Protein-DNA Interactions

CSE 591
WashU
Oct 10, 2018

Gary D. Stormo
Department of Genetics
Washington University in St.Louis

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

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

Outline of the talk

1. Basic Molecular Biology

Proteins bind DNA to regulate gene expression

Specificity of Transcription Factors (TFs)

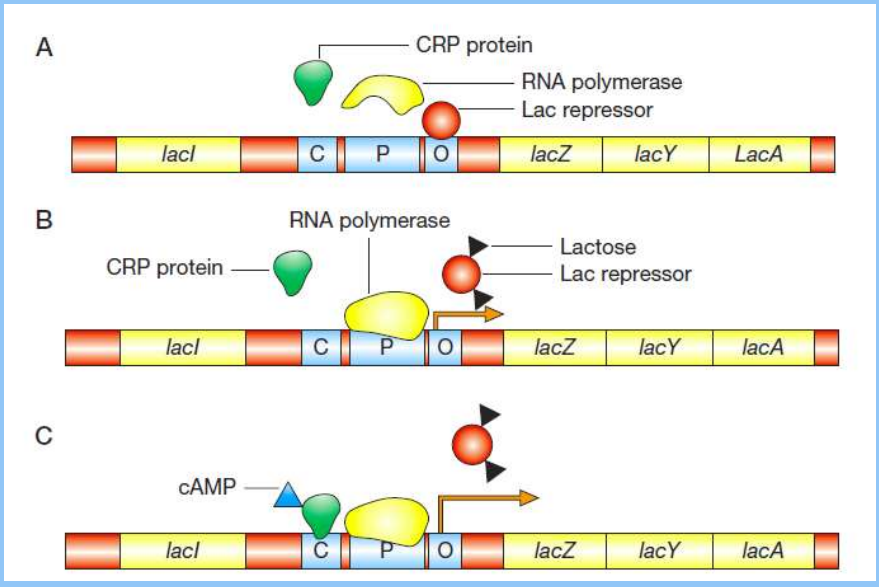
- Models and methods of determination
- Relationship to epigenetics

Open Computational Problems

Prediction of Specificity from Protein Sequence

- Prior work
- New challenges and possible approaches

Lactose regulatory system, Jacob and Monod, 1961



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

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

"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

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

TAATGTGAGTTAGCTCACTCAT
cgcTGTGAccgtGgTCgCagtT
tttTtTGAtcgtttTCaCattT
aaacgTGAtagccgTCaaacaa

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

TAATGTGAGTTAGCTCACTCAT
cgcTGTGAccgtGgTCgCagtT
tttTtTGAtcgtttTCaCattT
aaacgTGAtagccgTCaaacaa

After a few more examples, *nothing* was conserved!
Concept of "consensus sequence" emerged: have a preferred sequence but allow mismatches.
Still problematic, either low sensitivity or specificity.
New model was needed!

Representing TF Specificity with a Position Weight Matrix (PWM) Model
(aka: Weight Matrix, PSSM)

A:	-8	10	-1	2	1	-8
C:	-10	-9	-3	-2	-1	-12
G:	-7	-9	-1	-1	-4	-9
T:	10	-6	9	0	-1	11

PWM Model

Score = -24

….a	C	T	A	T	A	A	t	g …

A:	-8	10	**-1**	2	**1**	**-8**
C:	**-10**	-9	-3	-2	-1	-12
G:	-7	-9	-1	-1	-4	-9
T:	**10**	**-6**	9	**0**	-1	11

PWM Model

Score = 43

….a	c	T	A	T	A	A	T	g	t …

A:	-8	**10**	-1	**2**	**1**	-8
C:	-10	-9	-3	-2	-1	-12
G:	-7	-9	-1	-1	-4	-9
T:	**10**	-6	**9**	0	-1	**11**

A:	-8	10	-1	2	1	-8
C:	-10	-9	-3	-2	-1	-12
G:	-7	-9	-1	-1	-4	-9
T:	10	-6	9	0	-1	11

$$Score(S_i|W) = W \cdot S_i$$

PWM is a linear model:
- S_i encodes the sequence (which base occurs at each position)
- W weights those encoded features to provide the score
- Easy to add more features if they are necessary

A:	-8	10	-1	2	1	-8
C:	-10	-9	-3	-2	-1	-12
G:	-7	-9	-1	-1	-4	-9
T:	10	-6	9	0	-1	11

$$Score(S_i|W) = W \cdot S_i$$

PWM is a linear model:
- S_i encodes the sequence (which base occurs at each position)
- W weights those features to provide the score
- Easy to add more features if they are necessary

George Box: "All models are wrong.
Some models are useful."

Complete binding energy list vs model.

Note: Bioinformatics convention has higher scores for better binding sites. But lower energy corresponds to better binding sites.

Unfortunately both conventions are used in this talk, but it is usually clear which I am using at any time.

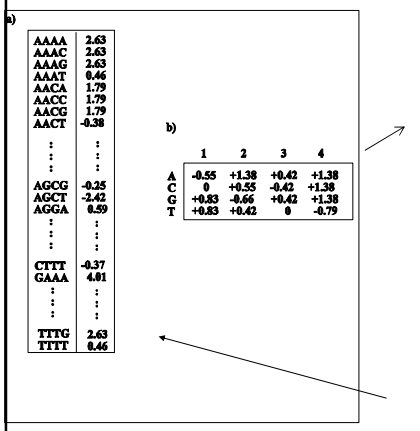
a)

AAAA	2.63
AAAC	2.63
AAAG	2.63
AAAT	0.46
AACA	1.79
AACC	1.79
AACG	1.79
AACT	-0.38
⋮	⋮
AGCG	-0.25
AGCT	-2.42
AGGA	0.59
⋮	⋮
CTTT	-0.37
GAAA	4.01
⋮	⋮
TTTG	2.63
TTTT	0.46

b)

	1	2	3	4
A	-0.55	+1.38	+0.42	+1.38
C	0	+0.55	-0.42	+1.38
G	+0.83	-0.66	+0.42	+1.38
T	+0.83	+0.42	0	-0.79

If simple additive model is inadequate, can use di-nucleotide or higher-order models. Some form of a matrix model must be correct because the binding data itself is a 1D matrix (vector).



