DNA binding proteins and motif analysis

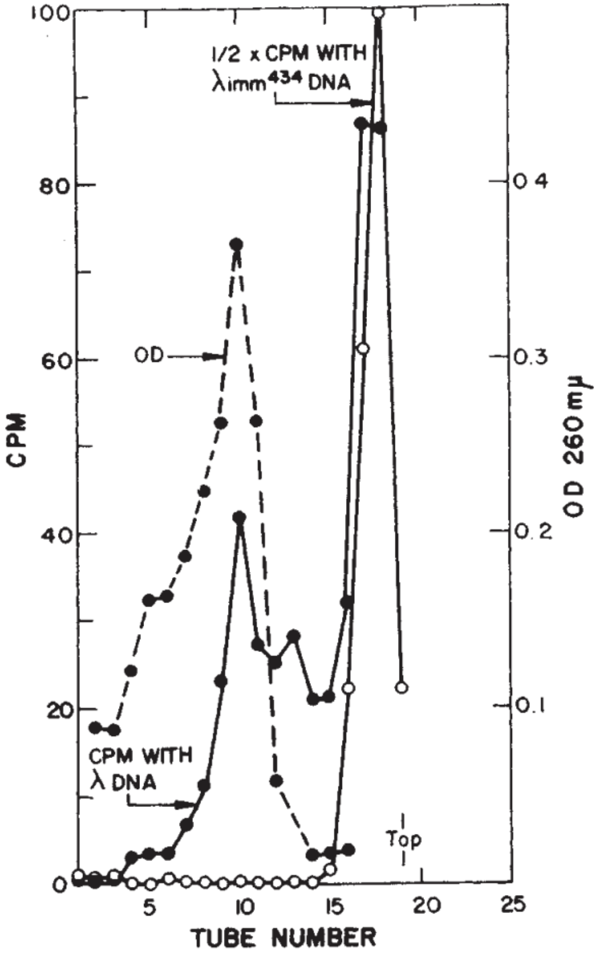
Bio 5488

Michael Meers

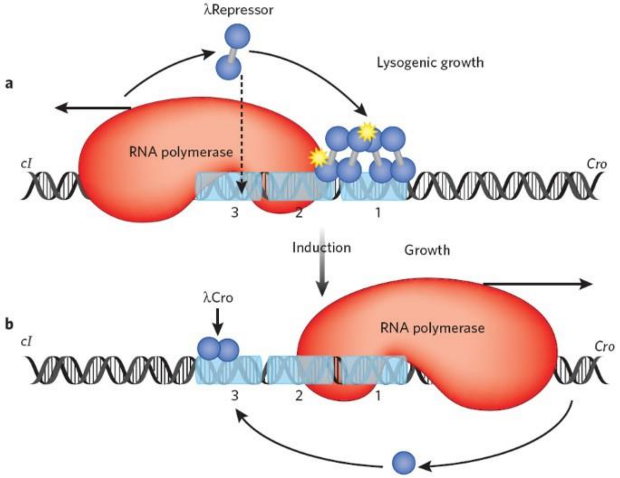
2/9/2026

Protein-DNA Interactions

- Jacob and Monod 1961: Repressors encoded in Lac Operon regulate rate of protein synthesis
- Open question: How do repressors regulate synthesis? Sequence-specific **DNA binding,** sequence-specific **mRNA inhibition**, or **tRNA interference**?
- **Specific protein-DNA interaction** between lamba repressor and lambda DNA (Ptashne 1967)



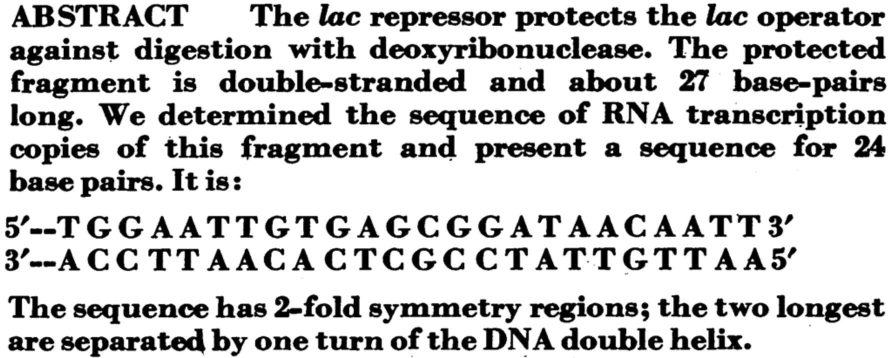
Ptashne (1967)



Ptashne (2011)

Protein-DNA Interactions

- Are interactions sequence-specific?
- Identification of **Lac repressor binding sequence** (Gilbert and Maxam 1973)
- Identification of the Lambda repressor binding sequence (Maniatas and Ptashne 1974)
- Identification of a **common sequence involved in prokaryotic transcription** (Pribnow 1975)

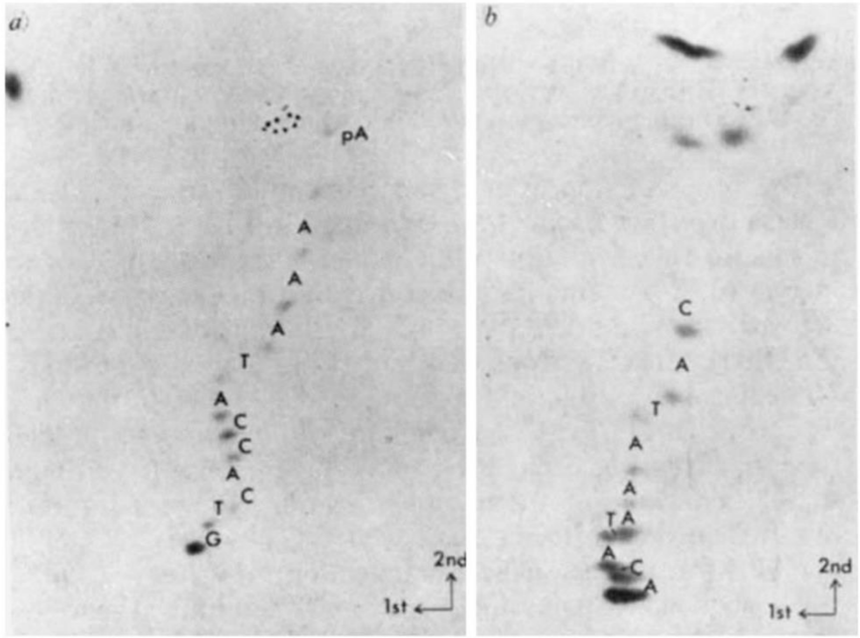


ABSTRACT The *lac* repressor protects the *lac* operator against digestion with deoxyribonuclease. The protected fragment is double-stranded and about 27 base-pairs long. We determined the sequence of RNA transcription copies of this fragment and present a sequence for 24 base pairs. It is:

5′--TGGAATTGTGAGCGGATAACAATT 3′
3′--ACCTTAACACTCGCCTATTGTTAA5′

The sequence has 2-fold symmetry regions; the two longest are separated by one turn of the DNA double helix.

