Operating System of the Cell

Regulation of Gene Expression

Gold Lab Symposium

May 14-15, 2026

Gary D. Stormo
Erlanger Professsor, emeritus
Department of Genetics
Edison Center for Genome Sciences
and Systems Biology

Washington University in St.Louis

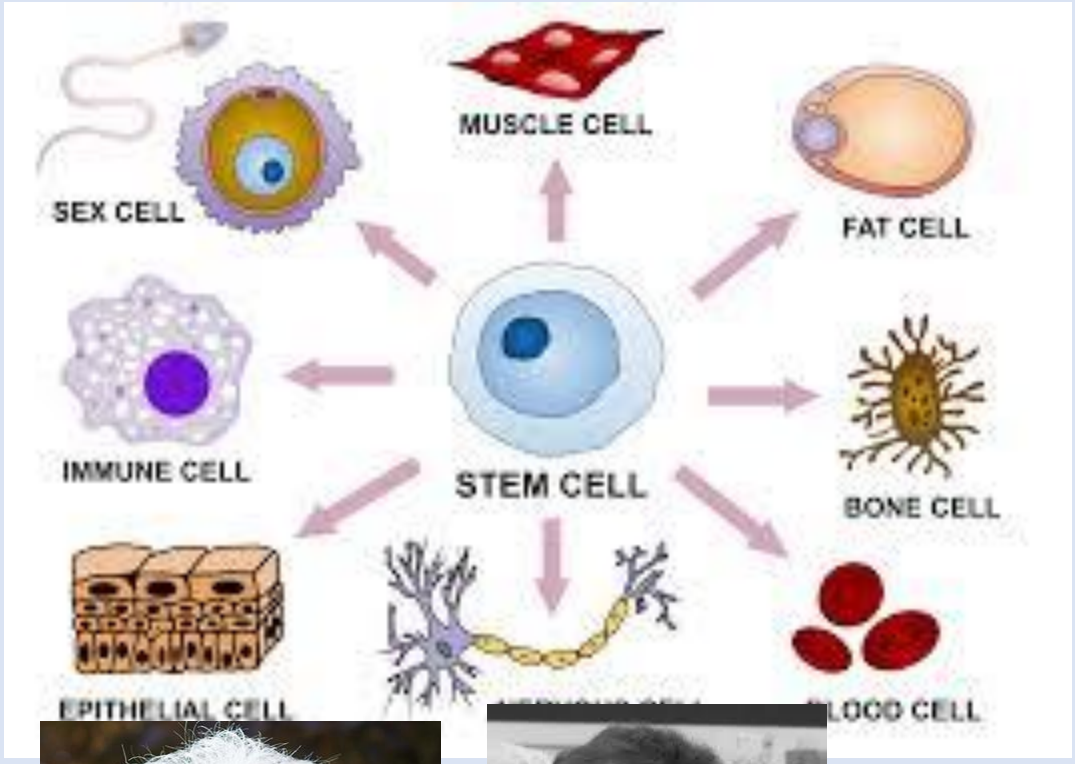
Advice from Max Delbruck for speaking to a diverse audience with unknown level of background knowledge:

"Assume your audience has zero knowledge but infinite intelligence."

Operating System of the Cell

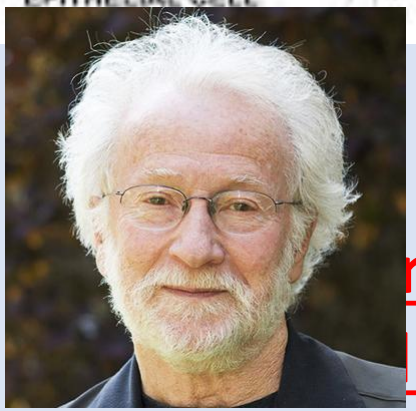
Regulation of Gene Expression

Consider the various cell types within the human body. Although they all share identical DNA, their appearances and functions differ significantly. Each cell type activates only a particular subset of the entire gene repertoire.



My Primary Interest:

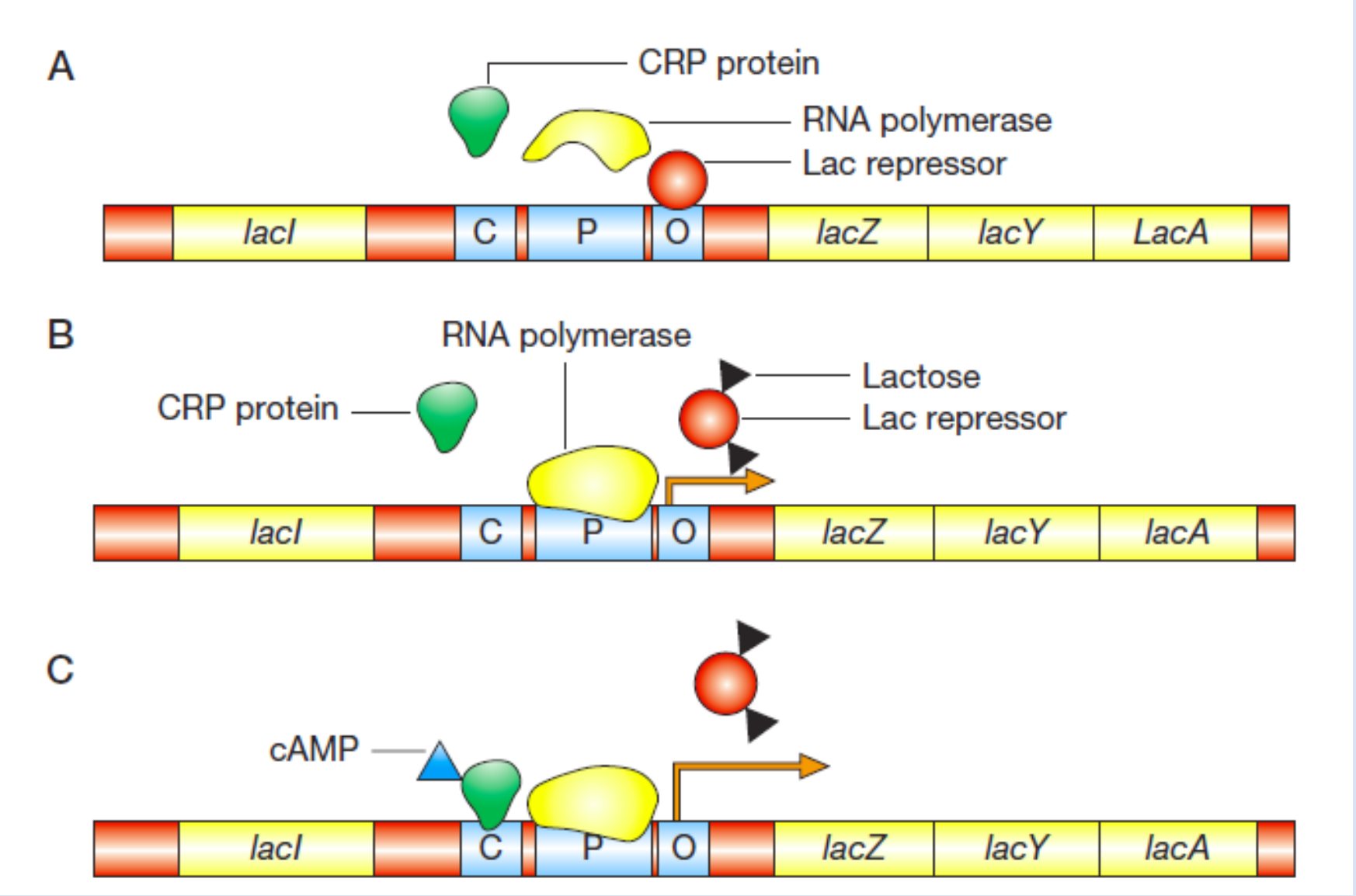
Regulation of Gene Expression



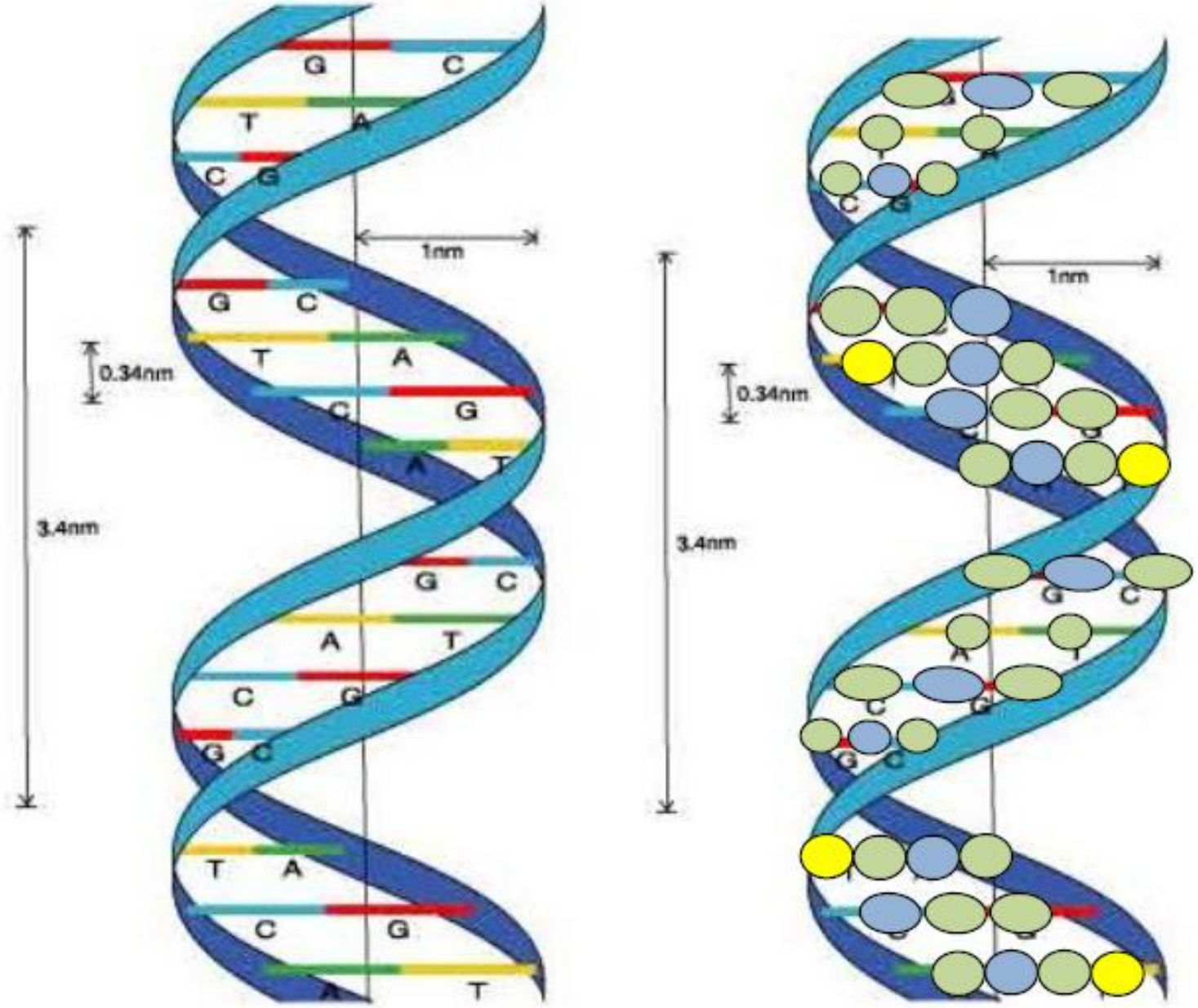
Larry Gold

Tom Schneider

Lactose regulatory system, Jacob and Monod, 1961

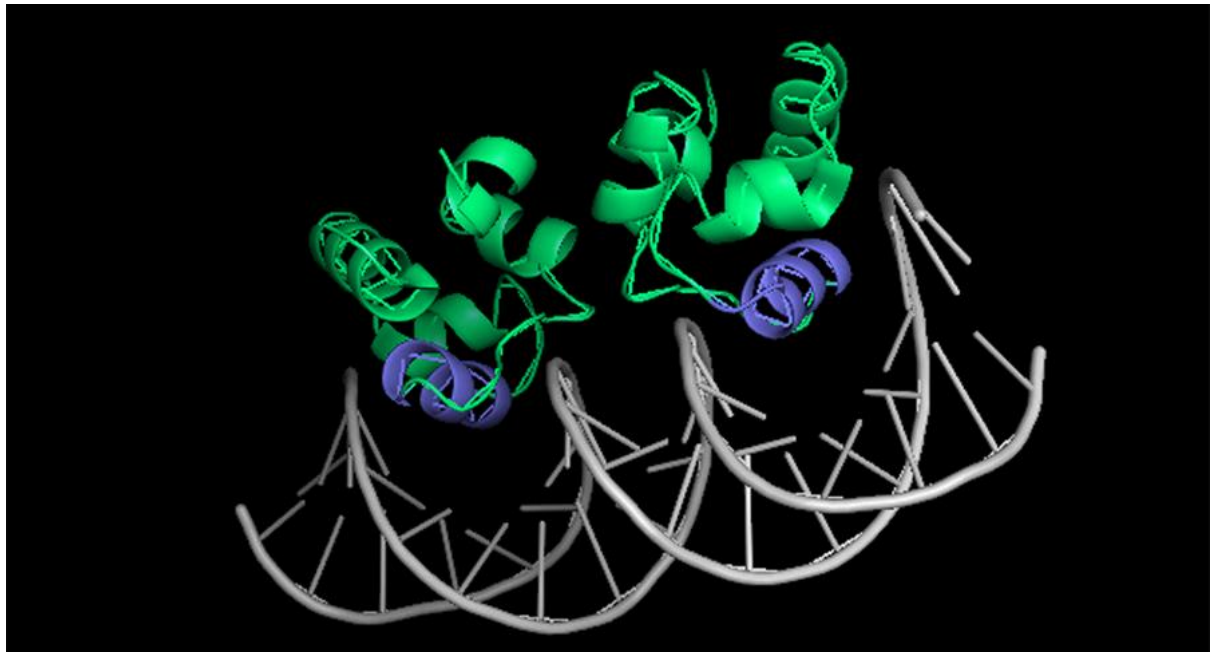


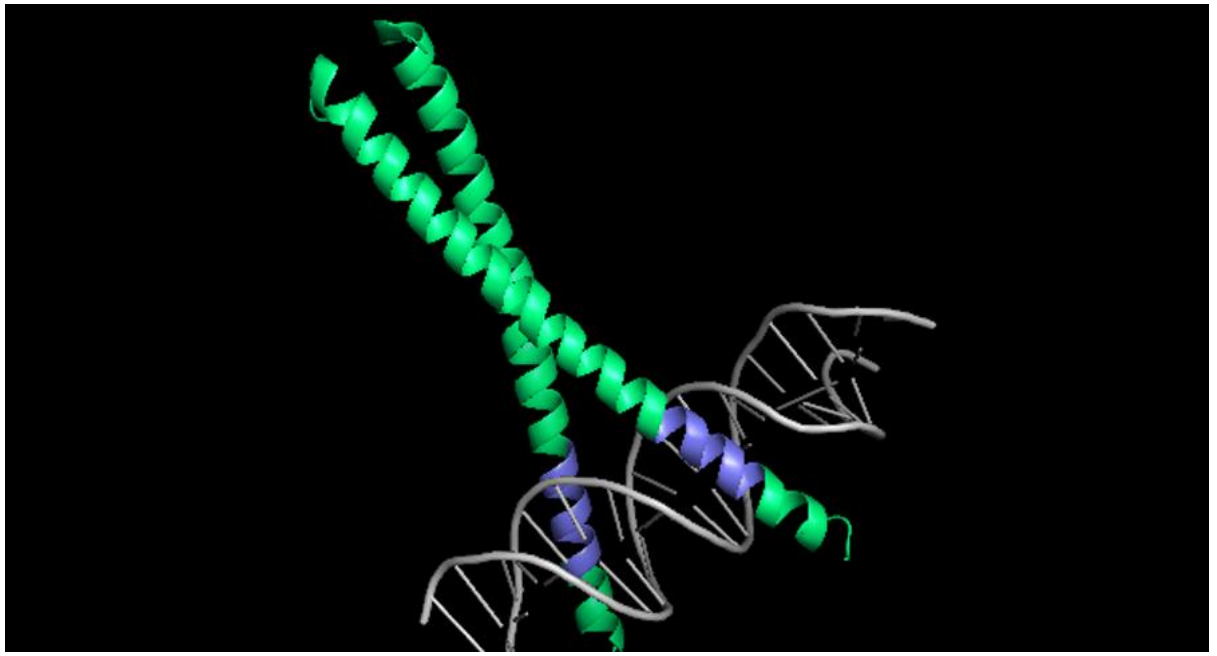
Stormo, Introduction to Protein-DNA Interactions, 2013, CSH Press

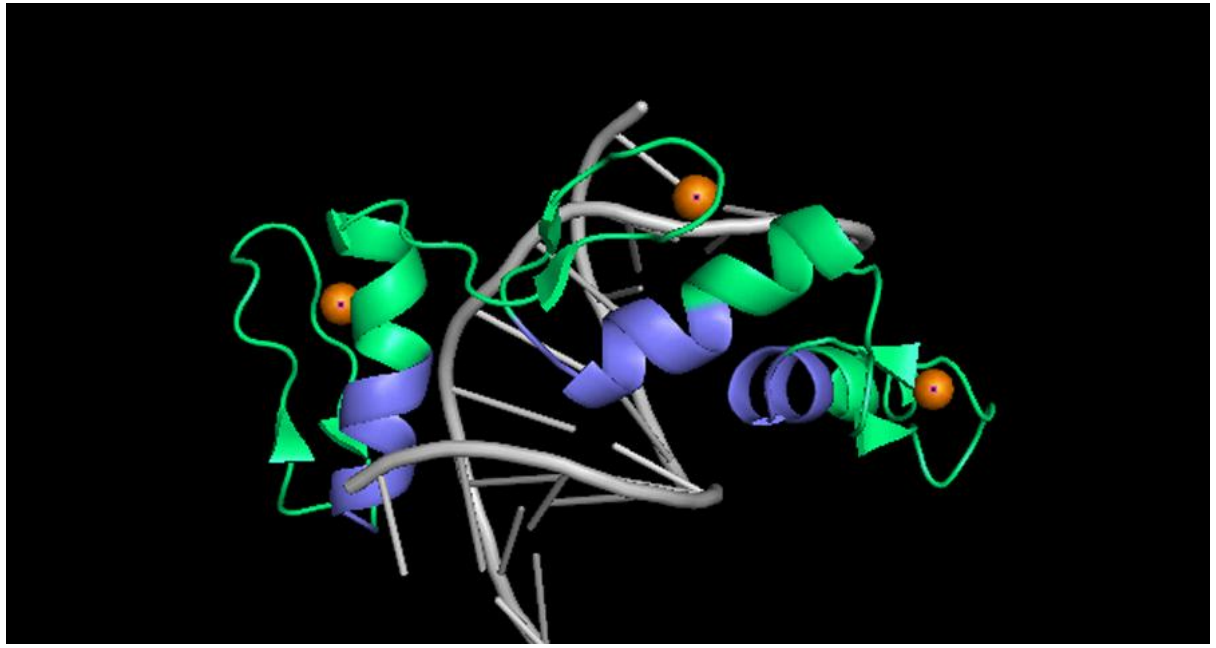


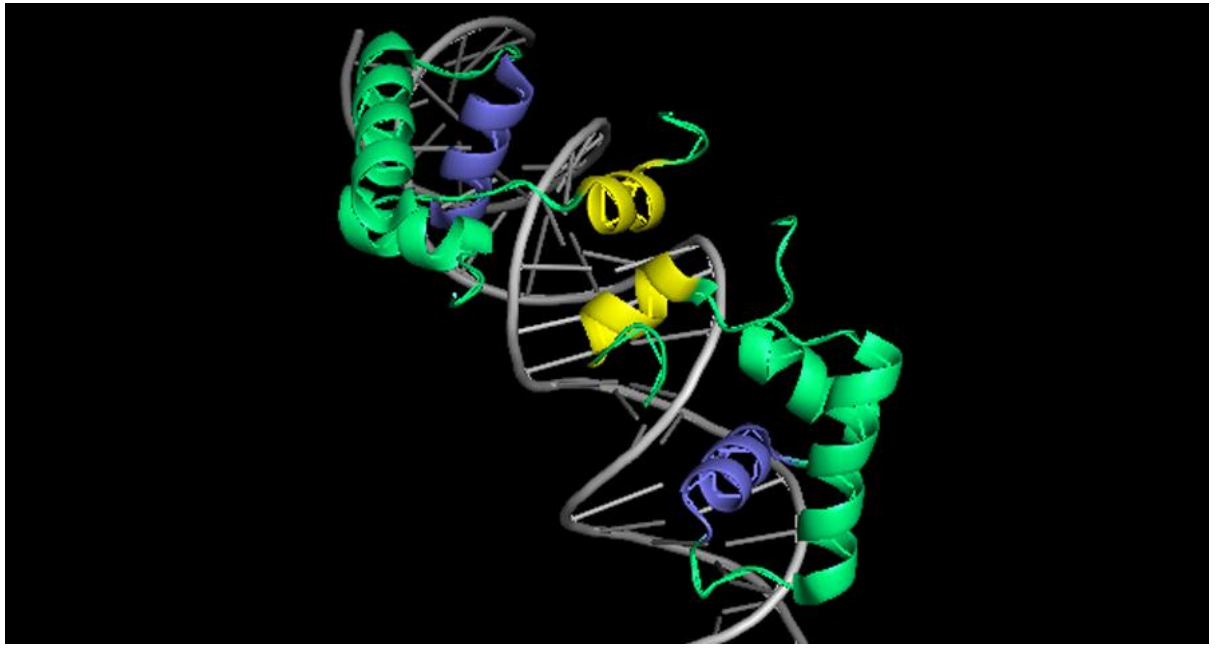
The surface of DNA contains chemical features that can be “read” by the chemical surface of proteins

Stormo, Introduction to Protein-DNA Interactions, 2013, CSH Press









Two Revolutions were occurring during my graduate school days

1. DNA sequencing revolution
   DNA sequencing invented (1977)
   Allowed new type of data for molecular biology research

2. Computer revolution
   Symbolized by development of personal computers (IBM PC 1975, Microsoft 1975, Apple 1976)

CRP binding site:

“Under conditions of low glucose, turn on the expression of the adjacent gene.”