	1	2	3
AA	0.83	0.42	1.38
AC	0	-0.42	1.38
AG	-1.21	0.42	1.38
AT	-0.13	0	-0.79
↓	↓	↓	↓
TT	1.25	0	-0.79

↓

	1	2
AAA	1.25	1.38
AAC	0.41	1.38
AAG	1.25	1.38
AAT	0.83	-0.79
↓	↓	↓
TTT	1.25	-0.79

Other extensions possible, such as structure-based parameters (Rohs, et al)

Parameter estimation

- Various methods for determining parameters:
  - Discriminant learning
  - Probabilistic modeling (i.e. log-odds)
    - Basis of most motif discovery algorithms
  - **Regression on quantitative data**
  - **Binding energy models**

Stormo (2013) Quantitative Biology 1:115-130

Probabilistic modeling based on known sites

PFM (PPM, PWM)

PWM (PSSM)

A.

A	9	214	63	142	118	8
C	22	7	26	31	52	13
G	18	2	29	38	29	5
T	193	19	124	31	43	216

N(b,i)

B.

A	0.04	0.88	0.26	0.59	0.49	0.03
C	0.09	0.03	0.11	0.13	0.22	0.05
G	0.07	0.01	0.12	0.16	0.12	0.02
T	0.80	0.08	0.51	0.13	0.18	0.89

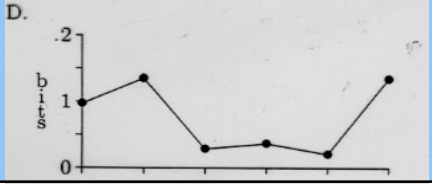
F(b,i)

C.

A	-2.76	1.82	0.06	1.23	0.96	-2.92
C	-1.46	-3.11	-1.22	-1.00	-0.22	-2.21
G	-1.76	-5.00	-1.06	-0.67	-1.06	-3.58
T	1.67	-1.66	1.04	-1.00	-0.49	1.84

$W(b,i) = \log[F(b,i)/P(b)]$

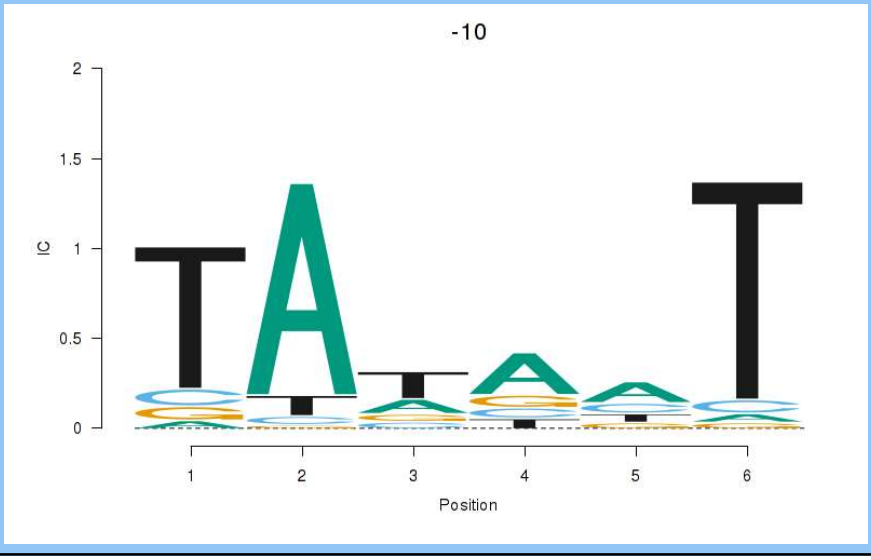
D.



$I(i) = \sum F(b,i)W(b,i)$

Motif discovery by Finding sites with max I

Classic Logo (from Tom Schneider):
Height of column at each position is Information Content
Each base in proportion to its frequency



