Transcription Factors

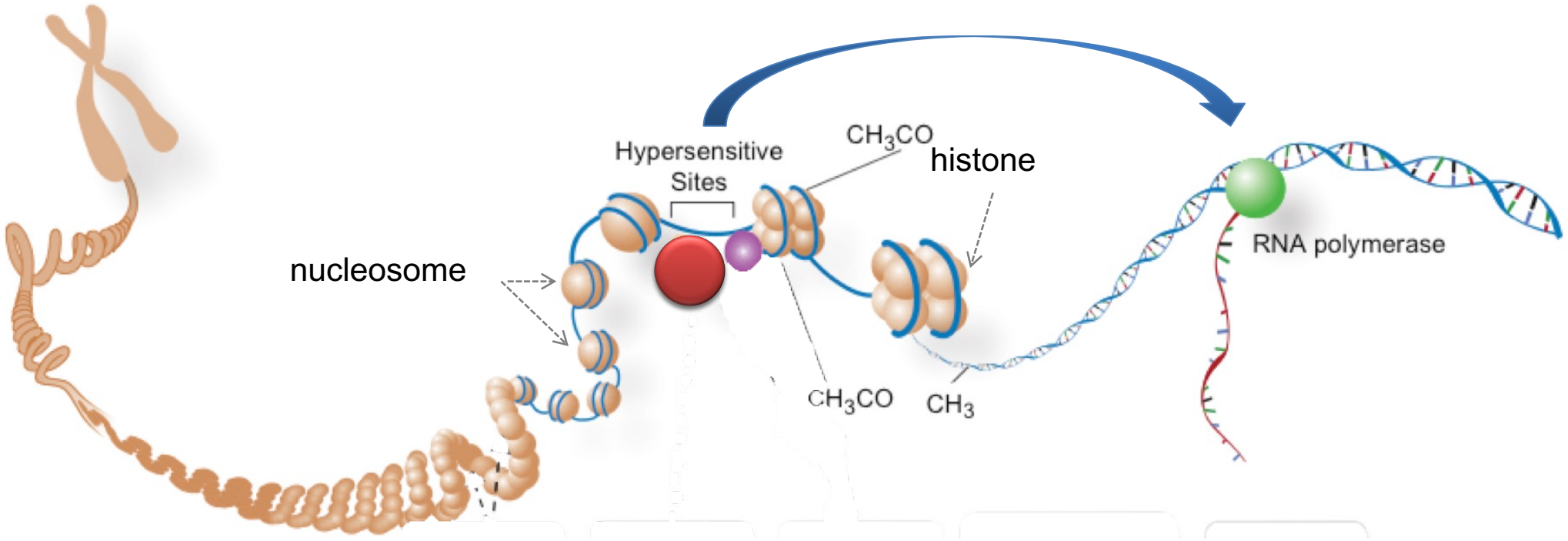
Chongzhi Zang

Acknowledgements: Some materials in the slides are borrowed from Harvard STAT 115 course taught by X. Shirley Liu.