TAATGTGAGTTAGCTCACTCAT

“Under conditions of low glucose, turn on the expression of the adjacent gene.”

TAATGTGAGTTAGCTCACTCAT
cgcTGTGAccgtGgTCgCagtT

Eventually no positions are completely conserved

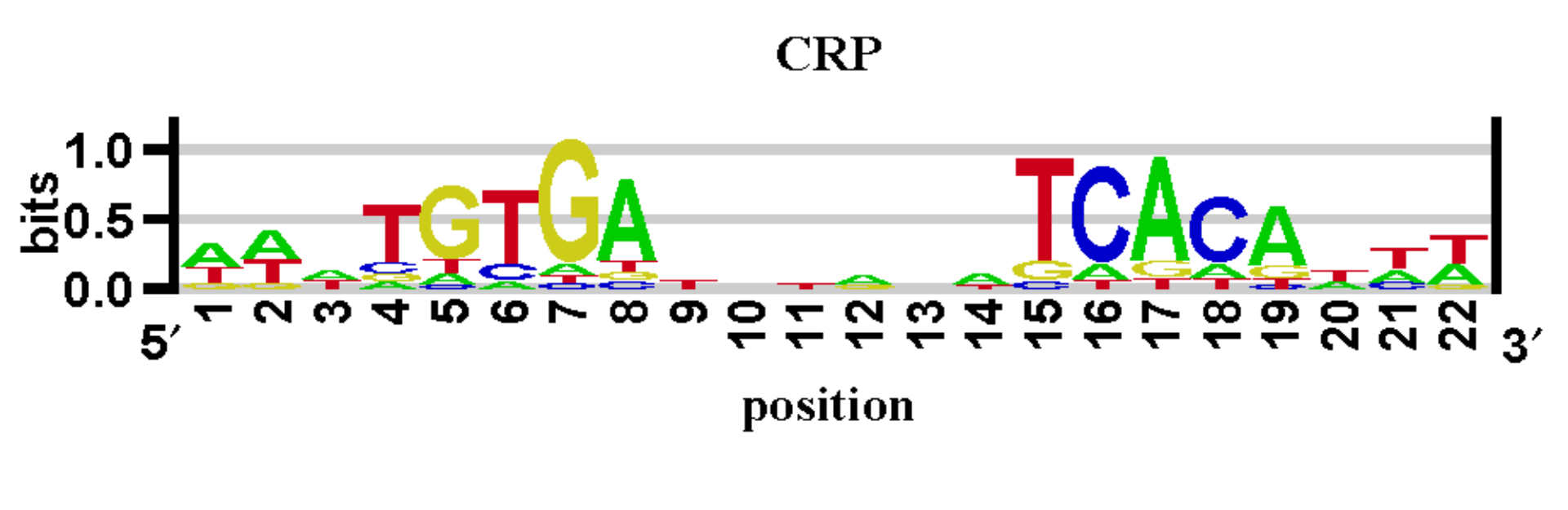
TAATGTGAGTTAGCTCACTCAT
cgcTGTGAccgtGgTCgCagtT
tttTtTGAtcgtttTCaCattT
aaacgTGAtagccgTCaaacaa
:

TGTGAnnnnnnTCACA

“consensus sequence” can be used to search allowing for mismatches

Eventually no positions are completely conserved

TAATGTGAGTTAGCTCACTCAT
cgcTGTGAccgtGgTCgCagtT
tttTtTGAtcgtttTCaCattT
aaacgTGAtagccgTCaaacaa

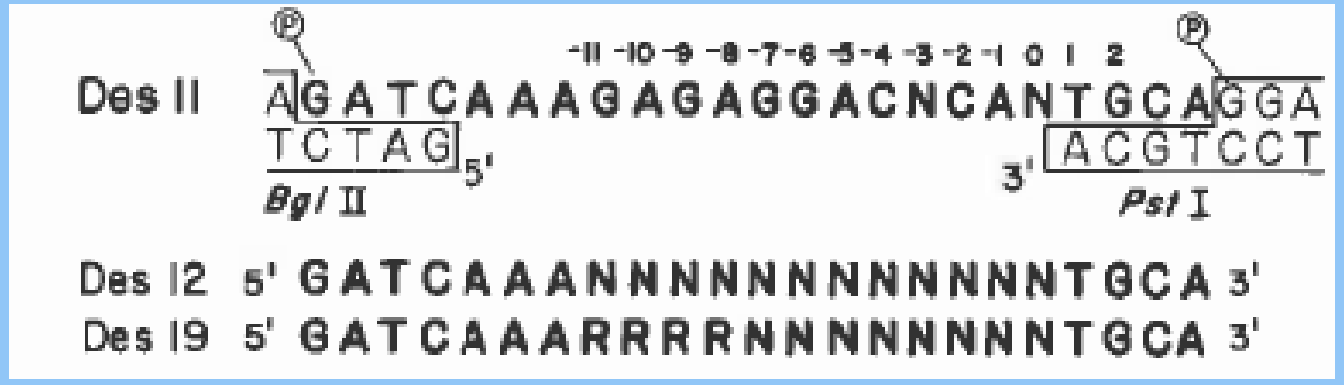


Third major technological advance: the ability to synthesize DNA

Sequence Specificity in Transcription and Translation, pages 341–350

RIBOSOME BINDING SITE SEQUENCES AND FUNCTION

John Childs[1], Keith Villanueba[2], Doug Barrick[2], Thomas D. Schneider[2,4], Gary D. Stormo[5], Larry Gold[2], Moshe Leitner[3,*4] and Marvin Caruthers[3]



Marv Caruthers in Chemistry at Univ of Colorado invented DNA synthesis method in 1981

Next Challenge

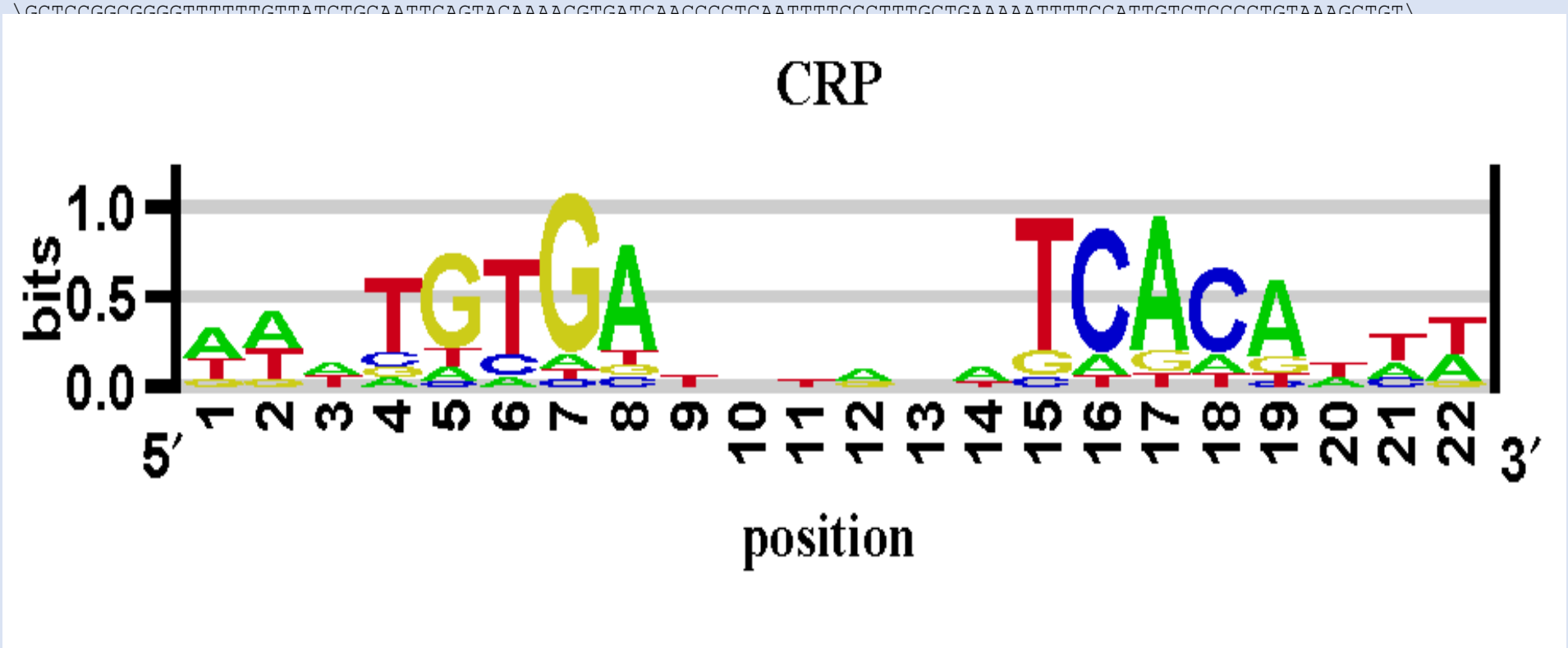
Outline of motif discovery problem

Fig. 6.1. A general schematic of the motif finding problem. Each *long thin line* represents a single DNA sequence. The *dark segments* within each line represent the binding sites whose positions are unknown in advance and we are trying to discover.