Outline of motif discovery problem

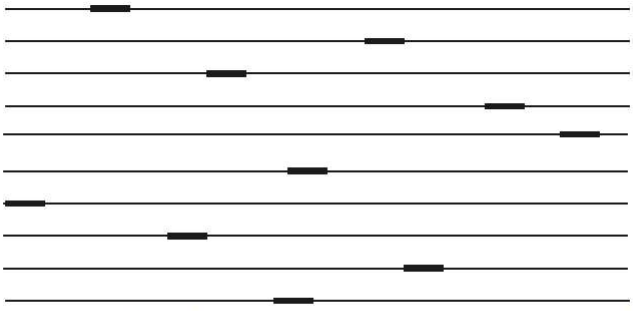


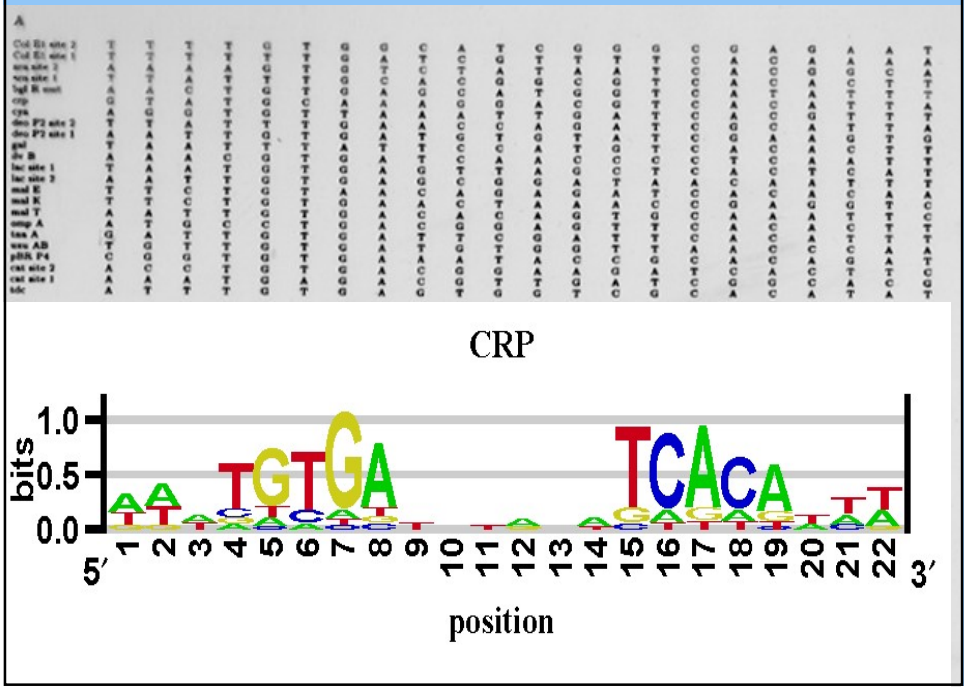
Fig. 6.1. A general schematic of the motif finding problem. Each *long thin line* represents a single DNA sequence. The *dark segments* within each line represent the binding sites whose positions are unknown in advance and we are trying to discover.

Genes regulated by CRP in *E. coli*

```
CE1CG
\TAATGTTTGTGCTGGTTTTTGTGGCATCGGGCGAGAATAGCGCGTGGTGTGAAAGACTGTTTTTTTGATCGTTTTCACAAAAATGGAAGTCCACAGTCTTGACAG\
ECOARABOP
\GACAAAAACGCGTAACAAAAGTGTCTATAATCACGGCAGAAAAGTCCACATTGATTATTTGCACGGCGTCACACTTTGCTATGCCATAGCATTTTTATCCATAAG\
ECOBGLR1
\ACAAATCCCAATAACTTAATTATTGGGATTTGTTATATATAACTTTATAAATTCCTAAAATTACACAAAGTTAATAACTGTGAGCATGGTCATATTTTTATCAAT\
ECOCRP
\CACAAAGCGAAAGCTATGCTAAAACAGTCAGGATGCTACAGTAATACATTGATGTACTGCATGTATGCAAAGGACGTCACATTACCGTGCAGTACAGTTGATAGC\
ECOCYA
\ACGGTGCTACACTTGTATGTAGCGCATCTTTCTTTACGGTCAATCAGCAAGGTGTTAAATTGATCACGTTTTAGACCATTTTTTCGTCGTGAAACTAAAAAAACC\
ECODEOP2
\AGTGAATTATTTGAACCAGATCGCATTACAGTGATGCAAACTTGTAAGTAGATTTCCTTAATTGTGATGTGTATCGAAGTGTGTTGCGGAGTAGATGTTAGAATA\
ECOGALE
\GCGCATAAAAAACGGCTAAATTCTTGTGTAAACGATTCCACTAATTTATTCCATGTCACACTTTTCGCATCTTTGTTATGCTATGGTTATTTCATACCATAAGCC\
ECOILVBPR
\GCTCCGGCGGGGTTTTTTGTTATCTGCAATTCAGTACAAAACGTGATCAACCCCTCAATTTTCCCTTTGCTGAAAAATTTTCCATTGTCTCCCCTGTAAAGCTGT\
ECOLAC
\AACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCAC\
ECOMALBA
\ACATTACCGCCAATTCTGTAACAGAGATCACACAAAGCGACGGTGGGGCGTAGGGGCAAGGAGGATGGAAAGAGGTTGCCGTATAAAGAAACTAGAGTCCGTTTA\
ECOMALBA
\GGAGGAGGCGGGAGGATGAGAACACGGCTTCTGTGAACTAAACCGAGGTCATGTAAGGAATTTCGTGATGTTGCTTGCAAAAATCGTGGCGATTTTATGTGCGCA\
ECOMALT
\GATCAGCGTCGTTTTAGGTGAGTTGTTAATAAAGATTTGGAATTGTGACACAGTGCAAATTCAGACACATAAAAAAACGTCATCGCTTGCATTAGAAAGGTTTCT\
ECOOMPA
\GCTGACAAAAAAGATTAAACATACCTTATACAAGACTTTTTTTTCATATGCCTGACGGAGTTCACACTTGTAAGTTTTCAACTACGTTGTAGACTTTACATCGCC\
ECOTNAA
\TTTTTTAAACATTAAAATTCTTACGTAATTTATAATCTTTAAAAAAAGCATTTAATATTGCTCCCCGAACGATTGTGATTCGATTCACATTTAAACAATTTCAGA\
ECOUXU1
\CCCATGAGAGTGAAATTGTTGTGATGTGGTTAACCCAATTAGAATTCGGGATTGACATGTCTTACCAAAAGGTAGAACTTATACGCCATCTCATCCGATGCAAGC\
PBR322
\CTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGAGTGCACCATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGGCGCTC\
TRN9CAT
\CTGTGACGGAAGATCACTTCGCAGAATAAATAAATCCTGGTGTCCCTGTTGATACCGGGAAGCCCTGGGCCAACTTTTGGCGAAAATGAGACGTTGATCGGCACG\
TDC
\GATTTTTATACTTTAACTTGTTGATATTTAAAGGTATTTAATTGTAATAACGATACTCTGGAAAGTATTGAAAGTTAATTTGTGAGTGGTCGCACATATCCTGTT\
```

