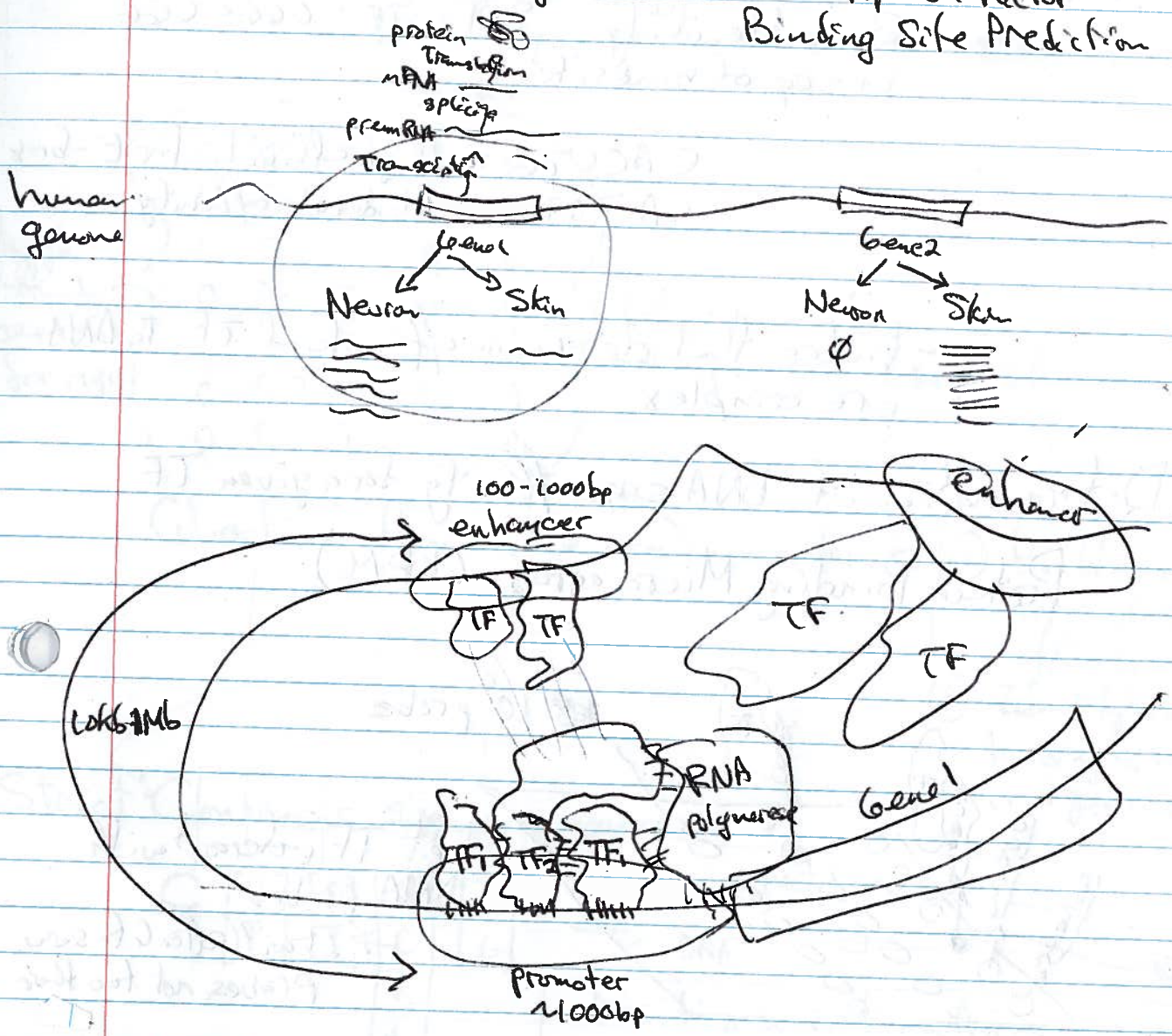
Gene Regulation + Transcription Factor Binding Site Prediction



Transcription factors: - Proteins that - bind specific DNA seq.
- Alter expression of nearby gene(s)
- activation
- repression

- In humans: 2000 different TFs
↳ each binds to different DNA sequences
Transcription Factor Binding sites

TFBS: - 6-15 bp long
- Some flexibility in seq. of bind site

E-box TF bind: CACGTG
SP1 TF: GGGCGGG

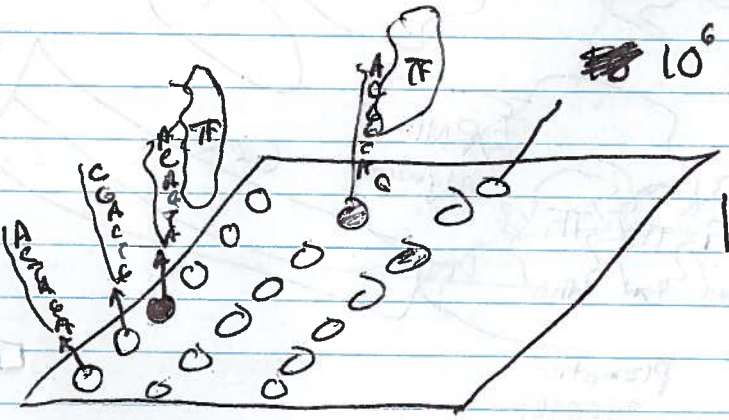
CACGTG: High affinity for E-box
CATGTG: Moderate affinity

⋮

- Rules that determine affinity of TF to DNA are complex

Determination of DNA seq. affinity for a given TF

Protein Binding Micro-array (PBM)



10^6 probes

Let TF interact with DNA probes.
→ It will attach to some probes, not to others

Take picture of PBM ⇒ infer affinity of TF to each probe

⇒

	intensity
ACAGTA	1000
ACGGTA	250
⋮	

Example: [redacted]

DNA sequences bound by TF [redacted] myc

not independent

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

high affinity sequences for myc

Representative sample of all possible sites TF

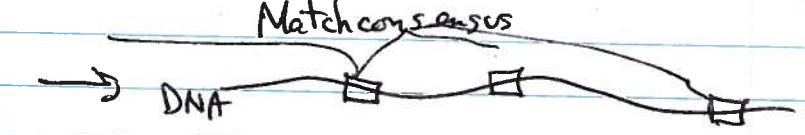
Question: How can we use in order to
① Build model for that TF's affinity
② Identify new binding sites in a genome

Strict Consensus sequence approach

$$C\begin{bmatrix}C\\T\end{bmatrix}\begin{bmatrix}A\\T\end{bmatrix}\begin{bmatrix}A\\C\\G\\T\end{bmatrix}\begin{bmatrix}A\\G\end{bmatrix}$$

or

$$C\begin{bmatrix}C\\T\end{bmatrix}[T]\begin{bmatrix}A\\C\\G\\T\end{bmatrix}\begin{bmatrix}A\\G\end{bmatrix}$$



Position weight matrices

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

Candidate sequence to score: C T A G G

$1 \times 3/7 \times 4/7 \times 1/7 \times 3/7 = 9/7^4 = 0.00\ldots$

How to identify matches for a given PWM in a given sequence

S= A C A G T G A C T T

For each starting position i in S

Calculate score of S[i, i+1, ... i+5-1] on PWM

If score > Threshold, then predict Binding
other no binding