Motif Discovery, only need “regulatory regions” of co-regulated genes

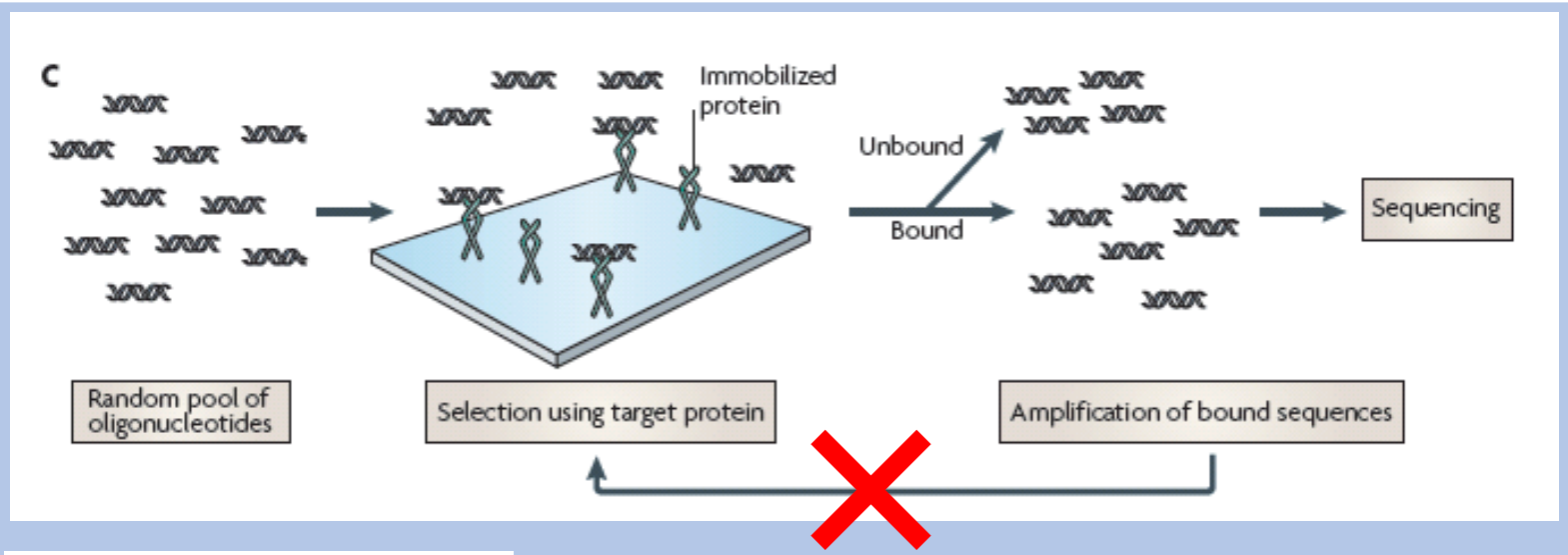
Input: Genes regulated by CRP in *E. coli*

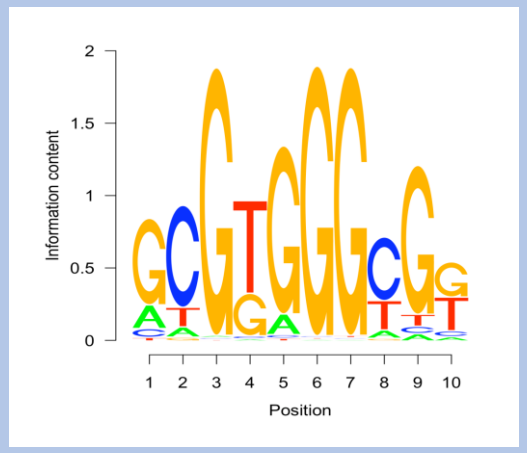
```
CE1CG
\TAATGTTTGTGCTGGTTTTTGTGGCATCGGGCGAGAATAGCGCGTGGTGTGAAAGACTGTTTTTTTGATCGTTTTCACAAAAATGGAAGTCCACAGTCTTGACAG\
ECOARABOP
\GACAAAAACGCGTAACAAAAGTGTCTATAATCACGGCAGAAAAGTCCACATTGATTATTTGCACGGCGTCACACTTTGCTATGCCATAGCATTTTTATCCATAAG\
ECOBGLR1
\ACAAATCCCAATAACTTAATTATTGGGATTTGTTATATATAACTTTATAAATTCCTAAAATTACACAAAGTTAATAACTGTGAGCATGGTCATATTTTTATCAAT\
ECOCRP
\CACAAAGCGAAAGCTATGCTAAAACAGTCAGGATGCTACAGTAATACATTGATGTACTGCATGTATGCAAAGGACGTCACATTACCGTGCAGTACAGTTGATAGC\
ECOCYA
\ACGGTGCTACACTTGTATGTAGCGCATCTTTCTTTACGGTCAATCAGCAAGGTGTTAAATTGATCACGTTTTAGACCATTTTTTCGTCGTGAAACTAAAAAAACC\
ECODEOP2
\AGTGAATTATTTGAACCAGATCGCATTACAGTGATGCAAACTTGTAAGTAGATTTCCTTAATTGTGATGTGTATCGAAGTGTGTTGCGGAGTAGATGTTAGAATA\
ECOGALE
\GCGCATAAAAAACGGCTAAATTCTTGTGTAAACGATTCCACTAATTTATTCCATGTCACACTTTTCGCATCTTTGTTATGCTATGGTTATTTCATACCATAAGCC\
ECOILVBPR
\GCTCCGGCGGGGTTTTTTGTTATCTGCAATTCAGTACAAAACGTGATCAACCCCTCAATTTTCCCTTTGCTGAAAAATTTTCCATTGTCTCCCCTGTAAAGCTGT\
```



Inferring Binding Energies from Selected Binding Sites

Yue Zhao, David Granas, Gary D. Stormo*





Further development and generalization of the algorithm

BEESEM: estimation of binding energy models using HT-SELEX data

Shuxiang Ruan[1], S. Joshua Swamidass[2] and Gary D. Stormo[1,*]

Many groups working to determine motifs for human TFs: Wolfe, Bulyk, Hughes,…
Jussi Taipale automated the process using robots

Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities

Arttu Jolma,[1,2] Teemu Kivioja,[1,3] Jarkko Toivonen,[3] Lu Cheng,[3] Gonghong Wei,[1] Martin Enge,[2] Mikko Taipale,[1] Juan M. Vaquerizas,[4] Jian Yan,[1] Mikko J. Sillanpää,[5] Martin Bonke,[1] Kimmo Palin,[3] Shaheynoor Talukder,[6] Timothy R. Hughes,[6] Nicholas M. Luscombe,[4] Esko Ukkonen,[3] and Jussi Taipale[1,2,7]

Genome Res. 2010 20: 861-873

~20 TFs

DNA-Binding Specificities of Human Transcription Factors

Arttu Jolma,[1,2,8] Jian Yan,[1,8] Thomas Whitington,[1] Jarkko Toivonen,[3] Kazuhiro R. Nitta,[1] Pasi Rastas,[3] Ekaterina Morgunova,[1] Martin Enge,[1] Mikko Taipale,[2] Gonghong Wei,[2] Kimmo Palin,[2] Juan M. Vaquerizas,[4] Renaud Vincentelli,[5] Nicholas M. Luscombe,[4] Timothy R. Hughes,[6] Patrick Lemaire,[7] Esko Ukkonen,[3] Teemu Kivioja,[1,2,3] and Jussi Taipale[1,2,*]

Cell 152, 327–339, January 17, 2013

~240 TFs

Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity

Cell 158, 1431–1443, September 11, 2014

Weirauch et al

>1000 TFs

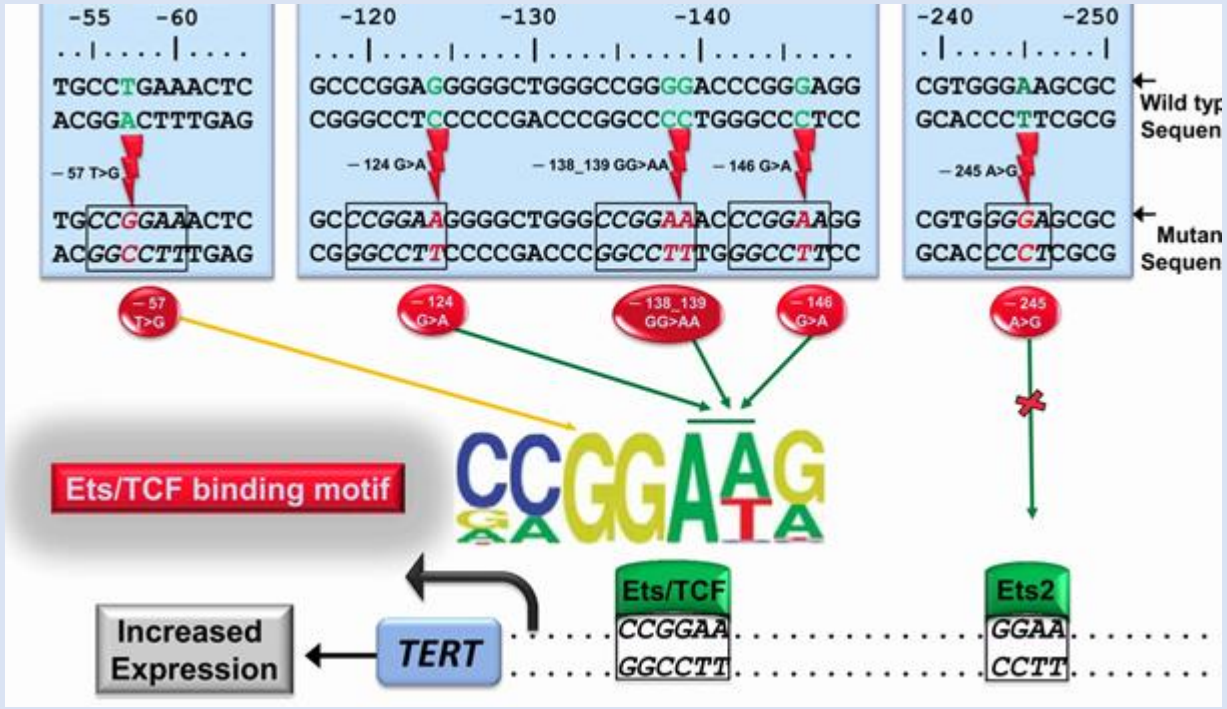
A Biomedical applications of TF motifs:

Genome-Wide Association Studies (GWAS) identify DNA variants associated (correlated) with phenotypes:

- normal human variations
- disease causing

The vast majority of GWAS hits (>90%) are in non-coding DNA. Suggests alterations in gene expression rather than gene function cause phenotypes

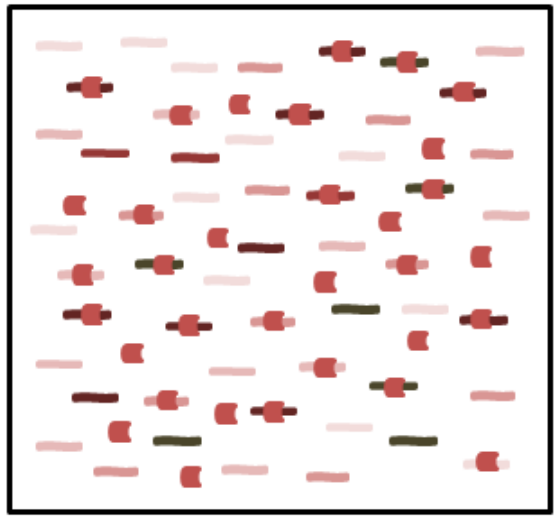
Telomerase Reverse Transcriptase (TERT) promoter mutations are the most frequent non-coding mutations in cancer, occur in over 50 types of cancer and are very common in some (80% melanoma, 70% glioblastoma…)
Create an ETS binding site that leads to high expression of the TERT protein, facilitating cancer growth.



