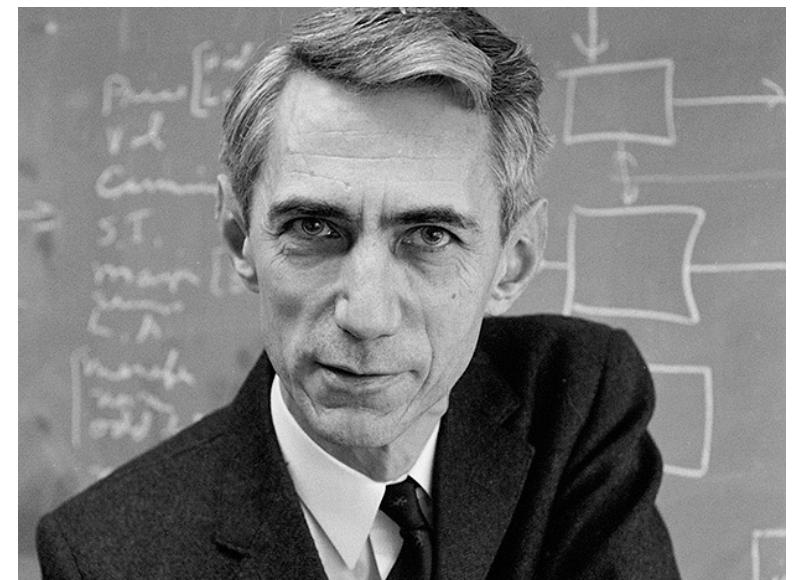
- Shannon entropy

$$H(X) = - \sum_i P(x_i) \log_2 P(x_i)$$

Expectation of Information Content

Information Content:

$$I(x) = - \log_2 P(x)$$



Claude Shannon
1916 – 2001

What is a motif?

- A recurring pattern; a distinctive pattern that occurs repeatedly.

Motif Number 1

- "the most often-painted building in America"
Rockport, Massachusetts



Sequence Motif

- What is a (biomolecular) sequence motif?
 - A pattern common to a set of DNA, RNA, or protein sequences that share a common biological property, such as functioning as binding sites for a particular protein

Sequence Motif Finding

- Computational motif finding:
 - Input data: a set of DNA sequences
 - e.g., upstream sequences of gene expression profile cluster
 - 20-1000 sequences, each 100-5000 bps long
 - Output: enriched sequence patterns (motifs)
- Ultimate goals for biology:
 - Which TFs are involved?
 - What are their binding motifs and effects (enhance / repress gene expression)?
 - Which genes are regulated by this TF?
 - Why is there disease when a TF goes wrong?
 - Are there binding partner / competitor for a TF?

Motif Representation

- Regular expression: Consensus CACAAAA
binary decision Degenerate CRCAAAW

Summary of single-letter code recommendations

Symbol	Meaning	Origin of designation
G	G	Guanine
A	A	Adenine
T	T	Thymine
C	C	Cytosine
R	G or A	puRine
Y	T or C	pYrimidine
M	A or C	aMino
K	G or T	Keto
S	G or C	Strong interaction (3 H bonds)
W	A or T	Weak interaction (2 H bonds)
H	A or C or T	not-G, H follows G in the alphabet
B	G or T or C	not-A, B follows A
V	G or C or A	not-T (not-U), V follows U
D	G or A or T	not-C, D follows C
N	G or A or T or C	aNy

A/G

A/T

Motif Representation

- Position Weight Matrix (PWM)
 - Position-Specific Scoring Matrix (PSSM)

Pos 123456789

GAGGTAAAC

TCCGTAAGT

CAGGTTGGA

ACAGTCAGT

TAGGTCACT

TAGGTACTG

ATGGTAACT

CAGGTATAAC

TGTGTGAGT

AAGGTAAGT

$$M = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 3 & 6 & 1 & 0 & 0 & 6 & 7 & 2 & 1 \\ 2 & 2 & 1 & 0 & 0 & 2 & 1 & 1 & 2 \\ 1 & 1 & 7 & 10 & 0 & 1 & 1 & 5 & 1 \\ 4 & 1 & 1 & 0 & 10 & 1 & 1 & 2 & 6 \end{bmatrix}$$

$$M = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}$$

Position Weight Matrix (PWM)

- Graphic representation: Sequence Logo



PWM:

$$M = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \\ 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \\ 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \\ 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix}$$

Sequence Logo:



- SeqLogo consists of stacks of symbols, one stack for each position in the sequence
- The overall height of the stack indicates the sequence conservation at that position (information content)
- The height of symbols within the stack indicates the relative frequency of nucleic acid at that position

$$R_i = \log_2(4) - H_i$$

$$H_i = - \sum_b f_{b,i} \times \log_2 f_{b,i}$$

Position Weight Matrix (PWM)