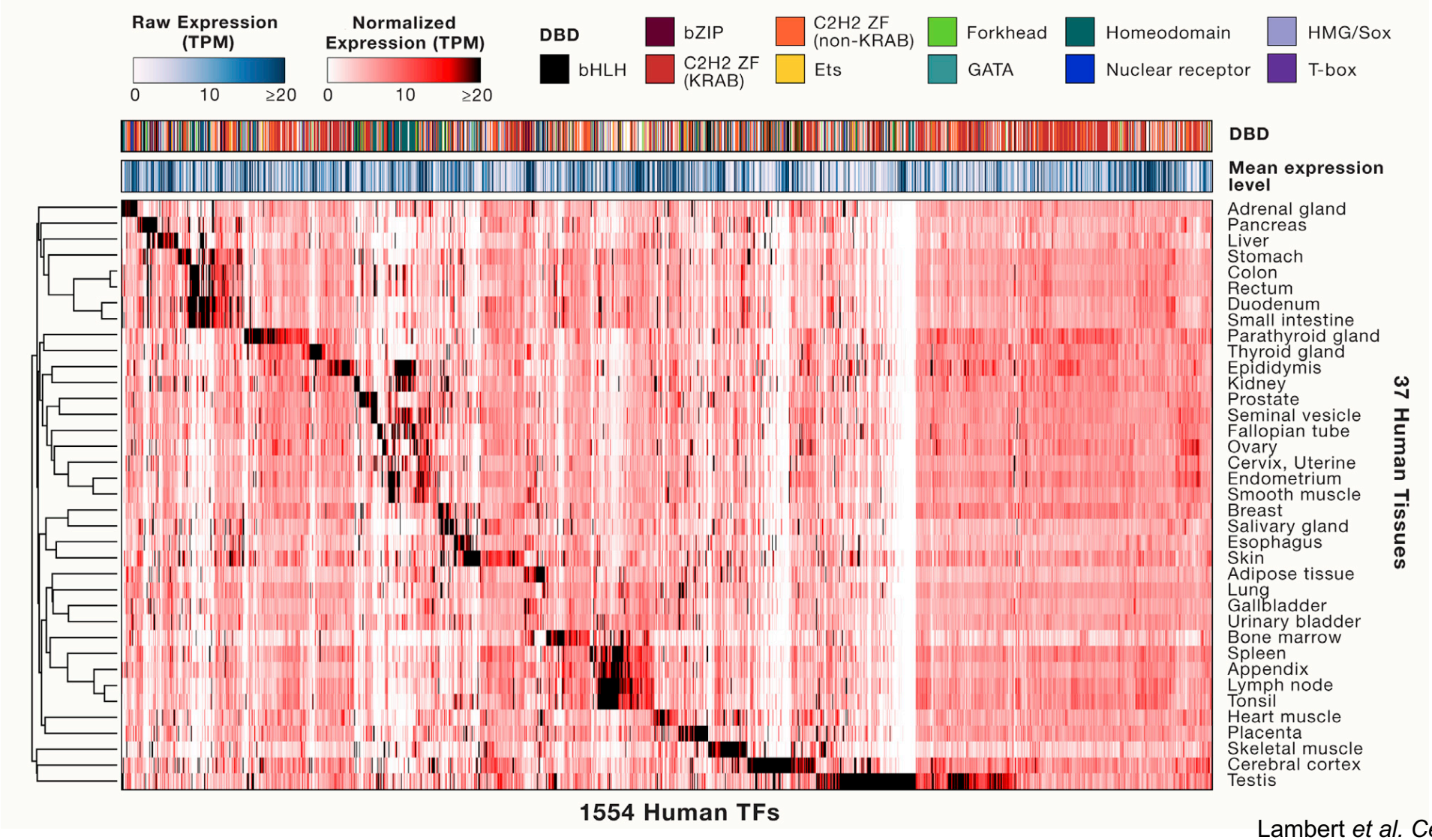
Outline

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

Transcription factors

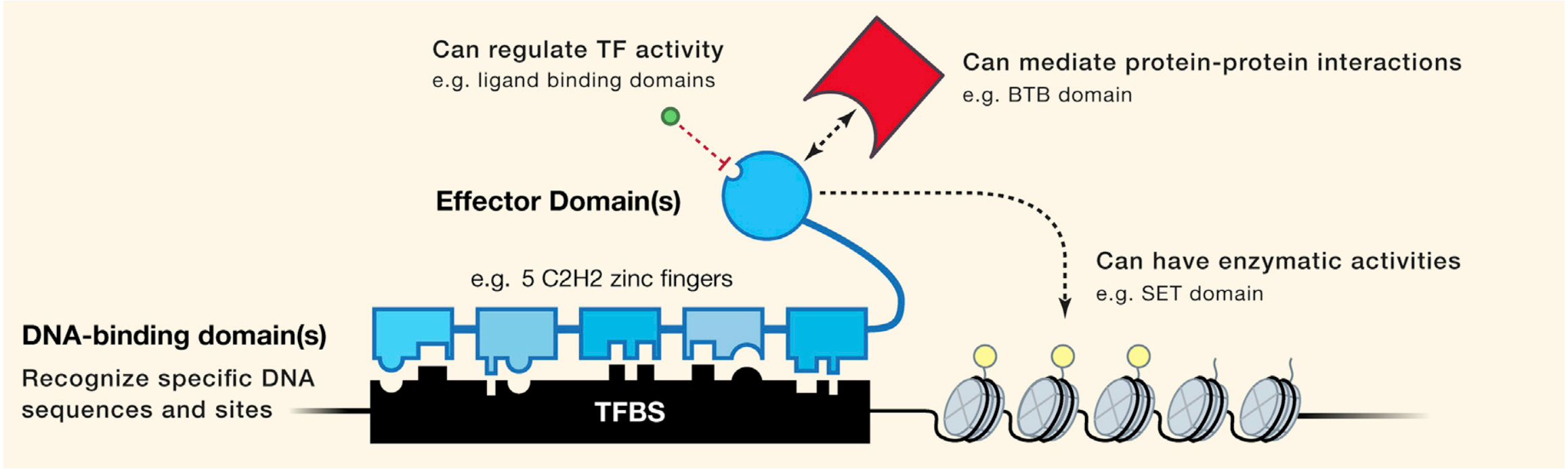


Many TFs exhibit tissue/cell-type-specific expression patterns