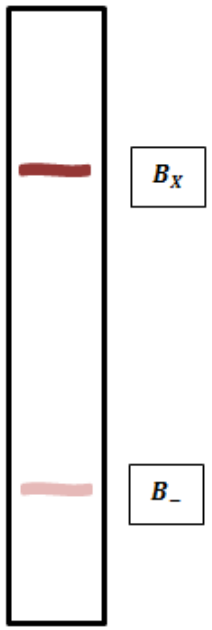
Nagore et al, Oncotarget, 2019, Vol. 10: 1546-1548

Spec-seq: Specificity by sequencing

A.



B.



$$\mathbf{P} + S_i \leftrightarrow \mathbf{P} \cdot S_i$$

$$K_A(S_i) = \frac{[\mathbf{P} \cdot S_i]}{[P][S_i]}$$

$$K_A(S_1) : K_A(S_2) : \ldots : K_A(S_n)$$

$$= \frac{[\mathbf{P} \cdot S_1]}{[S_1]} : \frac{[\mathbf{P} \cdot S_2]}{[S_2]} : \ldots : \frac{[\mathbf{P} \cdot S_n]}{[S_n]}$$

Specificity of the Lac repressor

Wild-type sequences

O1: **AATTGTGAG CGG ATAACAATT**

Randomized libraries

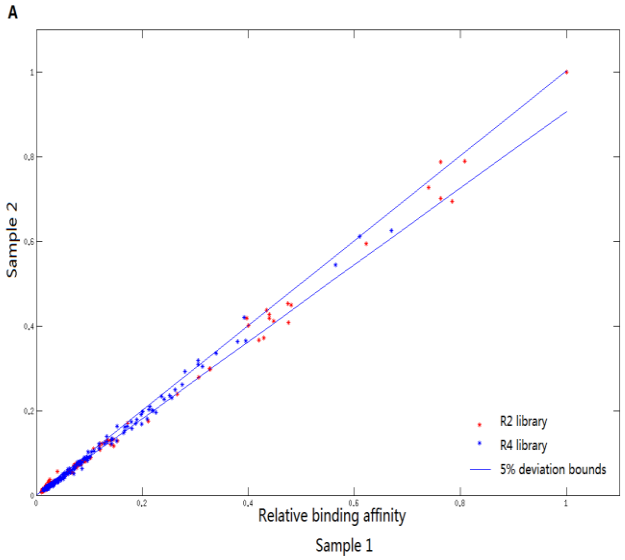
R2: **AATTGTNAN C G NTNACAATT**

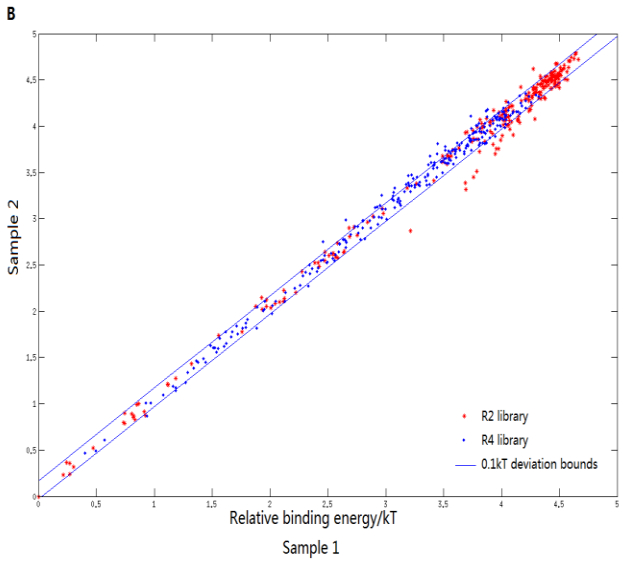
R3.1: **AATTGTGAG C N G NNNNCAATT**

R3.2: **AATTGNNNN C N G ATAACAATT**

R4: **AATTGTNAN CC GG NTNACAATT**

-10 -6 -1 -0 0 +0 +1 6 10





WT operator is asymmetric

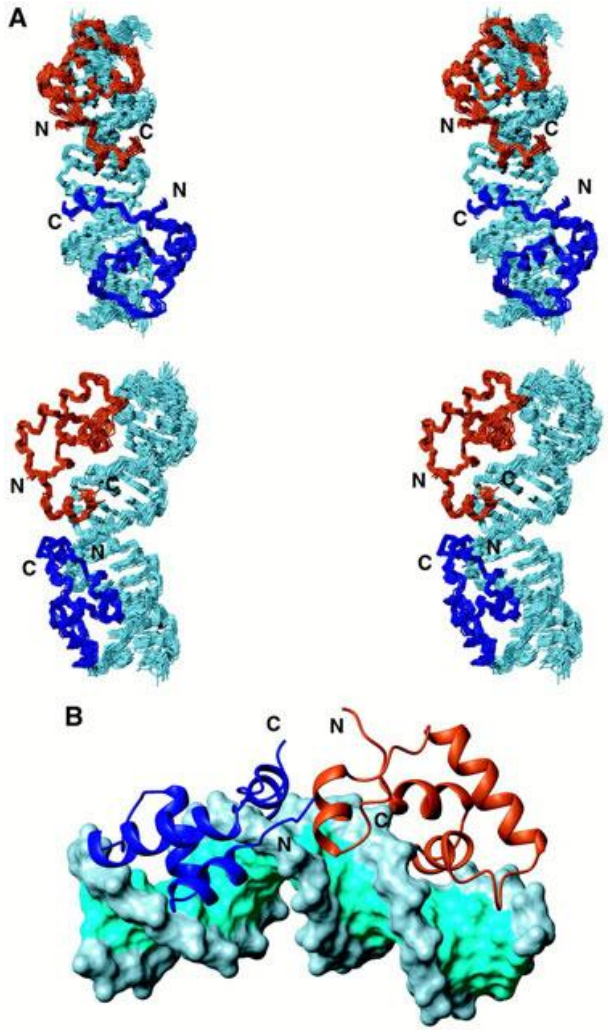
4 libraries: vary both sequence and spacing

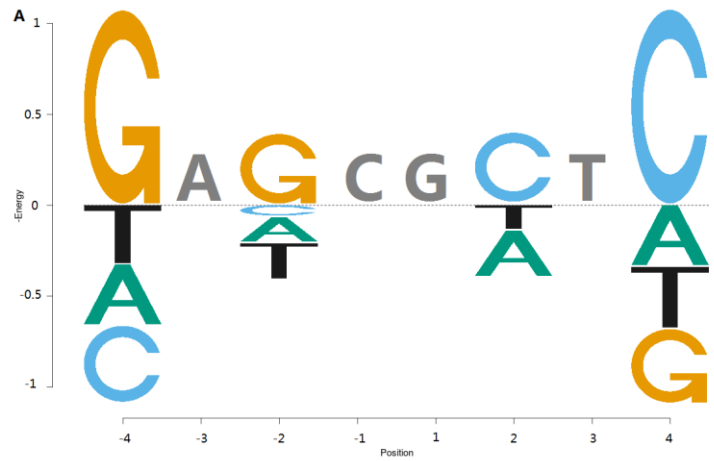
2560 different binding sites

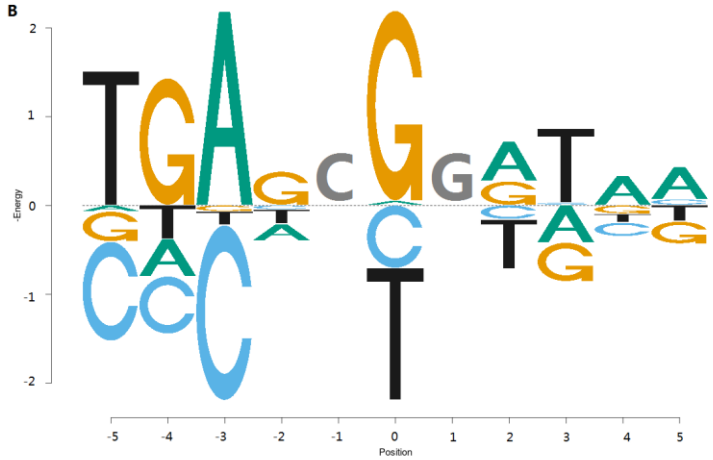
Highly reproducible:
~5% variance in affinity
~0.1kT variance in energy

Zuo, Stormo (2014) Genetics

Three-dimensional structure of the dimeric lac HP62–O1 operator complex.