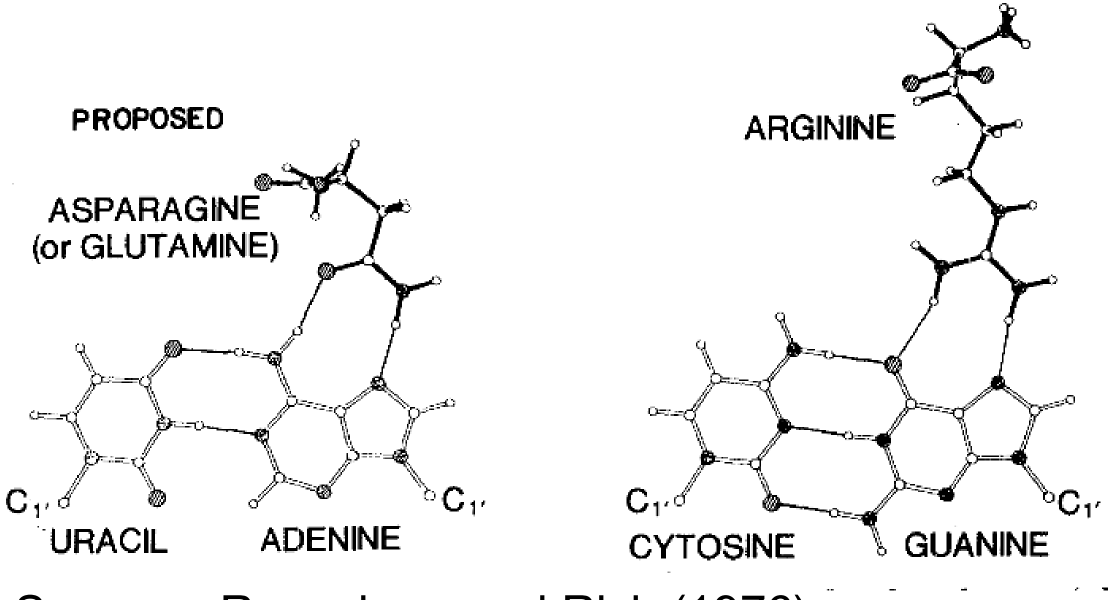
Gilbert and Maxam (1973)



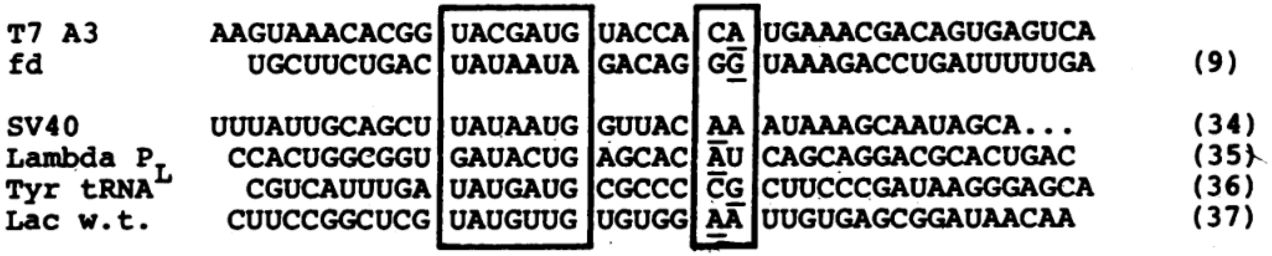
Maniatas et al. (1974)

Protein-DNA Interactions

- Is there a “recognition code” (specific sequences for specific proteins)?
- Perhaps Arginine and Asparagine/Glutamine can use dual H-bonding to distinguish bases (Seeman, Rosenberg, and Rich 1976)
- The Pribnow box (and TATA box) is **degenerate!**



Seeman, Rosenberg and Rich (1976)