- Motif Matching Score: Likelihood Ratio Score

$$M = \begin{matrix} A & \begin{bmatrix} 0.3 & 0.6 & 0.1 & 0.0 & 0.0 & 0.6 & 0.7 & 0.2 & 0.1 \end{bmatrix} \\ C & \begin{bmatrix} 0.2 & 0.2 & 0.1 & 0.0 & 0.0 & 0.2 & 0.1 & 0.1 & 0.2 \end{bmatrix} \\ G & \begin{bmatrix} 0.1 & 0.1 & 0.7 & 1.0 & 0.0 & 0.1 & 0.1 & 0.5 & 0.1 \end{bmatrix} \\ T & \begin{bmatrix} 0.4 & 0.1 & 0.1 & 0.0 & 1.0 & 0.1 & 0.1 & 0.2 & 0.6 \end{bmatrix} \end{matrix}$$

G A G G T A A A C

$$S = \log_2 \frac{\Pr(x \text{ from } \theta_m)}{\Pr(x \text{ from } \theta_0)}$$

Score for GAGGTAAAC = \log_2

$$\frac{p_m G \times p_m A \times p_m G \times p_m G \times p_m T \times p_m A \times p_m A \times p_m A \times p_m C}{p_0 G \times p_0 A \times p_0 G \times p_0 G \times p_0 T \times p_0 A \times p_0 A \times p_0 A \times p_0 C}$$

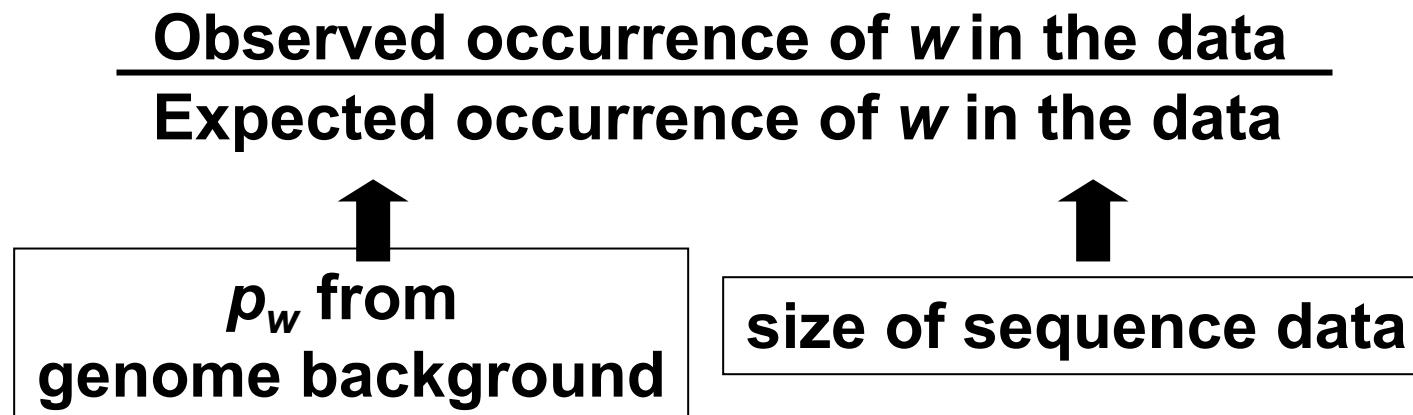
$$p_0(A, C, G, T) = [0.28, 0.22, 0.22, 0.28]$$

De Novo Sequence Motif Finding

- Goal: look for common sequence patterns enriched in the input data (compared to a background, e.g., genome)
- Deterministic approach: Regular expression enumeration
 - Pattern driven approach
 - Enumerate k-mers; check significance in dataset
- Probabilistic approaches: PWM update
 - Data driven approach, use data to refine motifs
 - Expectation-Maximization (E-M) approach
 - Gibbs Sampling

Regular Expression Enumeration

- Check over-representation for every w -mer
 - Expected w occurrence in data
 - Consider genome sequence + current data size
 - Observed w occurrence in data
 - Over-represented w is potential TF binding motif
- Suffix tree implementation of RE motif hits (e.g., WEEDER)



Regular Expression Enumeration

- Exhaustive, guaranteed to find global optimum, and can find multiple motifs
- Not as flexible with base substitutions, long list of similar good motifs, and limited with motif width

Probabilistic Approach

- Objects:
 - seq: sequence data to search for motif
 - θ_0 : non-motif probability (genome background) parameters
 - θ : motif probability matrix parameters
 - π : motif site locations
- Problem: $P(\theta, \pi | \text{seq}, \theta_0)$
- Approach: alternately estimate
 - π by $P(\pi | \theta, \text{seq}, \theta_0)$
 - θ by $P(\theta | \pi, \text{seq}, \theta_0)$
 - E-M and Gibbs sampler differ in the estimation methods

Expectation-Maximization: E Step

- E step: $\pi \mid \theta, \text{seq}, \theta_0$

TTGACGACTGCACGT

TTGAC

LR₁

TGACG

LR₂

GACGA

LR₃

ACGAC

LR₄

CGACT

LR₅

GACTG

LR₆

ACTGC

LR₇

CTGCA

LR₈

...