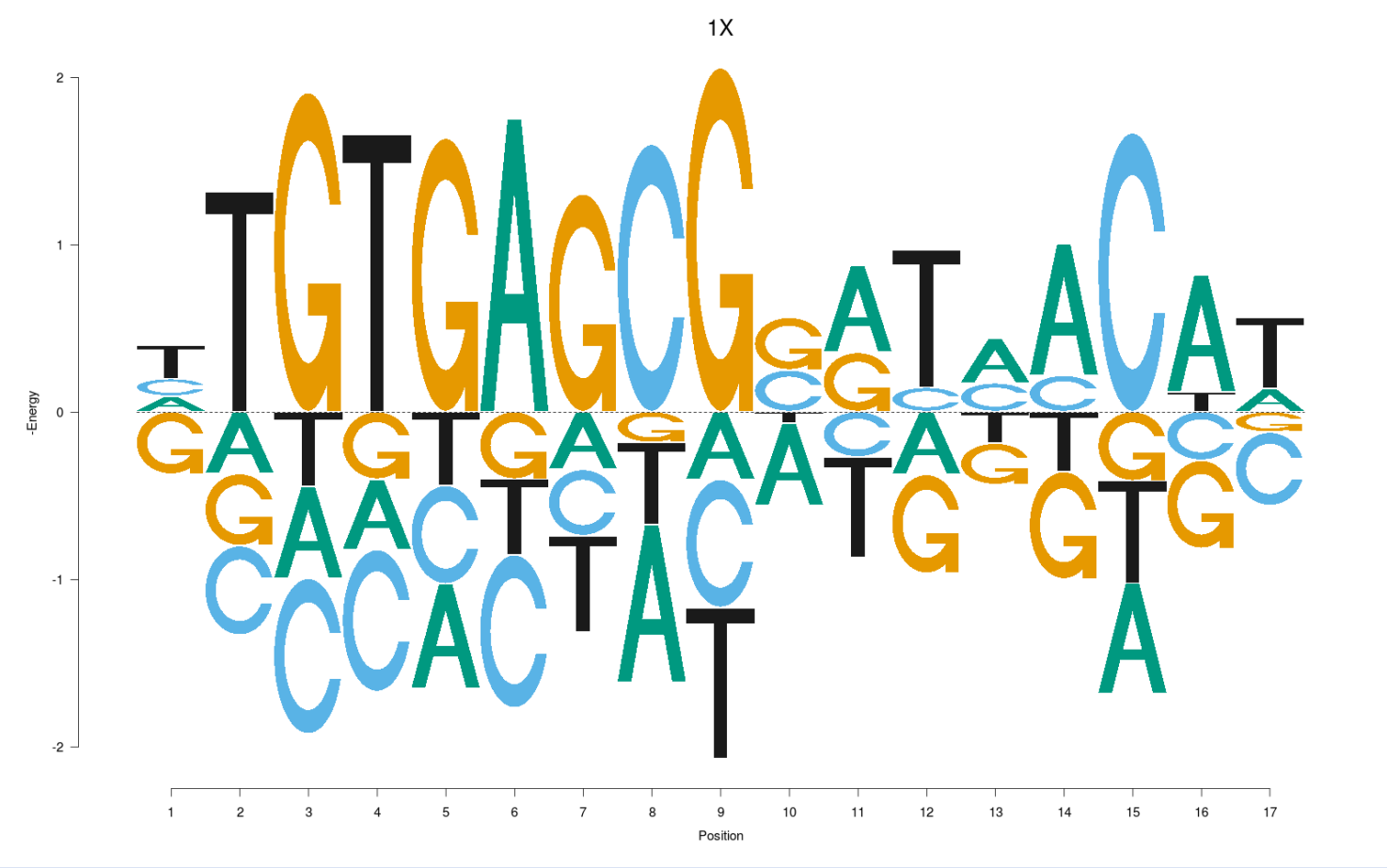




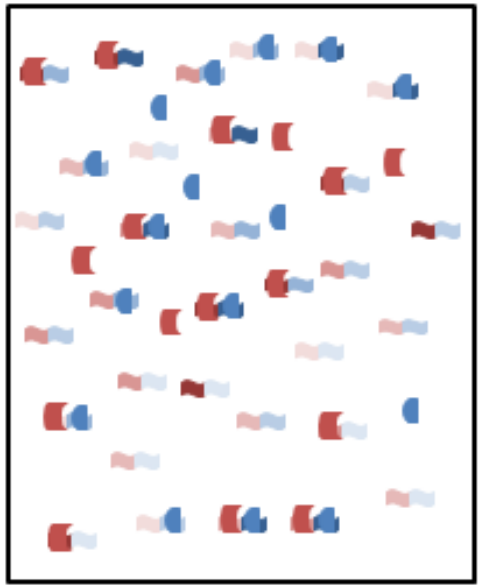


Kalodimos C G et al. EMBO J. 2002;21:2866-2876

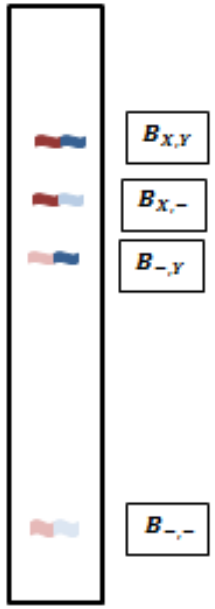
Specificity of complete “3 spacer” binding sites



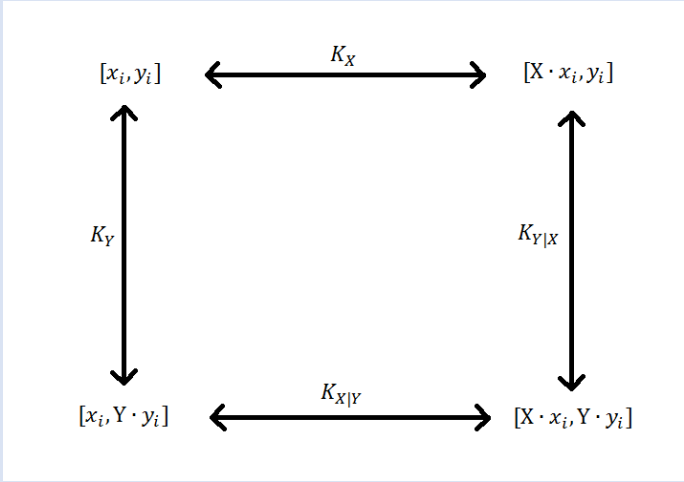
A.



B.



Coop-seq: get all of the important parameters in one experiment, including cooperativity



$$K_{\mathrm{X}}(x_1) : K_{\mathrm{X}}(x_2) : \ldots : K_{\mathrm{X}}(x_n) = \frac{N(x_1|B_{\mathrm{X},-})}{N(x_1|B_{-,-})} : \frac{N(x_2|B_{\mathrm{X},-})}{N(x_2|B_{-,-})} : \ldots : \frac{N(x_n|B_{\mathrm{X},-})}{N(x_n|B_{-,-})}$$

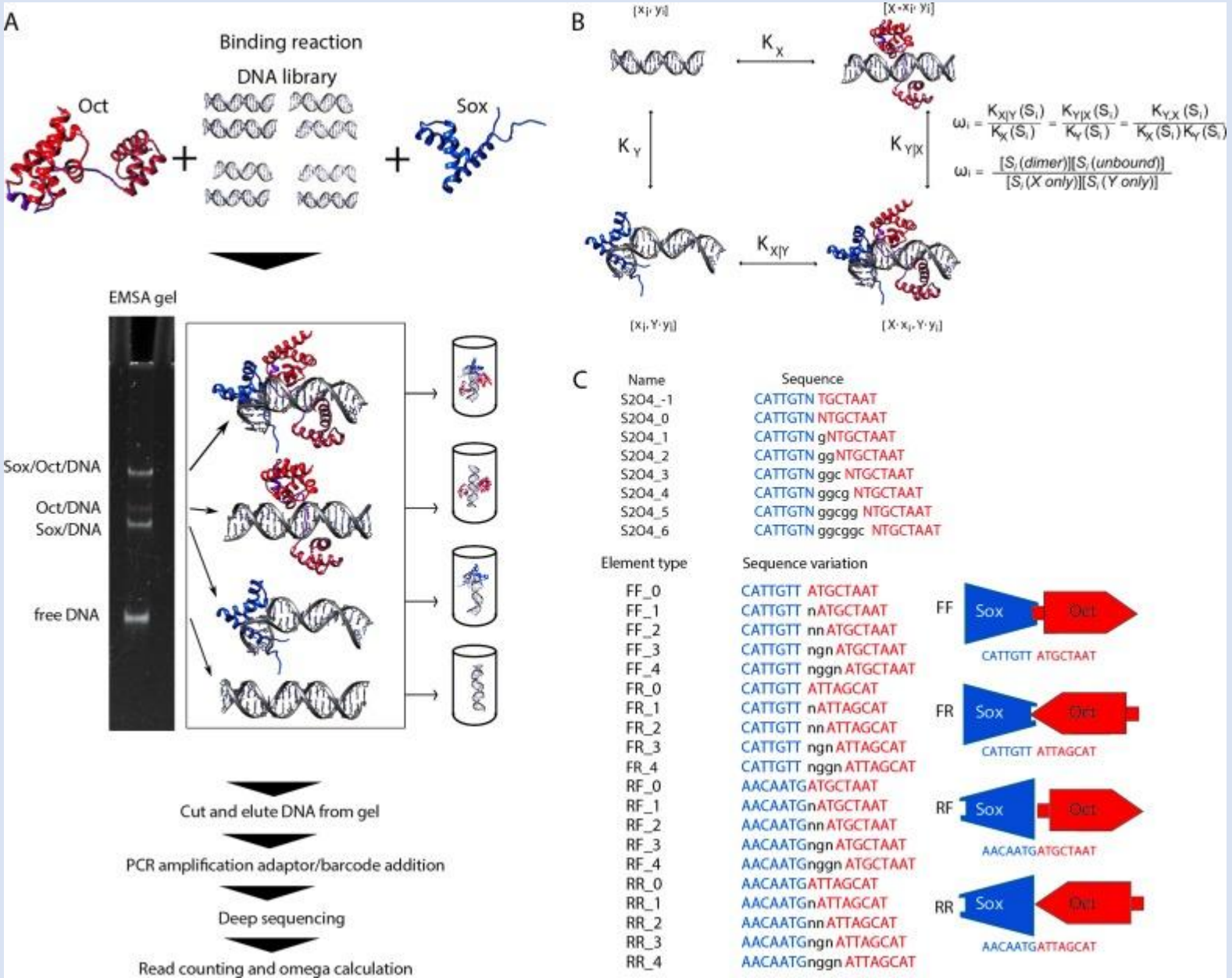
$$K_{\mathrm{X|Y}}(x_1) : K_{\mathrm{X|Y}}(x_2) : \ldots : K_{\mathrm{X|Y}}(x_n) = \frac{N(x_1|B_{\mathrm{X,Y}})}{N(x_1|B_{-,\mathrm{Y}})} : \frac{N(x_2|B_{\mathrm{X,Y}})}{N(x_2|B_{-,\mathrm{Y}})} : \ldots : \frac{N(x_n|B_{\mathrm{X,Y}})}{N(x_n|B_{-,\mathrm{Y}})}$$