Stormo and Hartzell, 1989, PNAS

Output: sites, logo



Inherent limitations of probabilistic models for protein-DNA binding specificity

Shuxiang Ruan, Gary D. Stormo*

PLOS COMPUTATIONAL BIOLOGY

July 7, 2017

Binding probabilities depend on the protein concentration

Positions are normalized independently, leading to apparent non-independence and mis-ordering of probabilities

Biophysical (energy) models are preferred

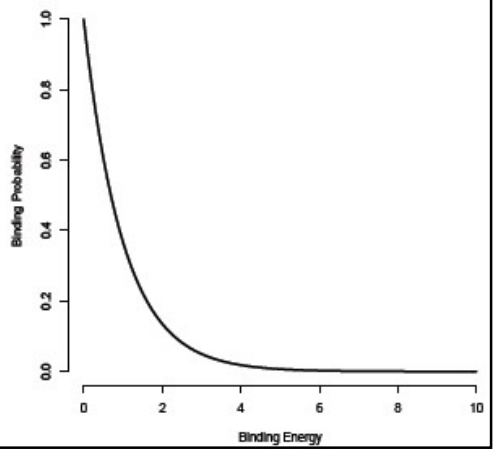
Log-odds method is equivalent to an energy model if the sites are from a Boltzmann distribution with binding probability $\propto e^{-E}$

posterior prior energy

$$F(S_i) = P(S_i)e^{-E_i} / Z$$

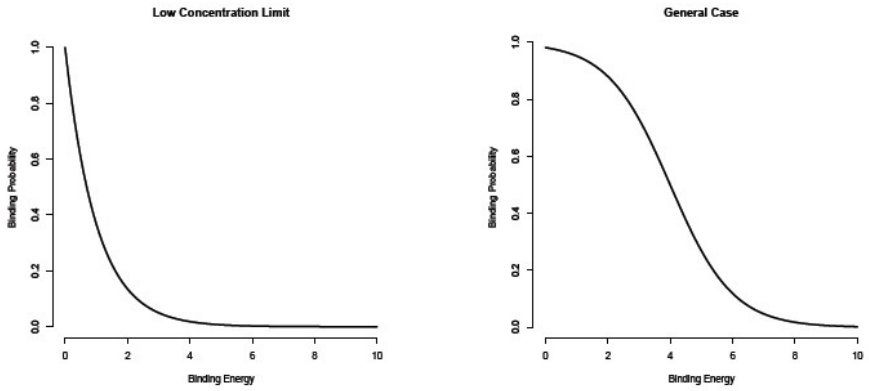
$$E_i = -\ln \frac{F(S_i)}{P(S_i)}$$

Log-odds relationship between **binding energy** and **probabilities**



Reality is a Fermi-Dirac distribution with Boltzmann a special case at the low concentration range

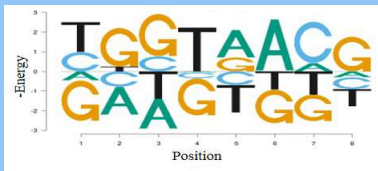
$$P(binding|S_i) = \frac{1}{1+e^{E(S_i)-\mu}}$$



Djordjevic et al, Genome Res. 2003 13:2381-90.

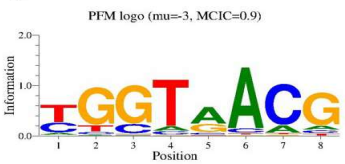
Limitations of Probabilistic Models

- PFMs depend on the protein concentration



Base	Position							
	1	2	3	4	5	6	7	8
A	1.98	3.37	3.66	2.25	0	0	1.69	2.04
C	0.55	2.73	1.39	2.50	2.07	2.56	0	2.35
G	3.56	0	0	4.09	0.68	4.28	3.43	0
T	0	1.73	3.53	0	2.78	3.51	3.27	2.60

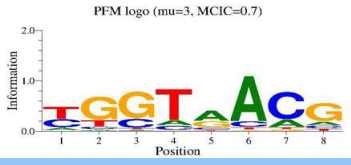
c



d

Base	Position							
	1	2	3	4	5	6	7	8
A	8	3	2	9	59	89	15	10
C	33	5	19	7	7	7	79	7
G	2	78	77	1	30	1	3	77
T	57	14	2	83	4	3	3	6

e



f

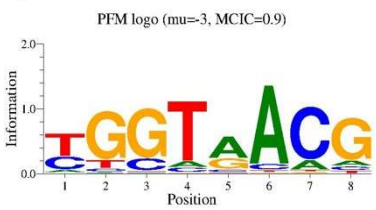
Base	Position							
	1	2	3	4	5	6	7	8
A	11	5	3	13	51	82	20	14
C	34	8	24	10	11	11	71	11
G	3	68	69	2	32	2	4	66
T	52	19	4	75	6	5	5	9

Ruan and Stormo (2017) PLoS Comp Bio

Limitations of PFMs

- Positions are normalized independently, leading to
- Apparent non-independence

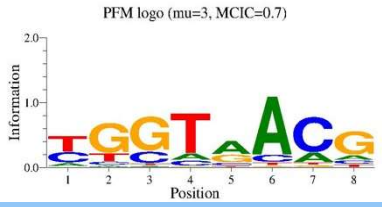
c



d

Base	Position							
	1	2	3	4	5	6	7	8
A	8	3	2	9	59	89	15	10
C	33	5	19	7	7	7	79	7
G	2	78	77	1	30	1	3	77
T	57	14	2	83	4	3	3	6

e



f

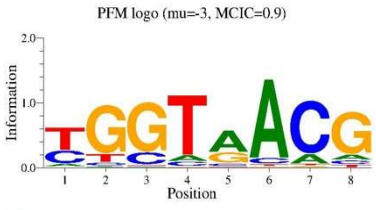
Base	Position							
	1	2	3	4	5	6	7	8
A	11	5	3	13	51	82	20	14
C	34	8	24	10	11	11	71	11
G	3	68	69	2	32	2	4	66
T	52	19	4	75	6	5	5	9