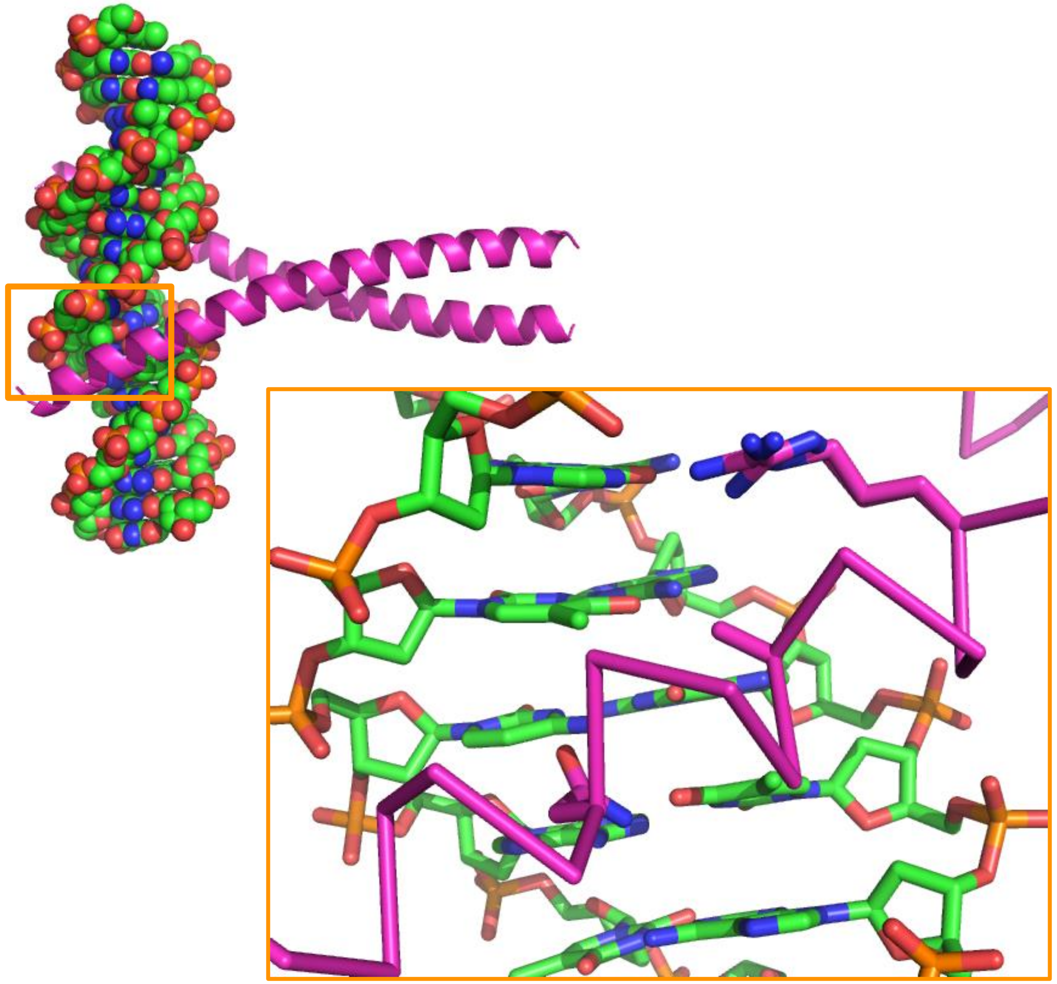


T7 A3	AAGUAAACACGG	UACGAUG	UACCA	CA	UGAAACGACAGUGAGUCA	
fd	UGCUUCUGAC	UAUAAUA	GACAG	GG	UAAAGACCUGAUUUUUGA	(9)
SV40	UUUAUUGCAGCU	UAUAAUG	GUUAC	AA	AUAAAGCAAUAGCA...	(34)
Lambda P_L	CCACUGGCGGU	GAUACUG	AGCAC	AU	CAGCAGGACGCACUGAC	(35)
Tyr tRNA	CGUCAUUUGA	UAUGAUG	CGCCC	CG	CUUCCCGAUAAGGGAGCA	(36)
Lac w.t.	CUUCCGGCUCG	UAUGUUG	UGUGG	AA	UUGUGAGCGGAUAACAA	(37)

Pribnow (1975)

Protein-DNA Interactions

- It has become clear **there is no universal code**. The interactions are degenerate in both directions.
- However, for a fixed mode of interaction (a single structural family of DNA-binding proteins), there is hope that partial weight matrices may be associated with key amino acid positions in the protein.
- Electrostatics, hydrogen bonds, water-mediated contacts, and hydrophobic packing
- In addition, sequence-specific DNA deformations (indirect readout) is often important
- This will require the determination of the binding preferences for many members of a family of TFs.



Sequence Motifs

- Motif: subsequence with some specific function
- May be in DNA, RNA, protein
- Function may be context dependent
  - Ribosome binding site must be transcribed
  - RNA, protein motifs may depend on structure
- May be gapped or ungapped
- Use model to search for (predict) new sites
  - Models may be simple sequences (regular expressions) or probabilistic patterns
- Modeling approach depends on data available
  - Quantitative/qualitative

Types of Motifs

Motif: Consensus Sequence Pattern

- May include degenerate bases and allow for mismatches
- *Search space is over possible patterns*

Weight Matrix (PWM, Profile, PSSM)

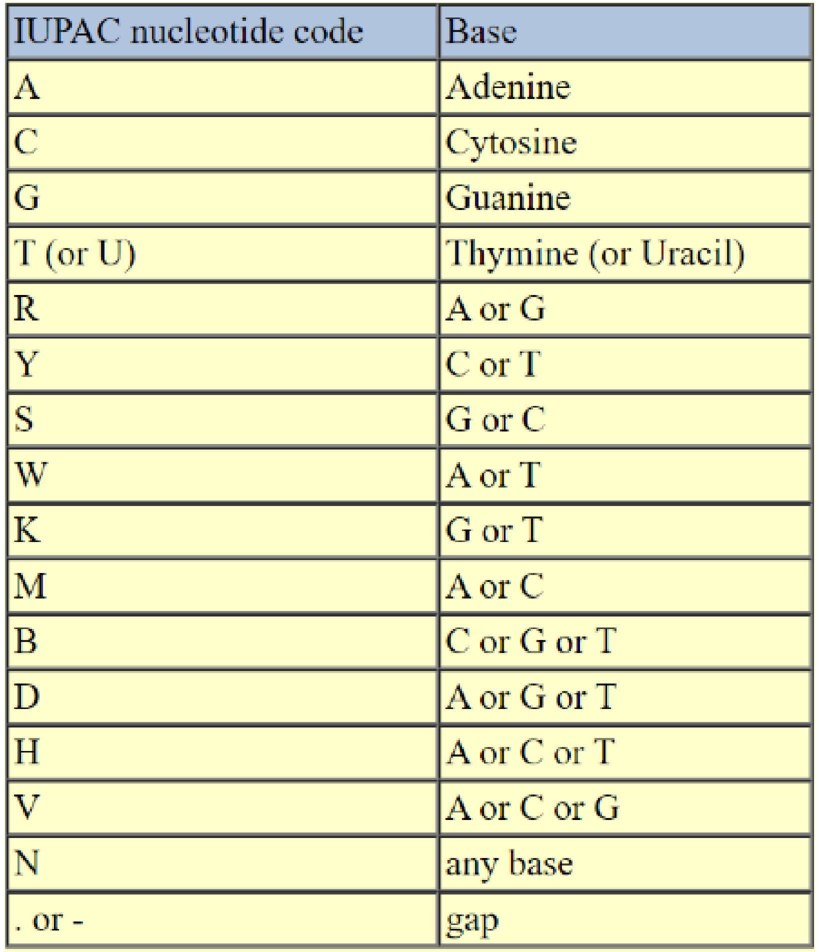
- Might go to higher order models
- *Search space is over possible alignments*

Pattern based algorithms

- Motif length l, mismatches m; N seqs, L long

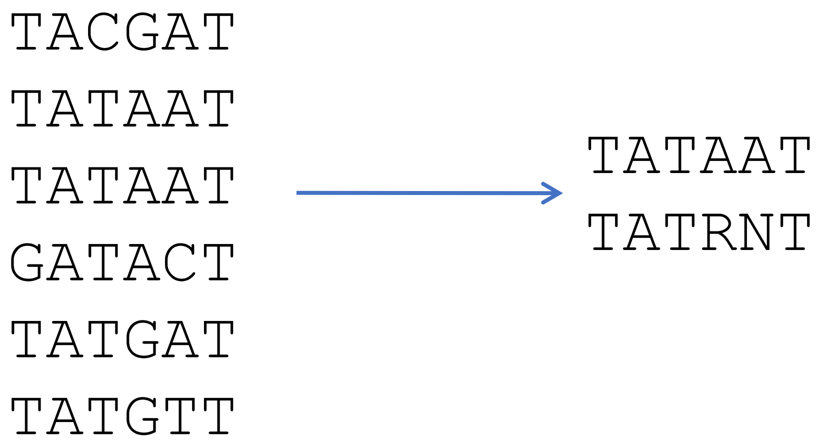
- 4^l patterns, search for most common (or most significant) allowing up to m mismatches
    - P-value from background distribution
    - Can allow for m mismatches
    - Can allow degenerate positions: 15^l patterns
    - Can just search using existing l-mers

- Can use suffix tree for efficient search of patterns allowing mismatches



IUPAC nucleotide code	Base
A	Adenine
C	Cytosine
G	Guanine
T (or U)	Thymine (or Uracil)
R	A or G
Y	C or T
S	G or C
W	A or T
K	G or T
M	A or C
B	C or G or T
D	A or G or T
H	A or C or T
V	A or C or G
N	any base
. or -	gap