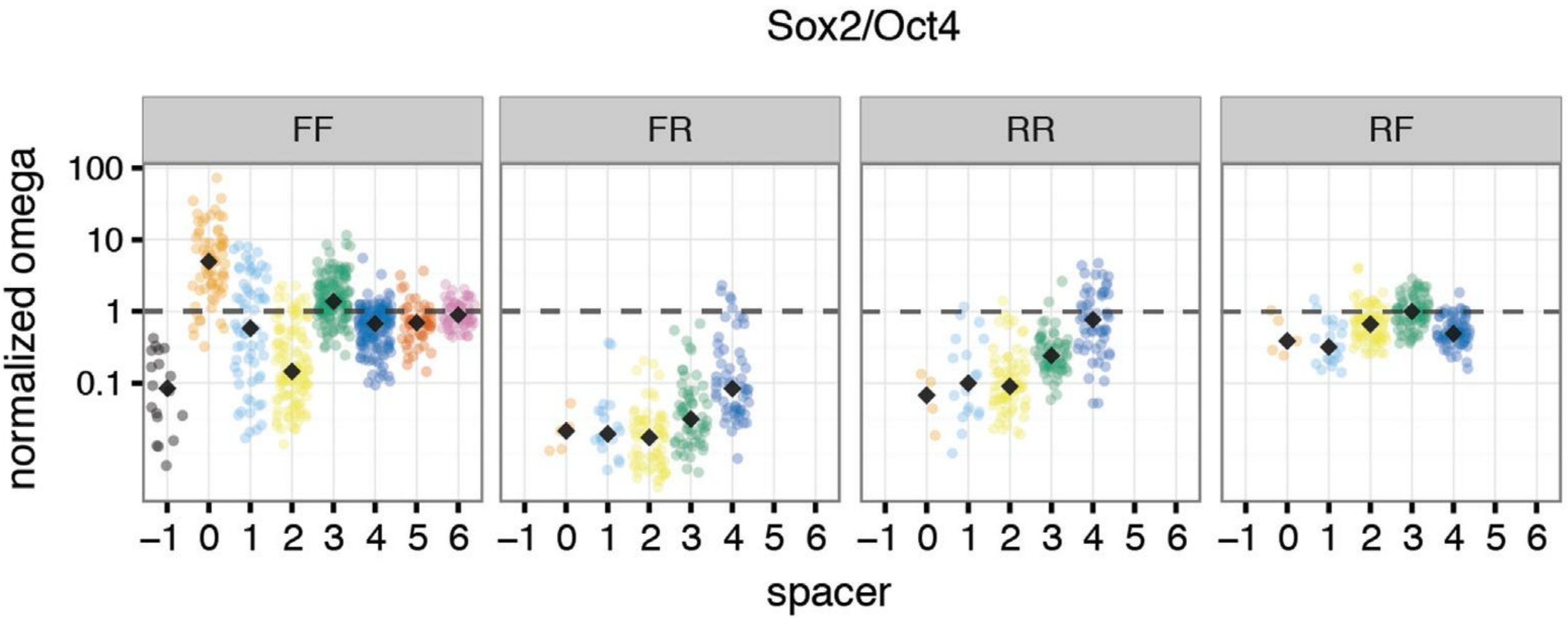
$$\omega_i = \frac{K_{\mathrm{X|Y}}(S_i)}{K_{\mathrm{X}}(S_i)} = \frac{K_{\mathrm{Y|X}}(S_i)}{K_{\mathrm{Y}}(S_i)} = \frac{K_{\mathrm{X,Y}}(S_i)}{K_{\mathrm{X}}(S_i) K_{\mathrm{Y}}(S_i)}$$

Stormo, Zuo, Chang, Briefings in Functional Genomics, 2015

Quantitative profiling of selective Sox/POU pairing on hundreds of sequences in parallel by Coop-seq

(NAR, 2017) Chang et al, collaboration with Ralf Jauch lab




Sox2/Oct4
FF
FR
RR
RF
normalized omega
100
10
1
0.1
−1 0 1 2 3 4 5 6
spacer

Aberrant homeodomain–DNA cooperative dimerization underlies distinct developmental defects in two dominant CRX retinopathy models

Yiqiao Zheng, Gary D. Stormo, and Shiming Chen

Genome Res. 2025 35(2):242-256.

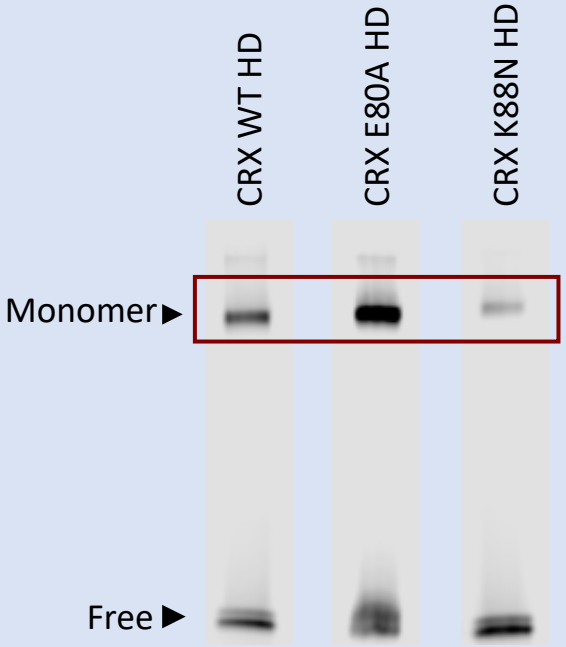
Two mutations in the CRX DNA-binding domain of the protein

E80A changes Glu at position 80 to Ala
K88N changes Lys at position 88 to Asn

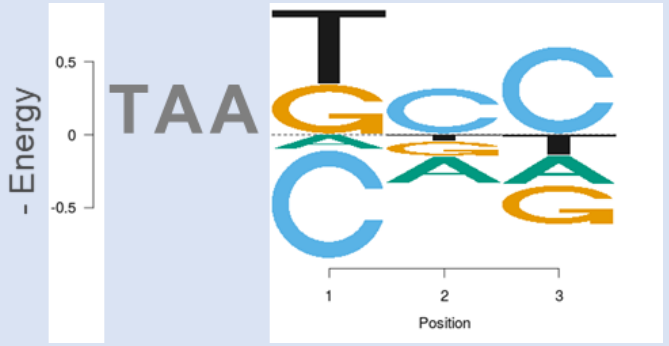
Spec-seq captures CRX DNA binding specificity changes at monomer motif