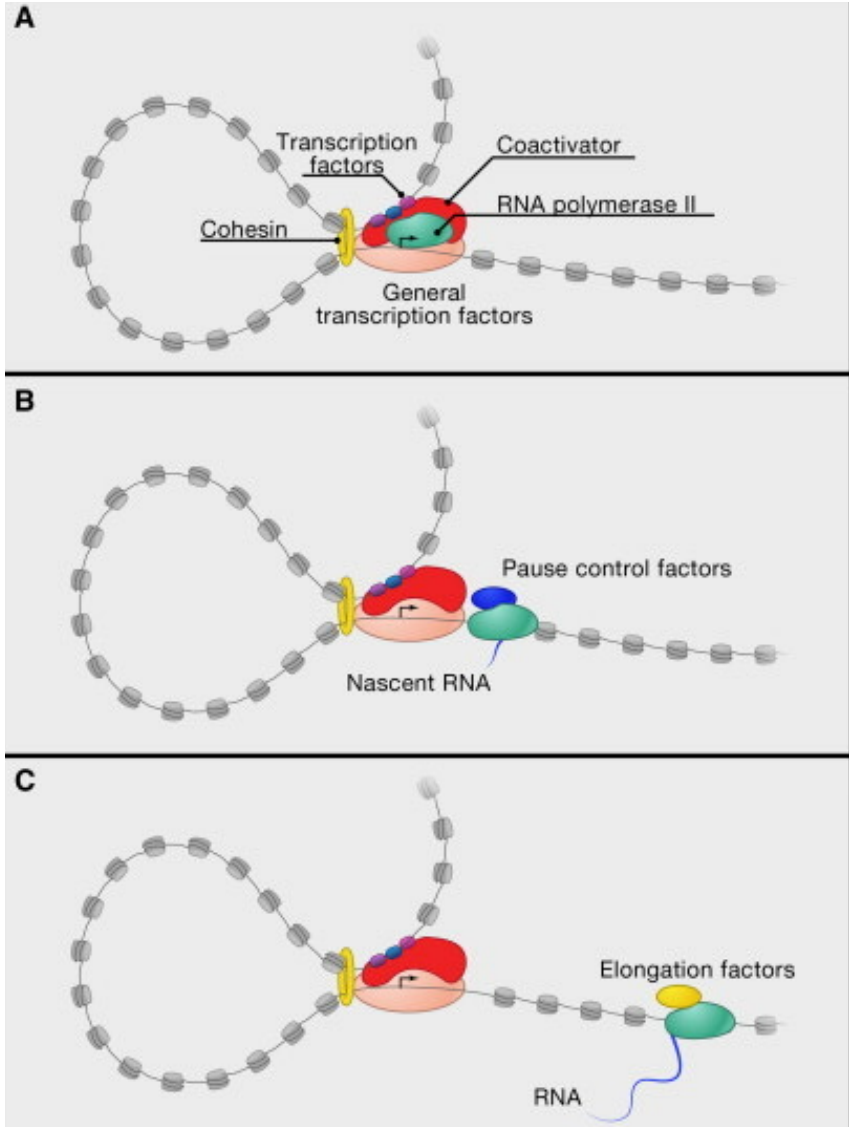
Lambert *et al. Cell* 2018

Transcription factors



Lambert *et al. Cell* 2018

Transcriptional regulation

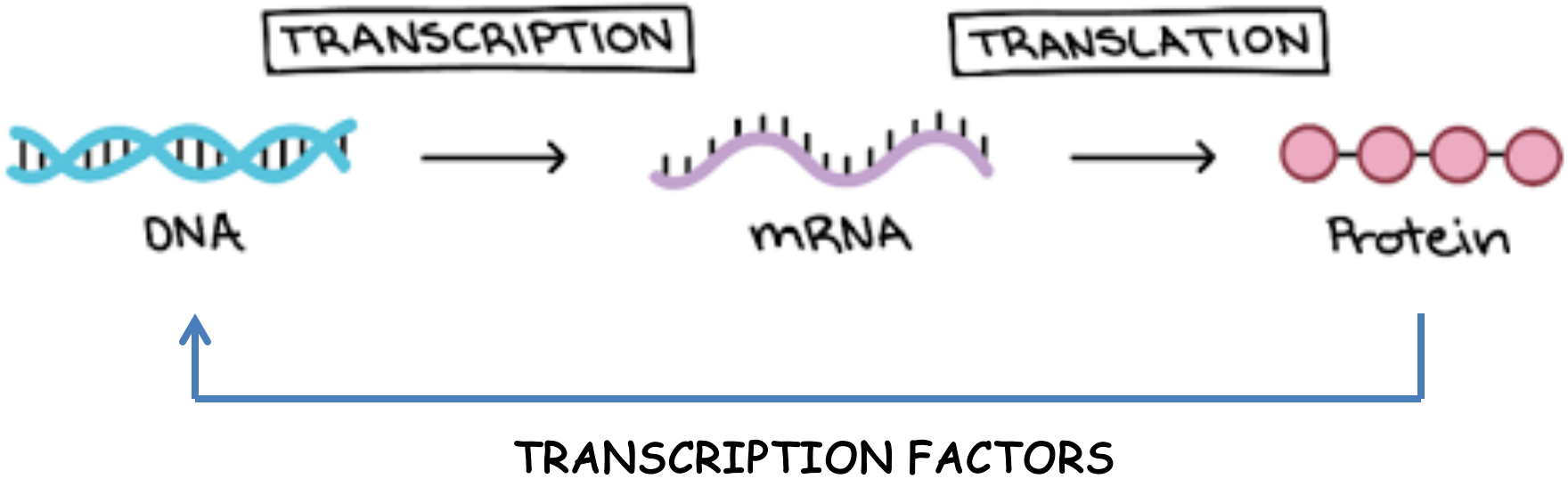


Lee & Young. *Cell* 2013

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



What is a motif?

Motif Number 1

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

Sequence Motif

- What is a motif?
  - A recurring pattern; a distinctive pattern that occurs repeatedly.