$< \Delta > = 2.5$

Limitations of PFMs

- Positions are normalized independently, leading to
- Apparent non-independence

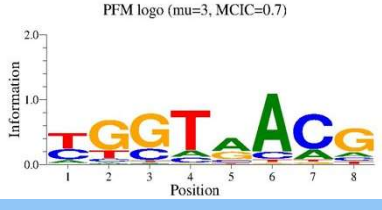
c



d

Base	Position							
	1	2	3	4	5	6	7	8
A	8	3	2	9	59	89	15	10
C	33	5	19	7	7	7	79	7
G	2	78	77	1	30	1	3	77
T	57	14	2	83	4	3	3	6

e



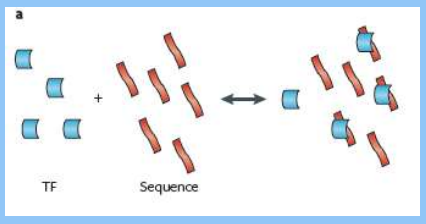
f

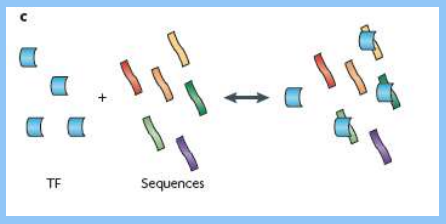
Base	Position							
	1	2	3	4	5	6	7	8
A	11	5	3	13	51	82	20	14
C	34	8	24	10	11	11	71	11
G	3	68	69	2	32	2	4	66
T	52	19	4	75	6	5	5	9

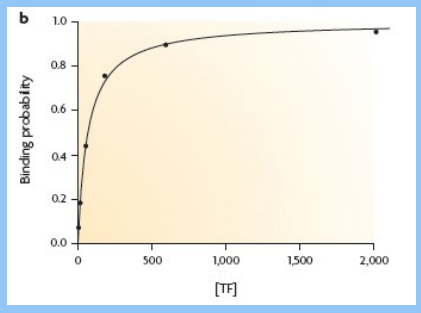
$< \Delta > = 5$

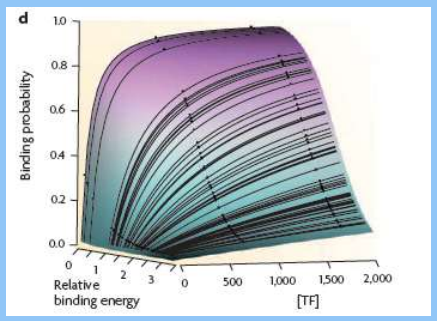
Measuring Specificity

(vs Affinity)

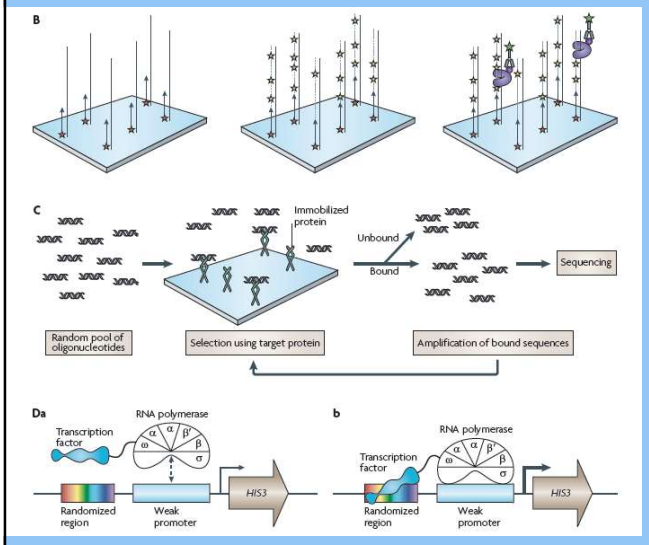




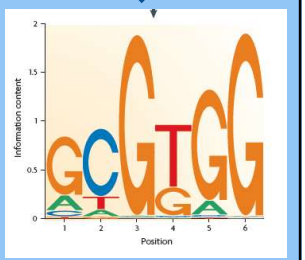




Modeling Specificity from high-throughput methods



Specificity Modeling



Stormo and Zhao, Nature Reviews Genetics, 2010

Diversity and Complexity in DNA Recognition by Transcription Factors
Gwenael Badis *et al.*
Science **324**, 1720 (2009);

Diverse sets:

>100 TFs

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

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

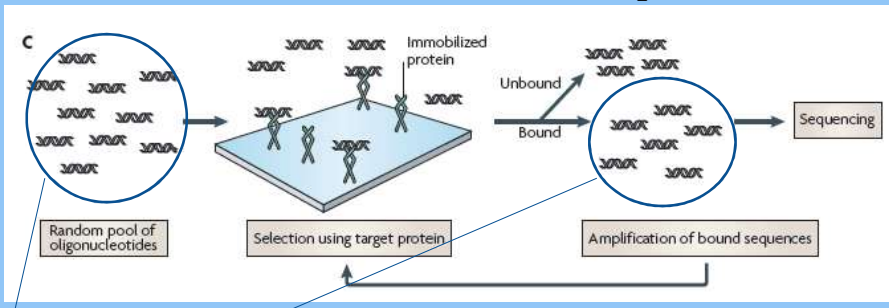
Shuxiang Ruan[1], S. Joshua Swamidass[2] and Gary D. Stormo[1,*]