Consensus Sequence Pattern



TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT

TATAAT
TATRNT

- Difficult to obtain an optimal consensus for identifying novel sites

- Relative frequency of bases at each positions lost

Weight Matrix Model

```
TACGAT
TATAAT
TATAAT
GATACT
TATGAT
TATGTT
```

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

- More information than a consensus sequence
- Many ways to determine the weights
- Assumes positional independence
- Requires significant data

Score a site

-24

….A **C** **T** **A** **T** **A** **A** T G T …

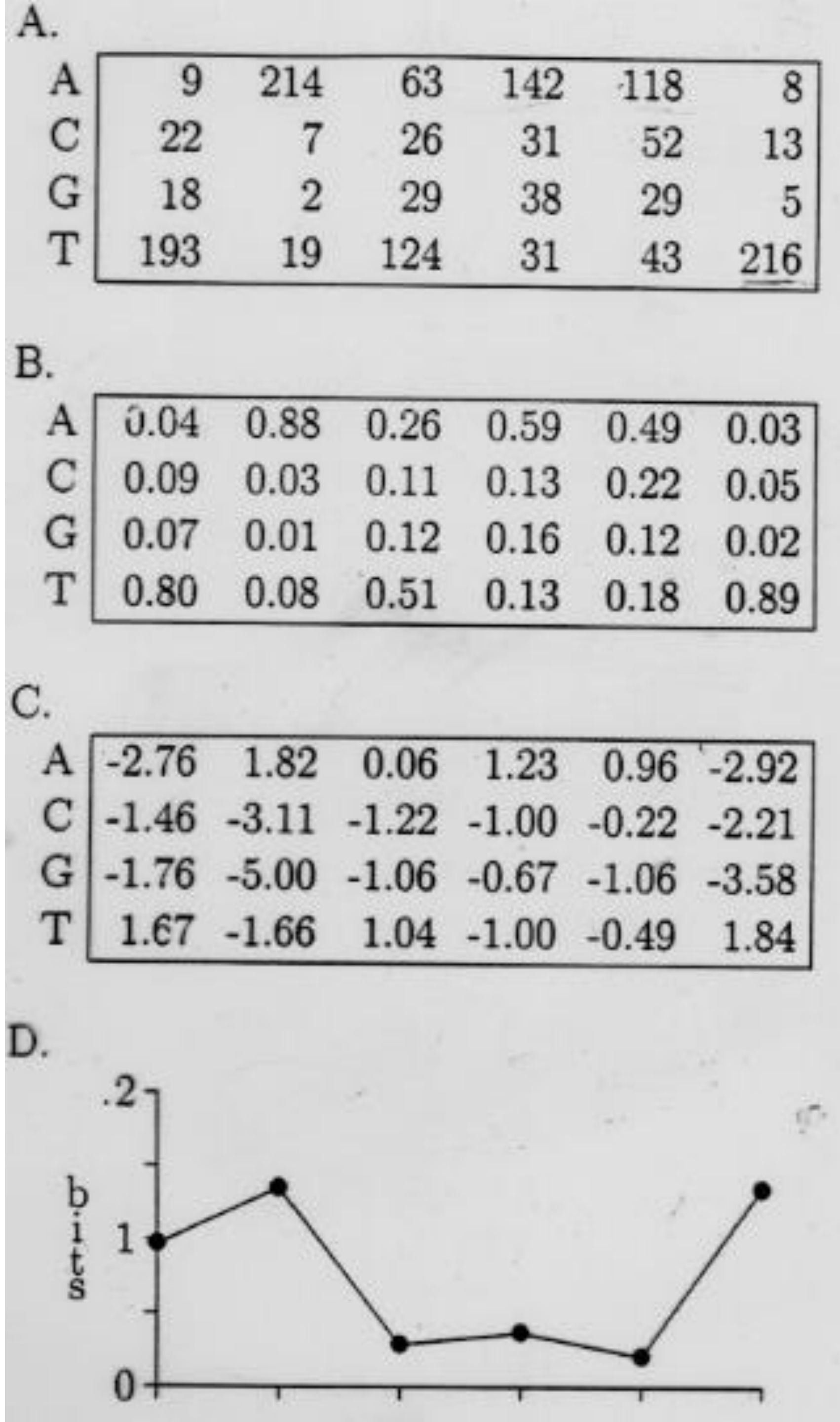
A:	-8	10	**-1**	2	**1**	**-8**
C:	**-10**	-9	-3	-2	-1	-12
G:	-7	-9	-1	-1	-4	-9
T:	**10**	**-6**	9	**0**	-1	11

Score a site

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.

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

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

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

N(b,j): Raw score

F(b,j): Weighted score

S(b,j) = log[F(b,j)/P(b)]
Probability-normalized log score

I(j) = ∑F(b,j)S(b,j)

j=jth position in sequence (column index)
b=base (A, C, G, or T) (row index)
(b,j) base b evaluated at position j

G. Stormo

Information Content

Matrix of Frequencies

A:	0.1	0.7	0.2	0.3	0.4	0.1
C:	0.1	0.1	0.1	0.3	0.2	0.1
G:	0.1	0.1	0.2	0.1	0.2	0.1
T:	0.7	0.1	0.5	0.3	0.2	0.7

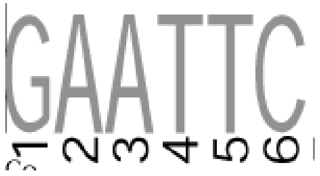
$$I_{seq} = \sum_{j} \sum_{b} f(b,j) \log_2 \frac{f(b,j)}{p(b)}$$

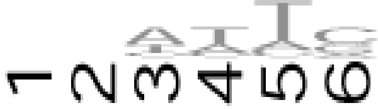
Sum is over columns j (the positions), and rows b (the bases)
to distinguish divergence of the empirical distribution (*f(b,j)*) from the background base distribution (*p(b)*)

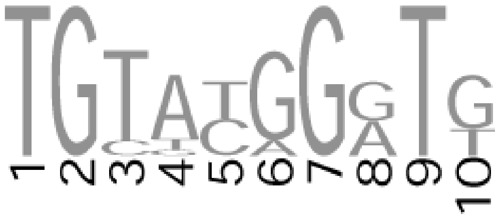
aka Relative Entropy, Kullback-Leibler Distance

Information Content

EcoR1	Random	Rap1
GAATTC	GCCTAC	TGTATGGGTG
GAATTC	ACATTC	TGTTCGGATT
GAATTC	TCATTC	TGCATGGGTG
GAATTC	CGACTC	TGTACAGGTG
GAATTC	GAATTC	TGTATGGATG
GAATTC	ATATCG	TGTTCGGGTT
GAATTC	GAAATG	TGTATGGGTG







Pseudocounts

Entries of zero in the count matrix cause big problems

- The log(0) is undefined (infinitely negative)
- Not enough observations to observe all possibilities