- 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

A/G **A/T**

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

IUPAC

Motif Representation

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

```
Pos   123456789
      GAGGTAAAC
      TCCGTAAGT
      CAGGTTGGA
      ACAGTCAGT
      TAGGTCATT
      TAGGTACTG
      ATGGTAACT
      CAGGTATAC
      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

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

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

Entropy

- From statistical physics

$$S = k_{\mathrm{B}} \ln \Omega$$

$$S_{\mathrm{B}} = -k_{\mathrm{B}} \sum_i p_i \ln(p_i)$$

(Boltzmann entropy)

Ludwig Boltzmann, who spent much of his life studying statistical mechanics, died in 1906, by his own hand. Paul Ehrenfest (Boltzmann's student), carrying on the work, died similarly in 1933. Now it's our turn to study statistical mechanics.

-- David L. Goodstein, in “States of Matter”

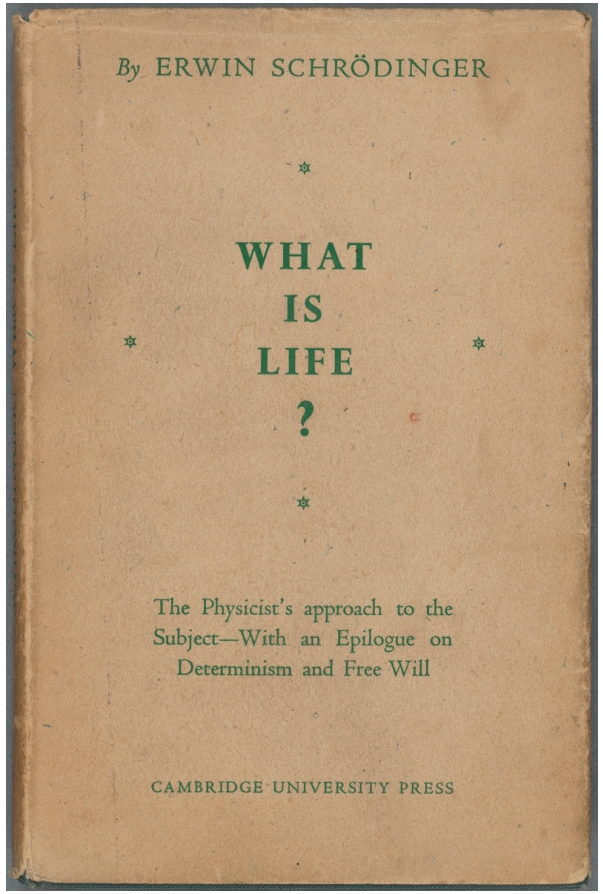
Entropy

- 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

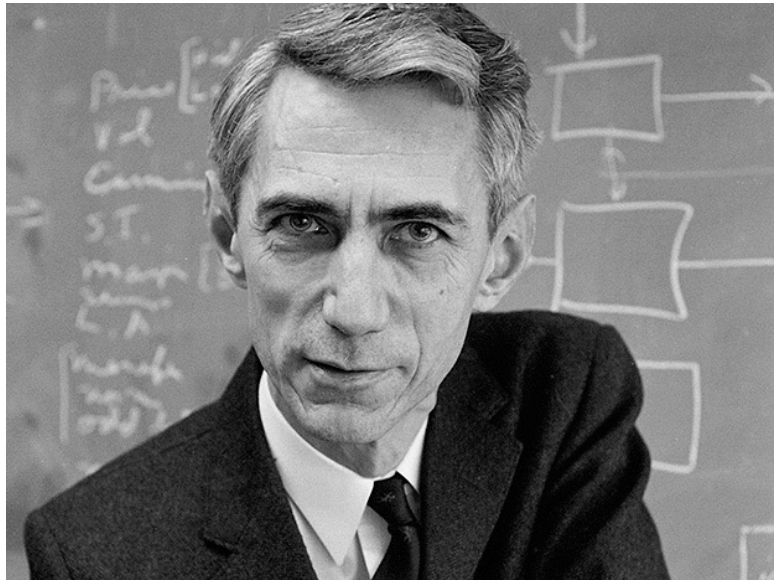
- Shannon entropy

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

Expectation of Information Content

Information Content:

$$I(x) = -\log_2 \mathrm{P}(x)$$



Claude Shannon
1916 – 2001

Position Weight Matrix (PWM)

- Graphic representation: Sequence Logo

PWM:

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

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

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

G A G G T A A A C

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

$$\Pr(x \text{ from } \theta) = \prod_{i=1}^{w} p(X_i|\theta)$$

Score for GAGGTAAAC **=** $\log_2$

$$\frac{p_mG \times p_mA \times p_mG \times p_mG \times p_mT \times p_mA \times p_mA \times p_mA \times p_mC}{p_0G \times p_0A \times p_0G \times p_0G \times p_0T \times p_0A \times p_0A \times p_0A \times p_0C}$$

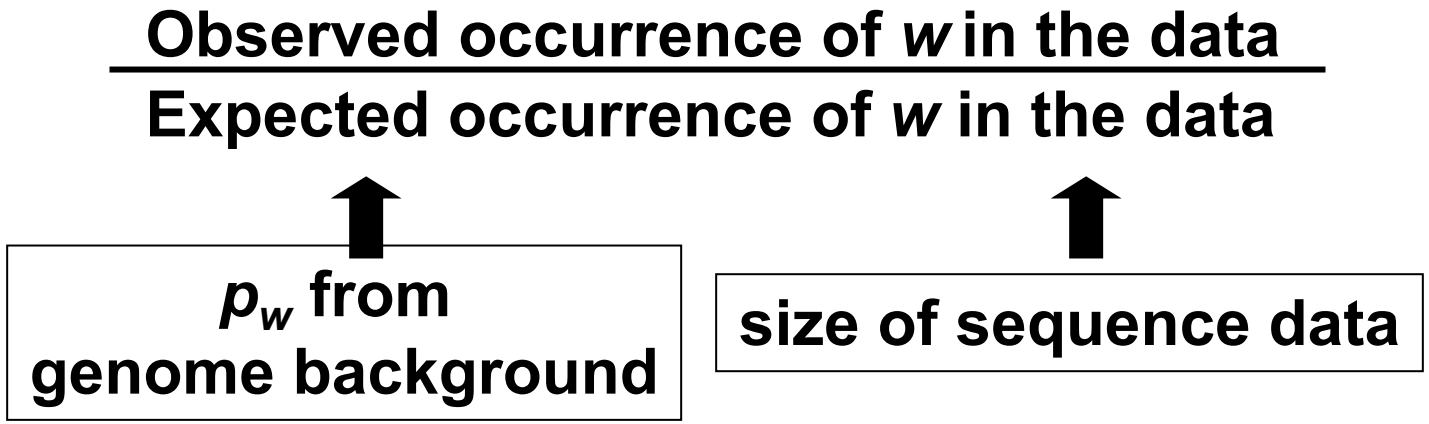
$$p_0(\text{A, C, G, T}) = [0.28,\ 0.22,\ 0.22,\ 0.28]$$

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: $\mathrm{P}(\theta, \pi \mid \mathrm{seq}, \theta_0)$
- Approach: alternately estimate
  - π by $\mathrm{P}(\pi \mid \theta, \mathrm{seq}, \theta_0)$
  - θ by $\mathrm{P}(\theta \mid \pi, \mathrm{seq}, \theta_0)$
  - 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
         . . .
```