Bioinformatics, 33(15), 2017, 2288–2295

Uses Expectation Maximization (EM) to simultaneously infer the binding site on each sequence and the parameters of the model (PWM)

Out performs all other algorithms on *in vitro* data, Comparable on *in vivo* data (ChIP-seq)

HT-SELEX (SELEX-Seq)

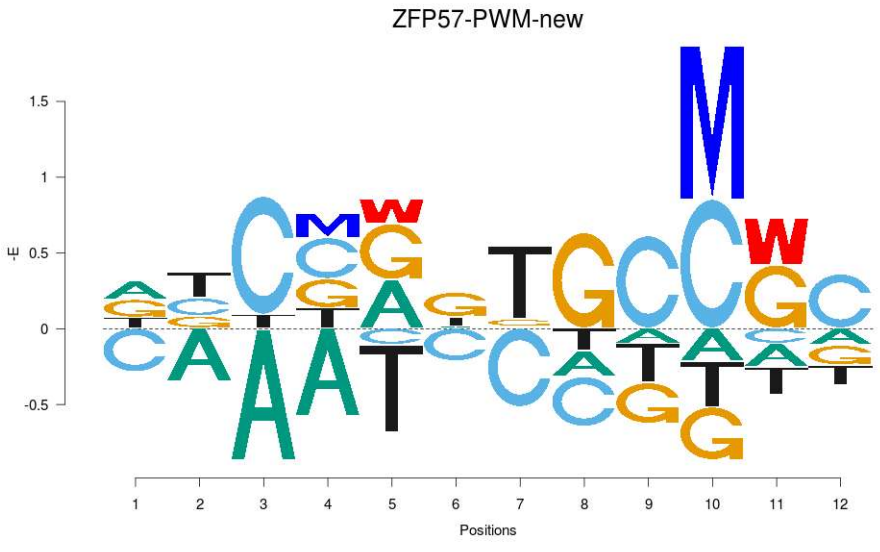


$$\frac{P(S_i|b)}{P(S_i)} \propto {}^{1}/_{(1+e^{E_i-\mu})}$$

Compared to reference sequence with E = 0

$$\frac{\frac{P(S_i|b)}{P(S_i)}}{\frac{P(S_{ref}|b)}{P(S_{ref})}} = \frac{(1+e^{-\mu})}{(1+e^{E_i-\mu})}$$

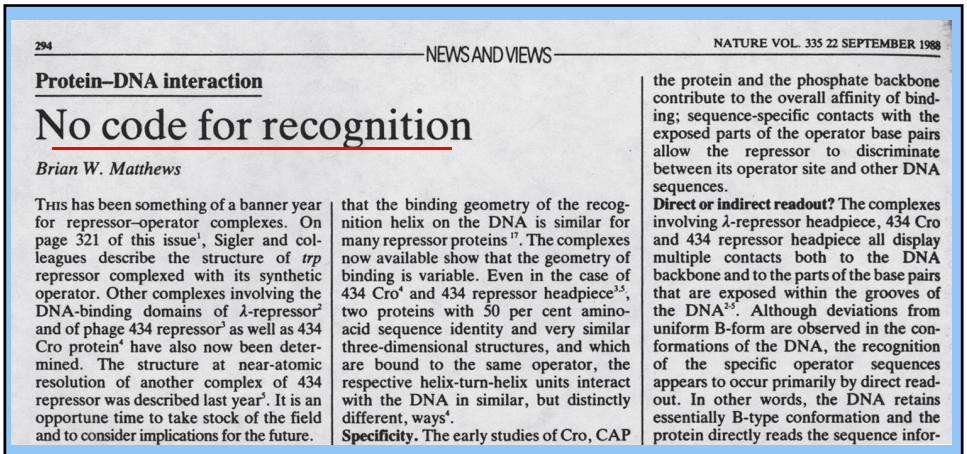
Can easily measure effects of methylation: M = mC; W = mC on opposite strand



Zuo et al, Sciences Advances, 2017

Open Problem: Specificity Prediction

- Goal:
  - Predict specificity from protein sequence
  - Decipher "Recognition Code" for Protein-DNA
- Strategy:
  - Infer interacting positions from covariation
  - Use machine learning methods (SVM,RF,KNN,NN,…) to develop predictive methods
- Evaluation:
  - Cross-validation



294 NEWS AND VIEWS NATURE VOL. 335 22 SEPTEMBER 1988

Protein–DNA interaction

No code for recognition

Brian W. Matthews

THIS has been something of a banner year for repressor–operator complexes. On page 321 of this issue[1], Sigler and colleagues describe the structure of *trp* repressor complexed with its synthetic operator. Other complexes involving the DNA-binding domains of λ-repressor[2] and of phage 434 repressor[3] as well as 434 Cro protein[4] have also now been determined. The structure at near-atomic resolution of another complex of 434 repressor was described last year[5]. It is an opportune time to take stock of the field and to consider implications for the future.

that the binding geometry of the recognition helix on the DNA is similar for many repressor proteins[1-7]. The complexes now available show that the geometry of binding is variable. Even in the case of 434 Cro[4] and 434 repressor headpiece[3-5], two proteins with 50 per cent amino-acid sequence identity and very similar three-dimensional structures, and which are bound to the same operator, the respective helix-turn-helix units interact with the DNA in similar, but distinctly different, ways[4].