Pseudocounts

A	0	17	5	3
T	10	0	5	2
G	4	3	5	5
C	6	0	5	10

Original count matrix

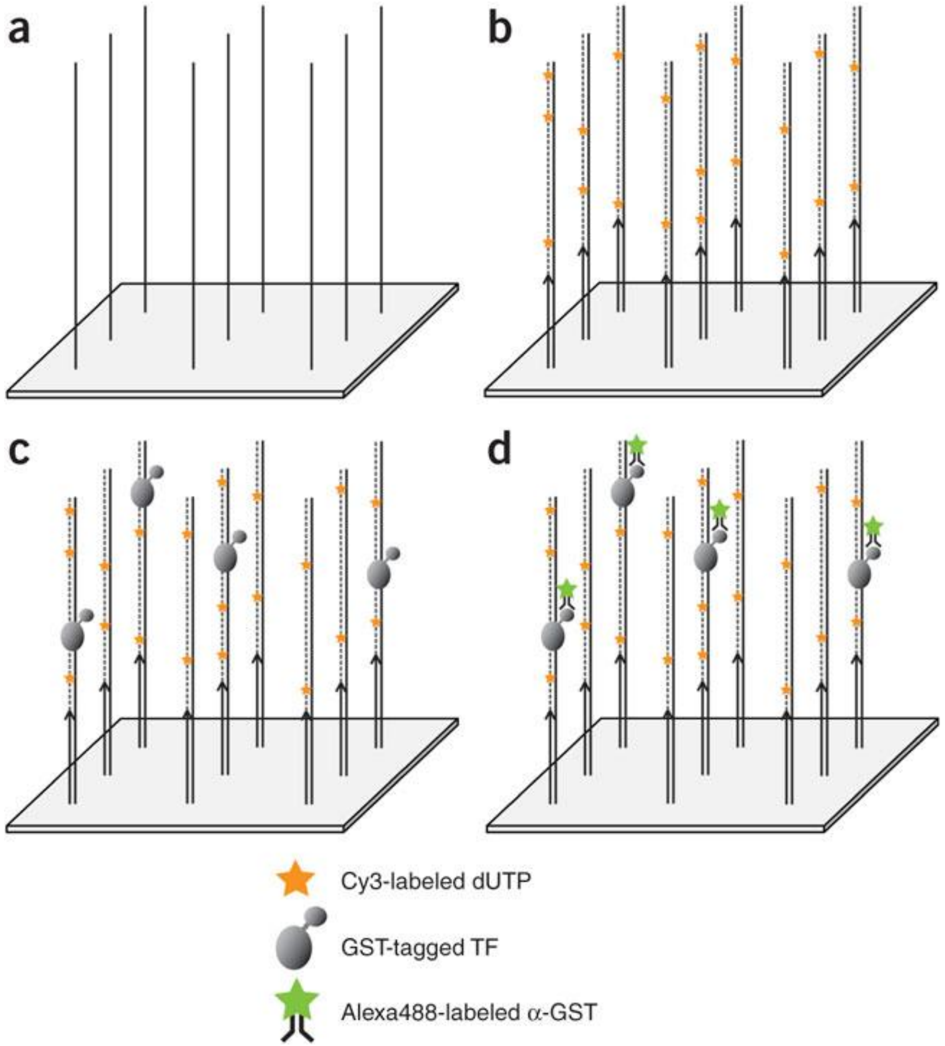
A	0.25	17.25	5.25	3.25
T	10.25	0.25	5.25	2.25
G	4.25	3.25	5.25	5.25
C	6.25	0.25	5.25	10.25

eg.1 Add 1 pseudocount per column

A	0.2	17.2	5.2	3.2
T	10.2	0.2	5.2	2.2
G	4.3	3.3	5.3	5.3
C	6.3	0.3	5.3	10.3

eg.2 Add 1 pseudocount per column according to background nucleotide frequencies
Assume %A=%T=20%

Methods for defining PWM: Protein Binding Microarrays



Custom arrays of 60-mer DNA sequences (~44,000 probes)

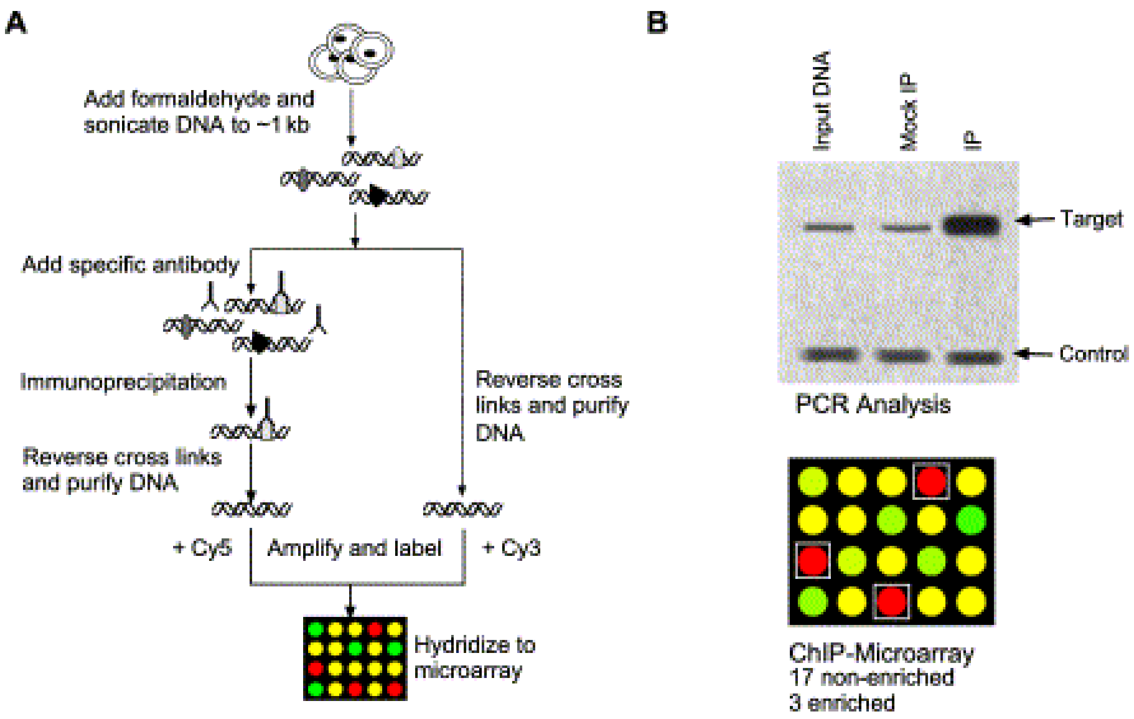
Contain all possible 10bp sequences

Each probe contains 27 10-mers

8-mers guaranteed to occur 16 times

Berger and Bulyk, Nature Protocols (2009)

Methods for defining PWM: ChIP-chip/ChIP-seq



Cross-link protein to DNA

Affinity purify protein-DNA complexes:

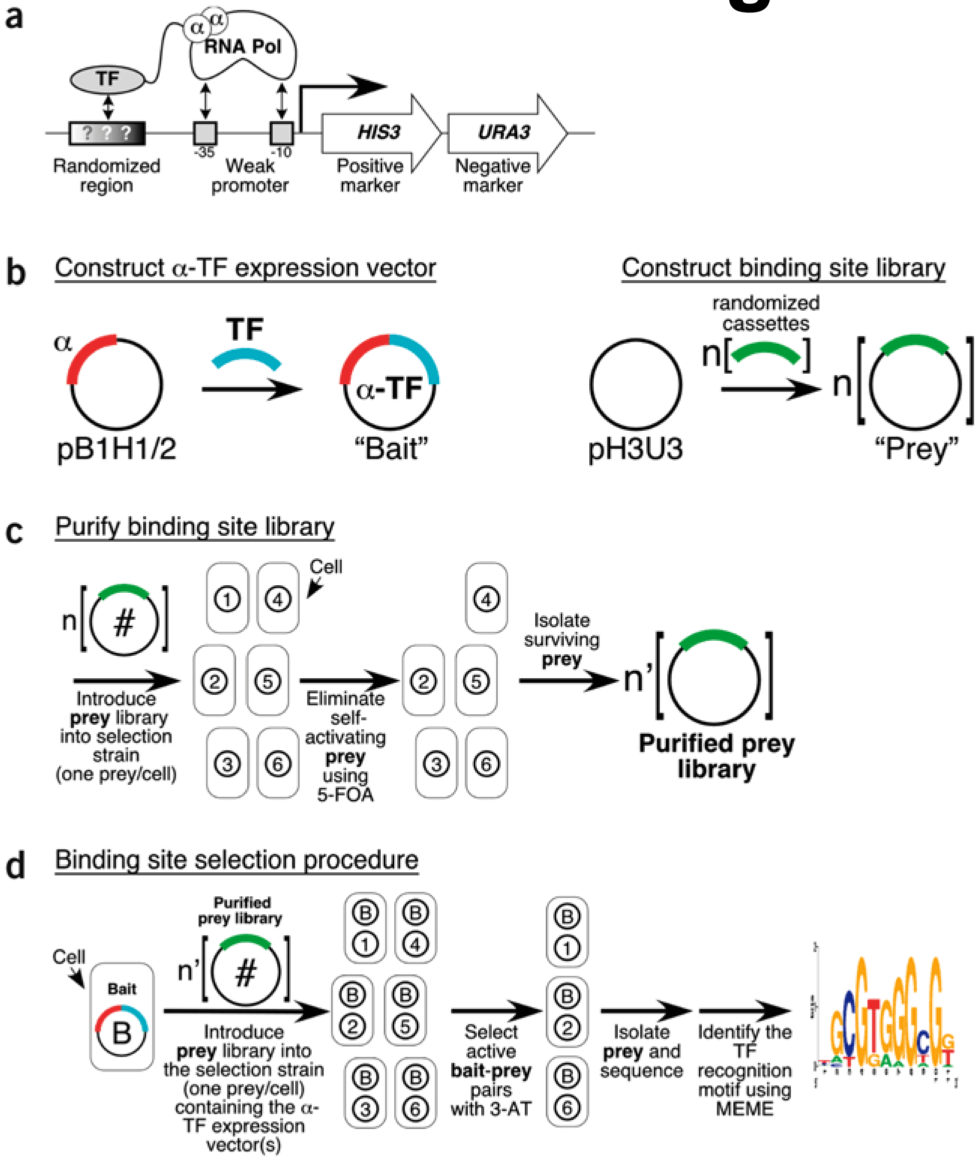
Ab to TF
Ab to tag on TF
affinity tag on TF

Reverse cross-links

Identify sequence by hybridization to microarray or by high-throughput sequencing

Buck and Lieb, Genomics (2004), Robertson, et. al., Nature Methods (2007)

Methods for defining PWM: Bacterial One-hybrid


a
TF
α
α
RNA Pol
? ? ?
-35
-10
HIS3
URA3
Randomized region
Weak promoter
Positive marker
Negative marker
b
Construct α-TF expression vector
α
TF
α-TF
pB1H1/2
"Bait"
Construct binding site library
randomized cassettes
n[ ]
n
pH3U3
"Prey"
c
Purify binding site library
n
#
Cell
1 4
2 5
3 6
Introduce prey library into selection strain (one prey/cell)
Eliminate self-activating prey using 5-FOA
4
2 5
3 6
Isolate surviving prey
n'
Purified prey library
d
Binding site selection procedure
Cell
Bait
B
Purified prey library
n'
#
Introduce prey library into the selection strain (one prey/cell) containing the α-TF expression vector(s)
B 1 B 4
B 2 B 5
B 3 B 6
Select active bait-prey pairs with 3-AT
B 1
B 2
B 6
Isolate prey and sequence
Identify the TF recognition motif using MEME

Genetic selection: survival is dependent on DNA-binding

TF of interest is fused to α-subunit of RNA polymerase

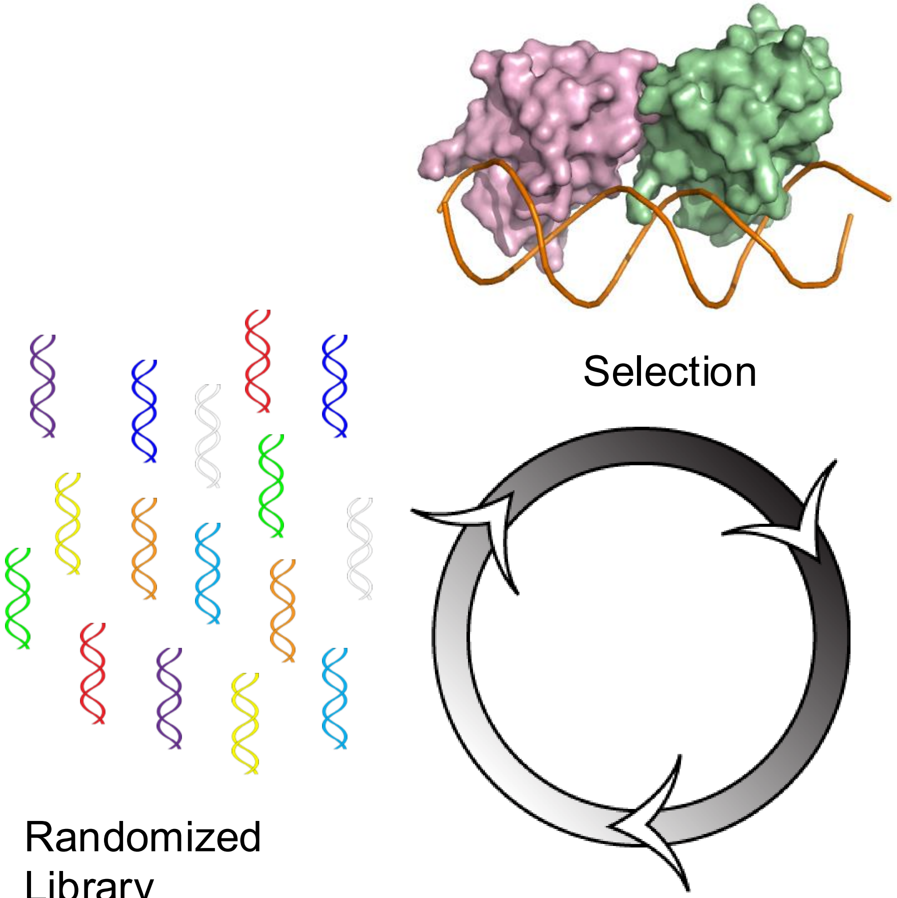
Randomized library of binding sites created and screened for autoactivation