CRX WT consensus TAATCC

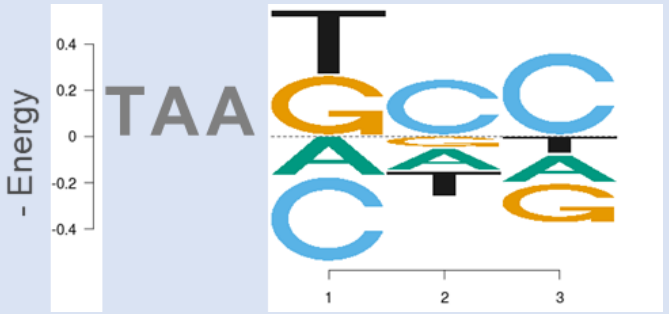
Monomer library … TAA NNN …



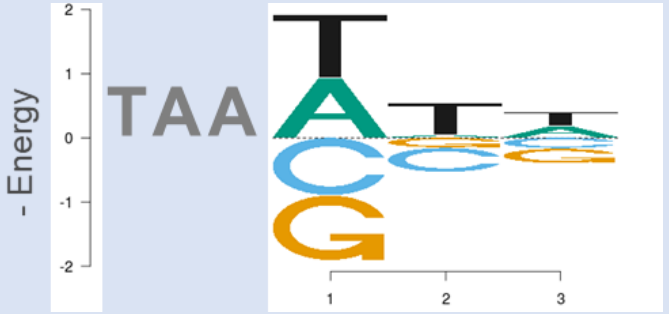
WT HD



E80A HD: **better tolerate nucleotide variants**



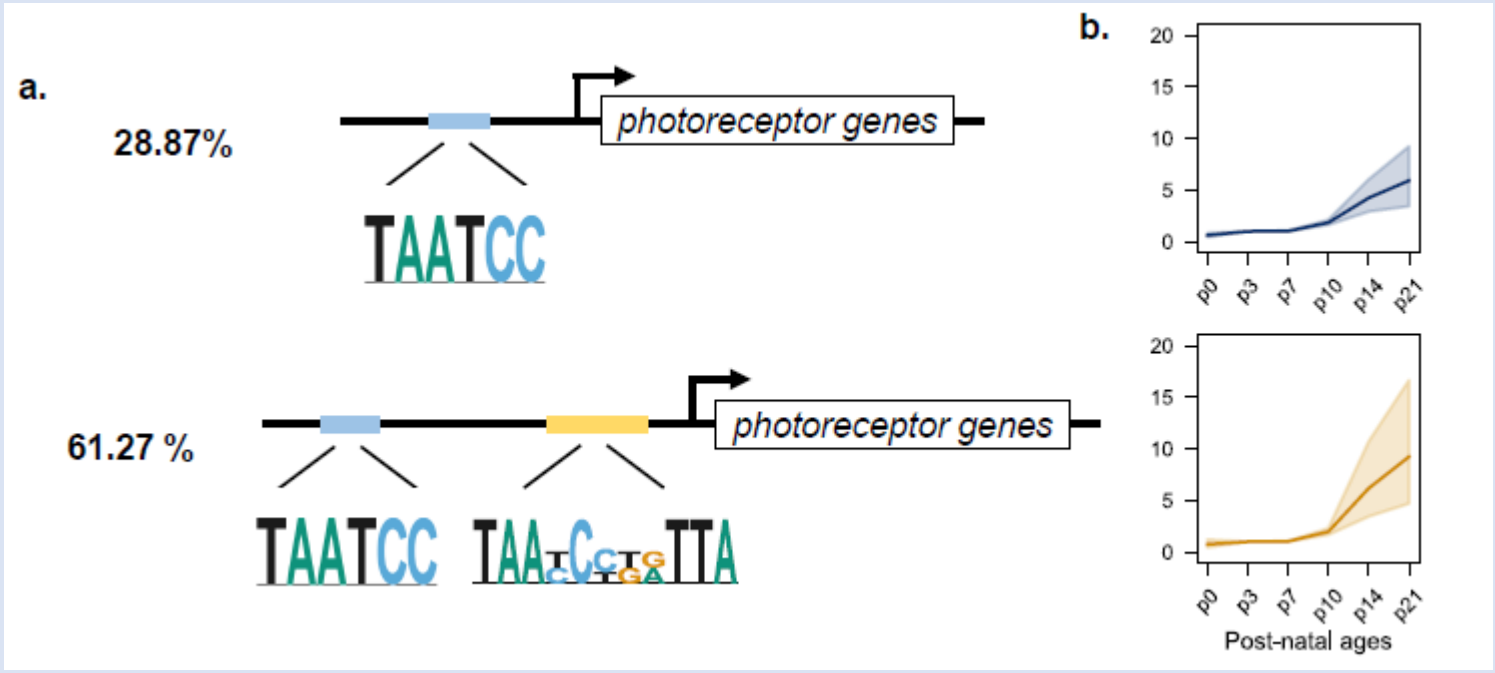
K88N HD: altered specificity

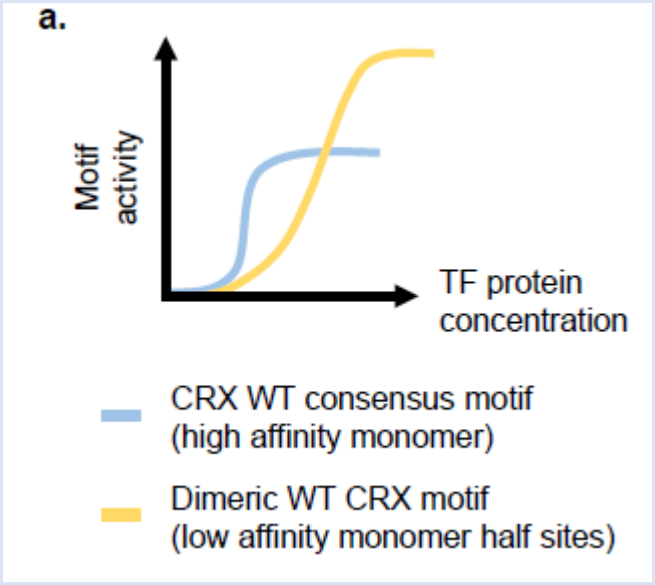


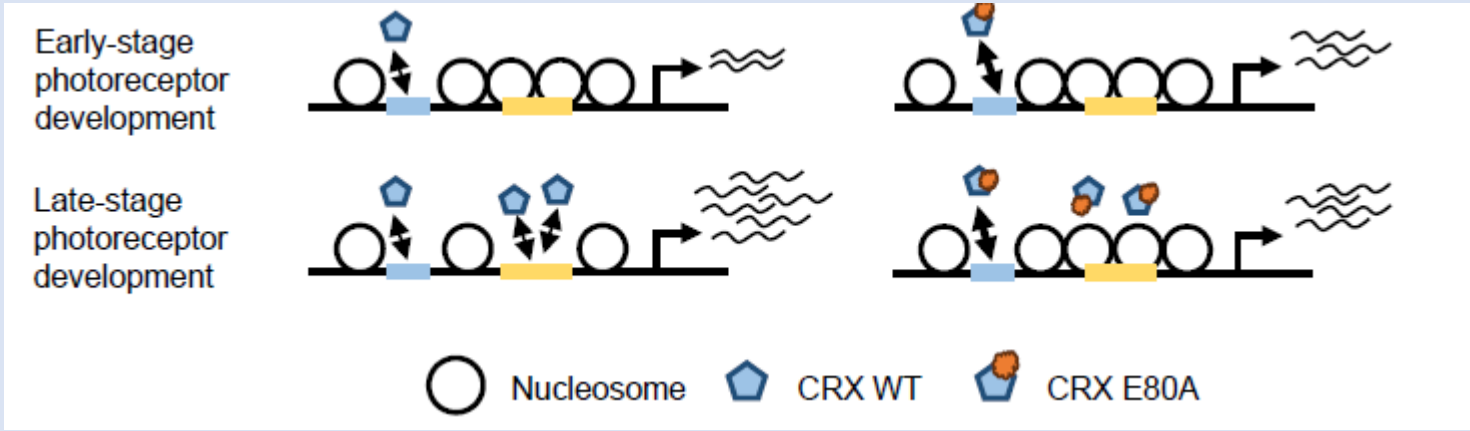
E80A K88N

Homeodomain

E80A has lost cooperativity







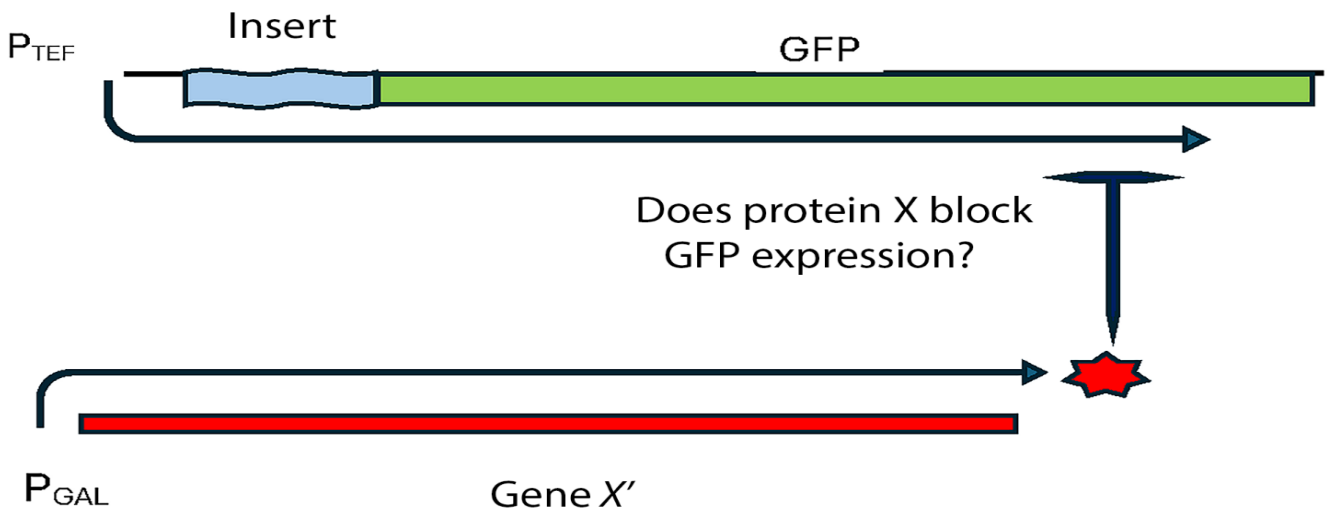
Regulation of Transcription Initiation is Ubiquitous
but post-transcriptional initiation is also common

Regulation can occur at
Transcription termination
RNA splicing
RNA translation
RNA degradation
Protein modification
Protein degradation

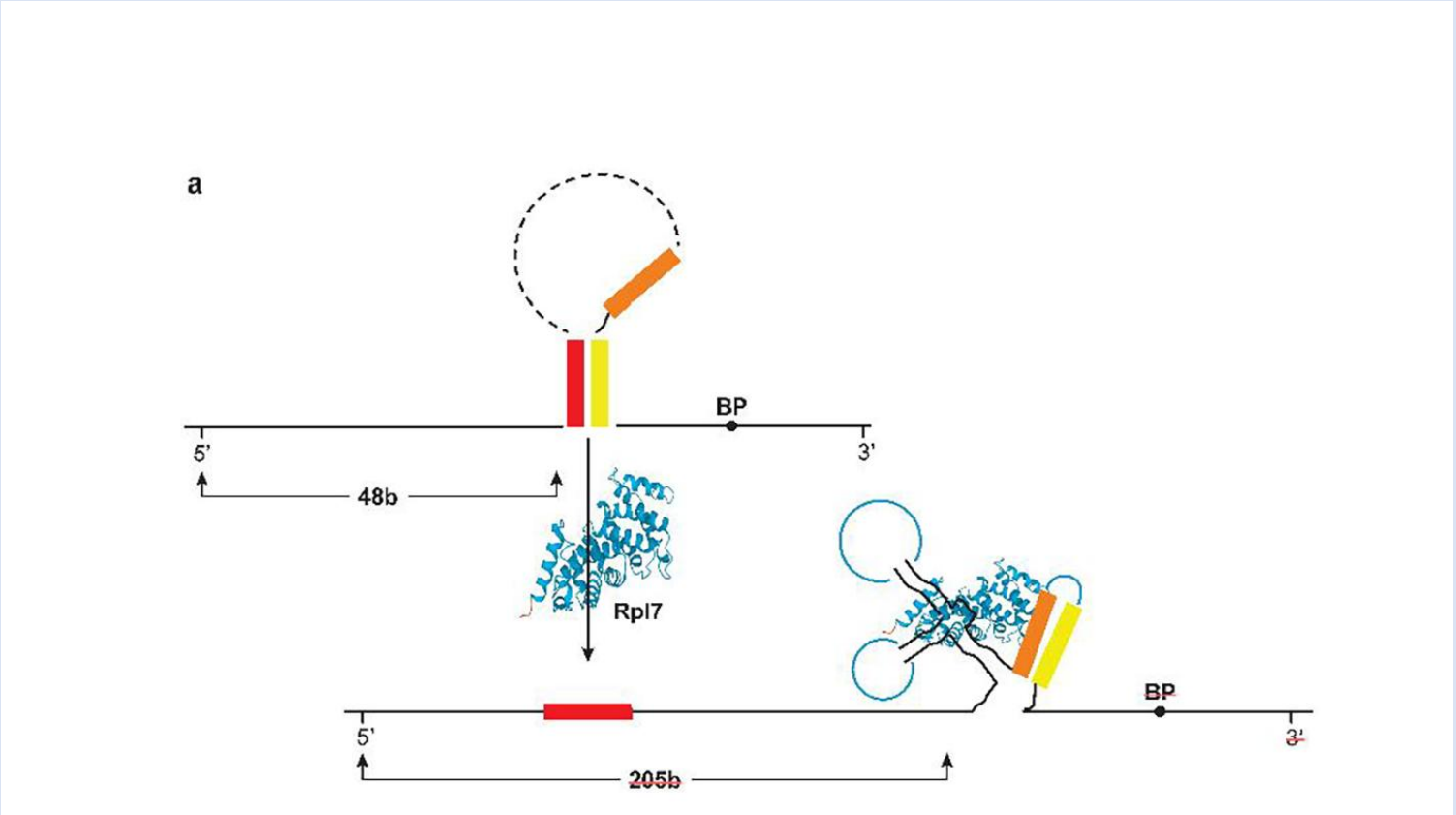
Auto-Regulation is of particular interest to me

A gene can control its own expression, setting its appropriate level (think cruise control)

Autoregulation of yeast ribosomal protein genes



Two alternative RNA structures. One serves as an enhancer to facilitate splicing and protein expression. The protein binds to the alternative structure to inhibit splicing and expression.



Granas et al, Nucleic Acids Res. 2025 53:gkaf739.

Acknowledgements

From the beginning:
Larry Gold
Andrzej Ehrenfeucht
Tom Schneider
Gerry Hertz, George Hartzell
Charles Lawrence, John Heumann
Alan Lapedes, Chris Workman
Sam Levy, Dana Fields, Eric Snyder

Wash U postdocs, students, research assistants, collaborators:

Fugen Li	Alan Kwan	Paul Taghert, Russ van Gelder
Takis Benos	Kai Tan	Jeff Milbrandt, Jeff Magee
Debraj GuhaThakurta	Yongmei Ji	Perren Cobb, Richard Hotchkiss, Mark Watson
Tao Zhao	Robin Matlib	Rafi Kopan, Buddy Brownstein, Tim Buchman
Yong Yin	Ritesh Agrawal	Jeff Gordon
Jiajian Liu	Liwei Chan	Charles Gu, Mike Province,
Guoyan Zhao	Xing Xu	Susan Dutcher
Burr Fontaine	Billy Li	Howard McLeod
Chris Man	Yiing Lin	Bob Waterston
Dana Homsi	Ting Wang	Alan Permutt, Jean Shaffer
Basab Roy	Yue Zhao	Mark Johnston, Barak Cohen
David Granas	Aaron Spivak	Jeremy Buhler, Weixong Zhang
Larry Schreifer	Ryan Christensen	Steve Beverley
Nnamdi Ihuegbu	Kenny Chang	Shin Imai, Christina Strong
Shuxiang Ruan	Zheng Zuo	Tim Ley, Jacqueline Payton,
Manishi Pandey	Shane Chu	Jim Havranek
Lala Motlhabi	Ronak Patel	Linda Sandell
Barrett Foat	Gurmukh Sahota	Ken Murphy, Xiaowei Wang

Outside collaborators:
Scot Wolfe, Chris Link, Javier Irazoqui, Ralf Jauch, Tony Gerber, Jan Gorodkin, Petko Petkov, Shiming Chen

Funding: NIH, DOE