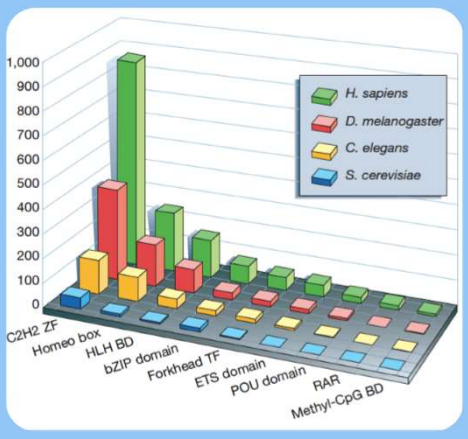
Specificity. The early studies of Cro, CAP

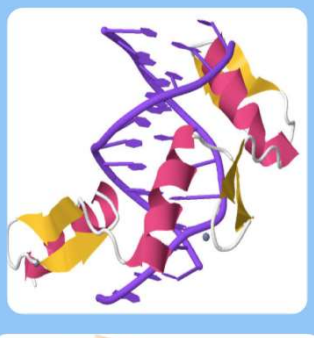
the protein and the phosphate backbone contribute to the overall affinity of binding; sequence-specific contacts with the exposed parts of the operator base pairs allow the repressor to discriminate between its operator site and other DNA sequences.

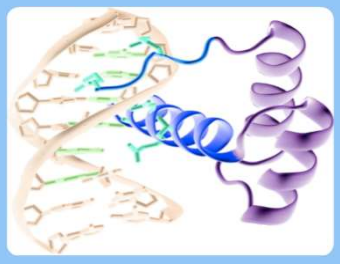
Direct or indirect readout? The complexes involving λ-repressor headpiece, 434 Cro and 434 repressor headpiece all display multiple contacts both to the DNA backbone and to the parts of the base pairs that are exposed within the grooves of the DNA[2-5]. Although deviations from uniform B-form are observed in the conformations of the DNA, the recognition of the specific operator sequences appears to occur primarily by direct readout. In other words, the DNA retains essentially B-type conformation and the protein directly reads the sequence infor-

But this only ruled out a universal, deterministic code, which we already knew didn't exist

Two Largest TF Families

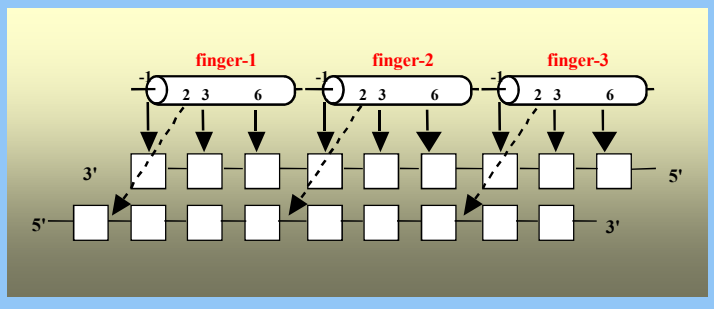






Tupler et al (2001) *Nature* **409**: 832-833

EGR family of transcription factors



Zn-finger (Cys_2His_2):

- 3 fingers, binding in a modular fashion
- target site: 4 bases long (for each finger)
- one base overlap in the target of each finger

Qualitative models...

- Usually in the form of a simple "binary" table.
- Difficult to expand to other than "one-to-one" model.
- By nature, unsuitable for quantitative predictions.

	Position in triplet		
	5'	Middle	3'
A	Gln 6	Asn 3 Ser 3 His 3	Gln -1
C	Ser 2	Asp 3 Thr 3 Val 3	Asp -1
G	Arg 6 Lys 6 Asp 2 Ser 2 Phe 2	His 3 Lys 3	Arg -1
T	Lys 2 Asp 2	Thr 3 Ala 3 Ser 3 Val 3	Leu -1 Thr -1 Asn -1

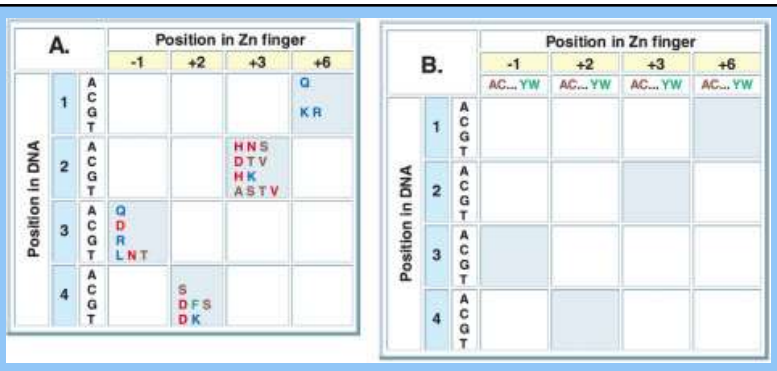
doi:10.1016/S0022-2836(02)00917-8 available online at http://www.idealibrary.com on IDEAL J. Mol. Biol. (2002) 323, 701–727

JMB

Probabilistic Code for DNA Recognition by Proteins of the EGR Family

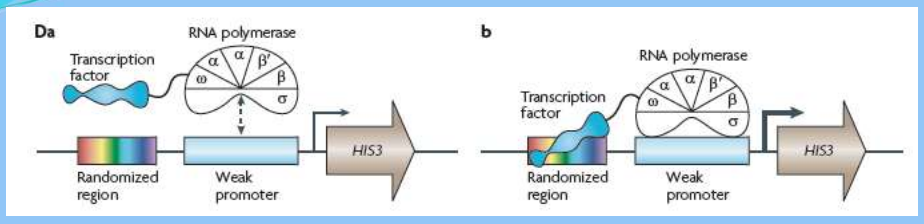
Panayiotis V. Benos[1], Alan S. Lapedes[2] and Gary D. Stormo[1]*