Co-transform with TF and select

Library complexity is limited by transformation efficiency (~10^9)

Meng and Wolfe, Nature Protocols (2006)

Methods for defining PWM: High-throughput SELEX



Selection

Randomized Library

Amplification

Next-gen Sequencing

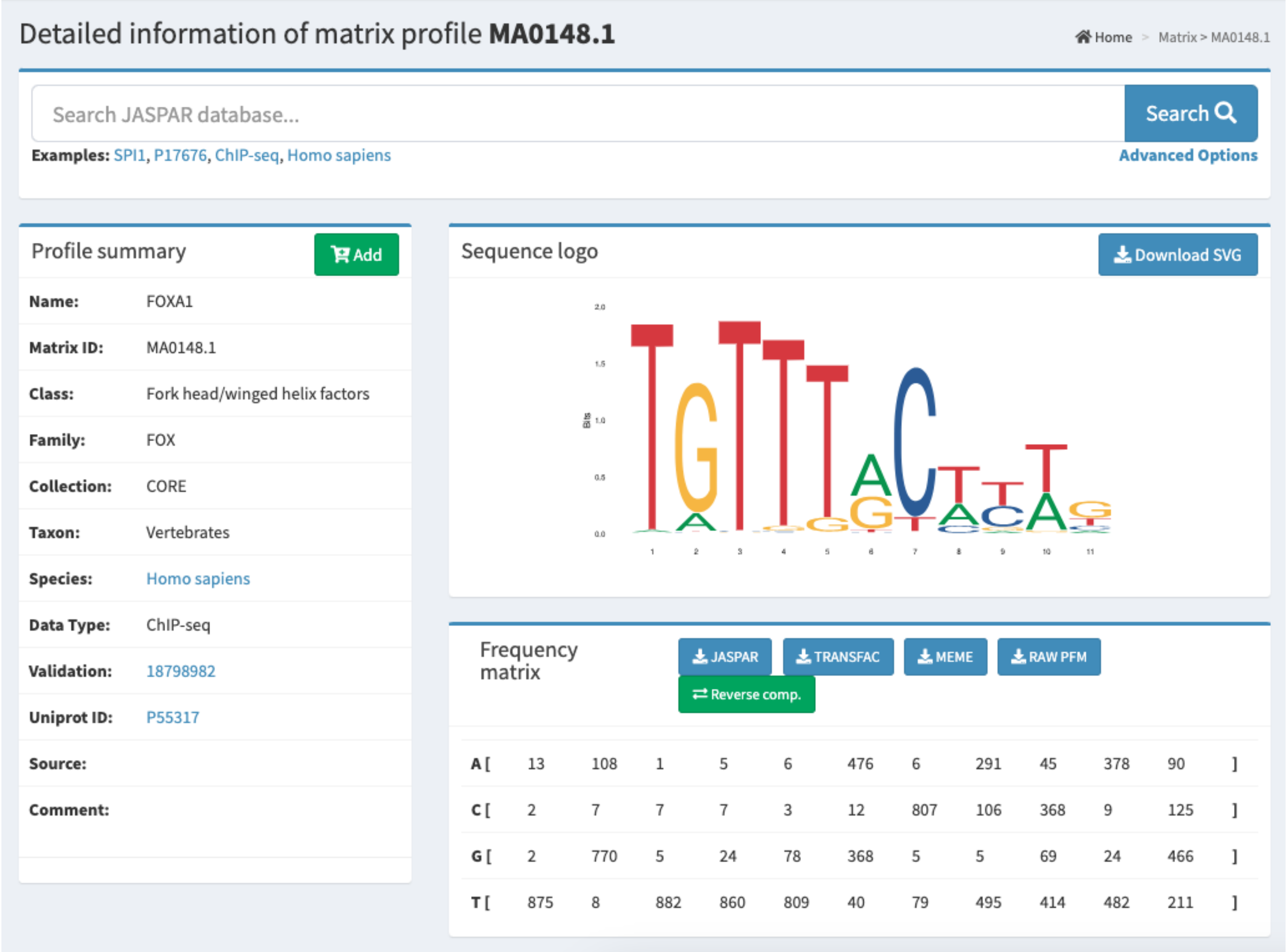
Purified protein

High complexity library

Zhao, Granas, and Stormo, PLOS Comput Biol (2009)

Public repositories of DNA protein binding data

- JASPAR: Since 2004, regularly updated open source repository
- ChIP, PBM, SELEX, etc.
- > 50 different species spanning most clades
- Extract PWMs for downstream analysis



Detailed information of matrix profile **MA0148.1**

Examples: SPI1, P17676, ChIP-seq, Homo sapiens

Profile summary

Name:	FOXA1
Matrix ID:	MA0148.1
Class:	Fork head/winged helix factors
Family:	FOX
Collection:	CORE
Taxon:	Vertebrates
Species:	Homo sapiens
Data Type:	ChIP-seq
Validation:	18798982
Uniprot ID:	P55317
Source:	
Comment:	