LR₁ = likelihood ratio =

$$\frac{P(\text{TTGAC} \mid \theta)}{P(\text{TTGAC} \mid \theta_0)}$$

Pos	A	C	G	T
1	0.7	0.1	0.01	0.2
2	0.01	0.01	0.8	0.1
3	0.32	0.02	0.3	0.18
4	0.03	0.42	0.1	0.47
5	0.2	0.5	0.1	0.2

$$p_\theta T \times p_\theta T \times p_\theta G \times p_\theta A \times p_\theta C \\ = 0.3 \times 0.3 \times 0.2 \times 0.3 \times 0.2$$

Expectation-Maximization

- E step: $\pi \mid \theta, \text{seq}, \theta_0$
- M step: $\theta \mid \pi, \text{seq}, \theta_0$

TTGACGACTGCACGT

TTGAC

LR₁

LR₁ × TTGAC

TGACG

LR₂

LR₂ × TGACG

GACGA

LR₃

LR₃ × GACGA

ACGAC

LR₄

LR₄ × ACGAC

CGACT

LR₅

...

GACTG

LR₆

ACTGC

LR₇

CTGCA

LR₈

...

- Scale ACGT at each position, θ reflects weighted average of π

Expectation-Maximization: M Step

TTGACGACTGCACGT

0.8 × TTGAC

0.2 × TGACG

0.6 × GACGA

0.5 × ACGAC

0.3 × CGACT

0.7 × GACTG

0.4 × ACTGC

0.1 × CTGCA

0.9 × TGCAC

...

$$T_1\% = \frac{0.8+0.2+0.9+\dots}{0.8+0.2+0.6+0.5+0.3+0.7+0.4+0.1+0.9+\dots}$$

$$G_2\% = \frac{0.2+0.3+0.9+\dots}{0.8+0.2+0.6+0.5+0.3+0.7+0.4+0.1+0.9+\dots}$$

$$C_5\% = \frac{0.8+0.5+0.4+0.9+\dots}{0.8+0.2+0.6+0.5+0.3+0.7+0.4+0.1+0.9+\dots}$$

Obtain updated θ

Expectation-Maximization

- E step: $\pi \mid \theta, \text{seq}, \theta_0$

TTGACGACTGCACGT

TTGAC

LR₁

TGACG

LR₂

GACGA

LR₃

ACGAC

LR₄

CGACT

LR₅

GACTG

LR₆

ACTGC

LR₇

CTGCA

LR₈

...

- M step: $\theta \mid \pi, \text{seq}, \theta_0$

LR₁ × TTGAC

LR₂ × TGACG

LR₃ × GACGA

LR₄ × ACGAC

...

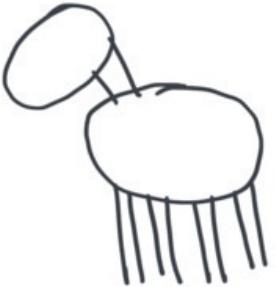
- Iterate until θ does not improve.
- Representative method:
MEME

HOW TO: DRAW A HORSE

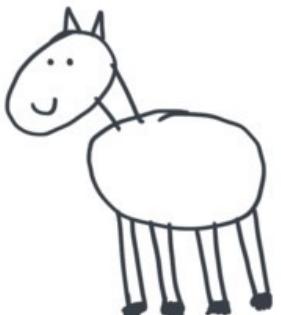
BY VAN OKTOP



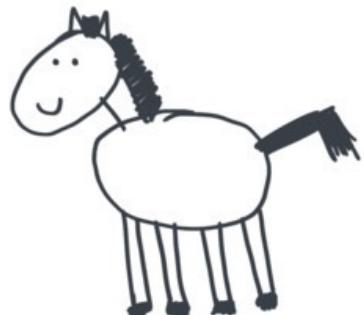
① DRAW 2 CIRCLES



② DRAW THE LEGS



③ DRAW THE FACE



④ DRAW THE HAIR



⑤ ADD
SMALL
DETAILS.

Summary

- Epigenome
- Entropy
- Motif
- Motif representation: PWM
- Motif finding: E-M