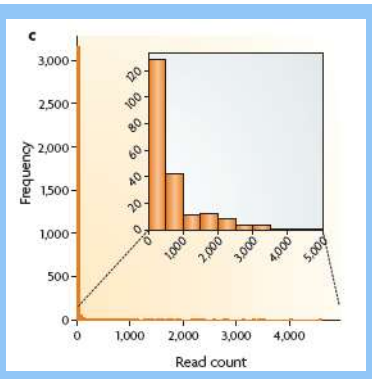
Estimate an energy model for zinc finger proteins based on a limited set of qualitative data:
Protein-DNA pairs from SELEX and phage-display Experimental datasets

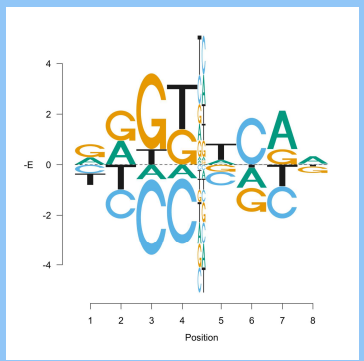


- **If** we knew the energy parameters, we could calculate the probability of obtaining any particular combination (protein-DNA) in any specific experiment (SELEX or phage-display)
- Instead, we can find the parameters that maximize the probability of obtaining the combinations we observed

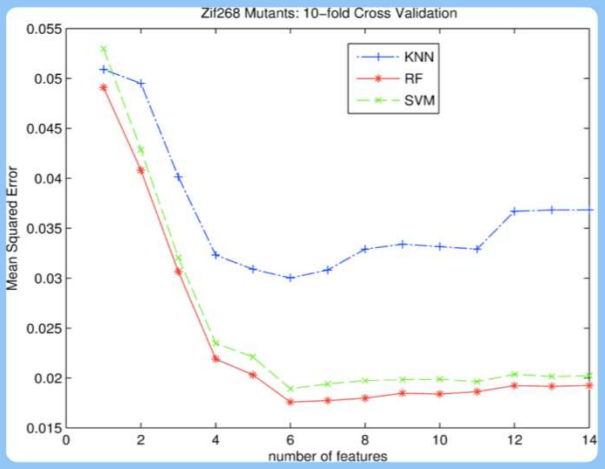
Bacterial-1-Hybrid (B1H)





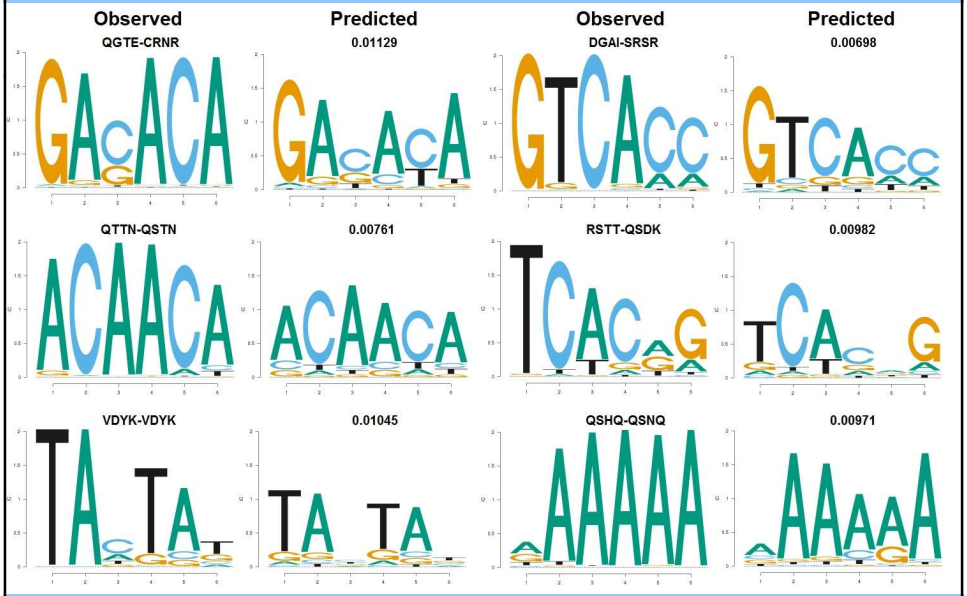


ZF: 10-fold Cross Validation



Feature Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Finger	3	3	2	2	2	3	2	2	3	3	3	2	3	2
Recognition Helix Position	-1	3	3	6	-1	2	2	1	1	6	5	5	4	4

Average Prediction Accuracy for ZFPs



http://stormo.wustl.edu/ZFNModels

Other Prior Work

- Mona Singh's group at Princeton
  - SVM, larger training set (2015)
- Tim Hughes' group at Univ Toronto
  - RF, larger and more diverse training set (2015)
- Brendon Frey's group at Univ Toronto
  - Deep learning on several TF families to get general predictors of protein specificity (2015)

Open Problems and Opportunities

- Larger, more diverse datasets now exist
  - Current predictions are much poorer on more diverse proteins – especially those with many ZFs of which only a subset may interact with DNA
  - Can we predict which fingers are used?
- We and others are collecting data about methylation sensitivity
  - No current models attempt to predict that
  - Are there "simple rules" for methylation sensitivity?
- Can Deep Learning on a more defined problem give both better predictions and mechanistic insights?

Acknowledgements

- Stormo Lab
- Dana Fields
- Takis Benos
- JJ Liu
- Ryan Christensen
- Yue Zhao
- Shuxiang Ruan
- Zheng Zuo
- David Granas

- Collaborators
- Alan Lapedes, Los Alamos Labs
- Scott Wolfe, Univ of Mass Med School
  - And several people from his lab
- Petko Petkov, Jackson Lab
- Funding: NIH