LR_1 = 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

$$p_0T \times p_0T \times p_0G \times p_0A \times p_0C$$
$$= 0.3 \times 0.3 \times 0.2 \times 0.3 \times 0.2$$

Expectation-Maximization

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

```
TTGACGACTGCACGT
TTGAC              LR1
 TGACG             LR2
  GACGA            LR3
   ACGAC           LR4
    CGACT          LR5
     GACTG         LR6
      ACTGC        LR7
       CTGCA       LR8
         ...
```

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

$LR_1 \times$ TTGAC

$LR_2 \times$ TGACG

$LR_3 \times$ GACGA

$LR_4 \times$ ACGAC

...

- 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
…

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

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

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

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

LR_1 × TTGAC

LR_2 × TGACG

LR_3 × GACGA

LR_4 × ACGAC

...

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

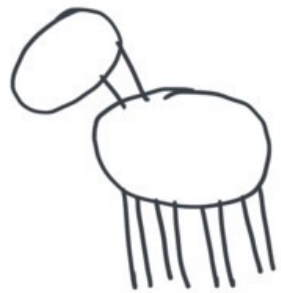
Summary

- Transcriptional regulation
- Motif
- Entropy
- Motif representation: PWM
- Motif finding: E-M

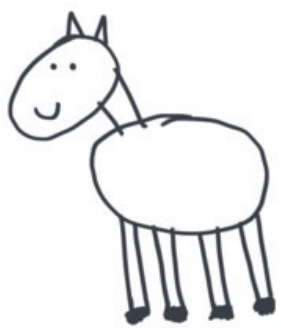
HOW TO:
DRAW A HORSE

BY VAN OKTOP

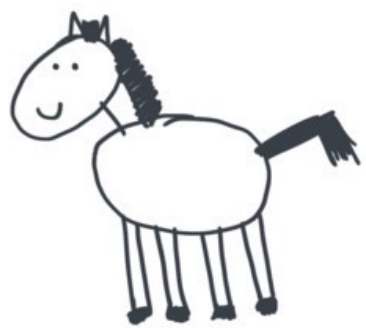
1 DRAW 2 CIRCLES



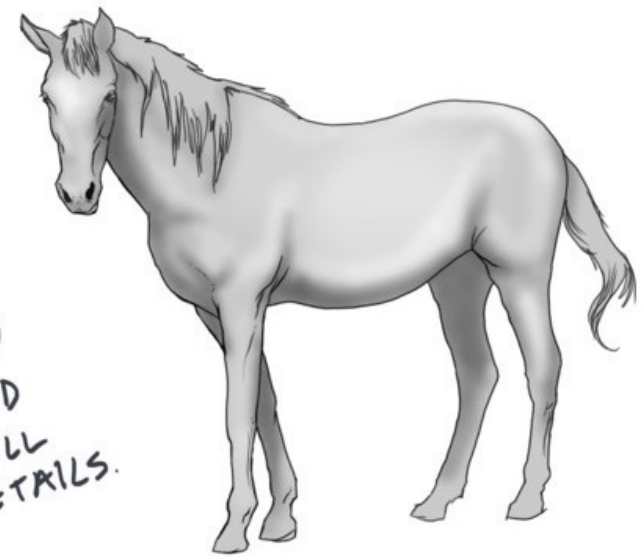
2 DRAW THE LEGS



3 DRAW THE FACE



4 DRAW THE HAIR



5 ADD SMALL DETAILS.