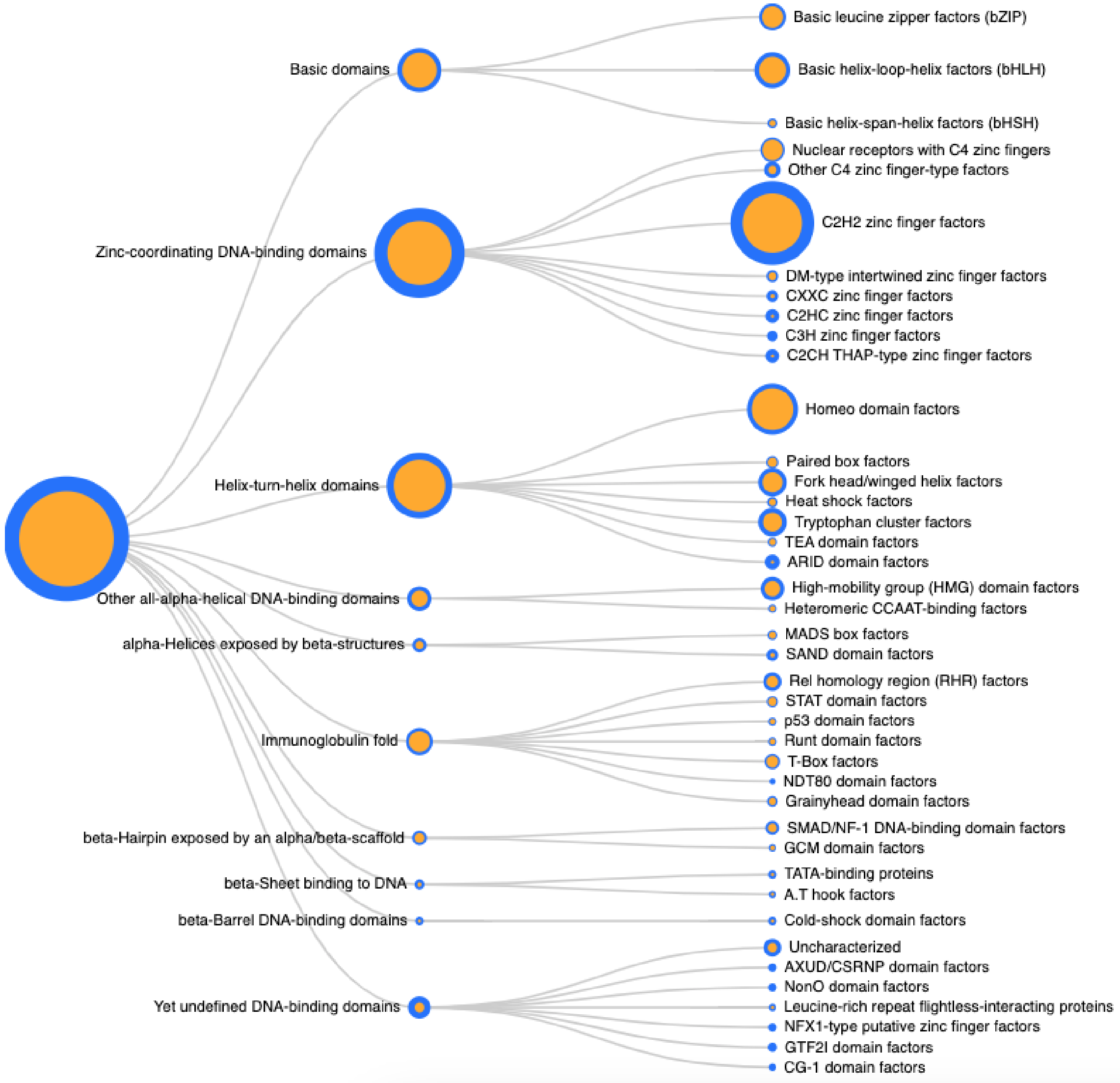
Sequence logo

Frequency matrix

A [	13	108	1	5	6	476	6	291	45	378	90	]
C [	2	7	7	7	3	12	807	106	368	9	125	]
G [	2	770	5	24	78	368	5	5	69	24	466	]
T [	875	8	882	860	809	40	79	495	414	482	211	]

Public repositories of DNA protein binding data

- HOmo sapiens COmprehensive MOdel Collection (HOCOMOCO)
- Human-specific (949 TFs)
- Coverage of nearly all human DNA binding domain classes


Basic domains
Zinc-coordinating DNA-binding domains
Helix-turn-helix domains
Other all-alpha-helical DNA-binding domains
alpha-Helices exposed by beta-structures
Immunoglobulin fold
beta-Hairpin exposed by an alpha/beta-scaffold
beta-Sheet binding to DNA
beta-Barrel DNA-binding domains
Yet undefined DNA-binding domains
Basic leucine zipper factors (bZIP)
Basic helix-loop-helix factors (bHLH)
Basic helix-span-helix factors (bHSH)
Nuclear receptors with C4 zinc fingers
Other C4 zinc finger-type factors
C2H2 zinc finger factors
DM-type intertwined zinc finger factors
CXXC zinc finger factors
C2HC zinc finger factors
C3H zinc finger factors
C2CH THAP-type zinc finger factors
Homeo domain factors
Paired box factors
Fork head/winged helix factors
Heat shock factors
Tryptophan cluster factors
TEA domain factors
ARID domain factors
High-mobility group (HMG) domain factors
Heteromeric CCAAT-binding factors
MADS box factors
SAND domain factors
Rel homology region (RHR) factors
STAT domain factors
p53 domain factors
Runt domain factors
T-Box factors
NDT80 domain factors
Grainyhead domain factors
SMAD/NF-1 DNA-binding domain factors
GCM domain factors
TATA-binding proteins
A.T hook factors
Cold-shock domain factors
Uncharacterized
AXUD/CSRNP domain factors
NonO domain factors
Leucine-rich repeat flightless-interacting proteins
NFX1-type putative zinc finger factors
GTF2I domain factors
CG-1 domain factors

Motif Finding Problem

- A fundamental problem in molecular biology
    - Specific protein and DNA binding
    - Transcription factor binding sites recognition

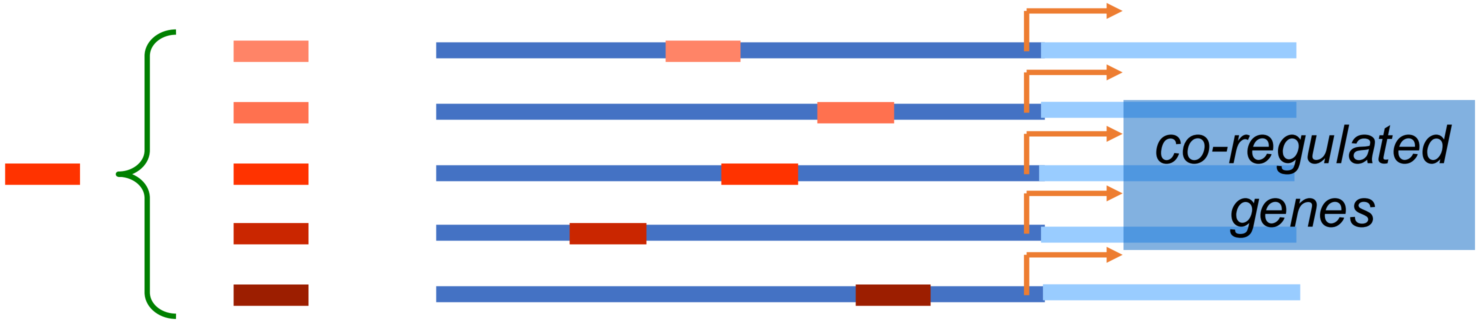
- Statistical definition:
    - Given some sequences, find over-represented substrings (motif discovery)

- Biological definition:
    - Given some co-regulated promoters, find transcription factor binding model
    - How do we use biology to improve motif finding algorithms?

- Many algorithms/programs developed
    - consensus, gibbs sampling, EM, projection, phylogenetic footprinting, etc.

Motif Finding Algorithms Class I

Single species, multiple genes (planted motif problem)



- random background sequences
- a proper description of a consensus motif → better models
- randomly plant copies of the motif into sequences
- define an objective function, and use a search algorithm to find the copies that give a good score

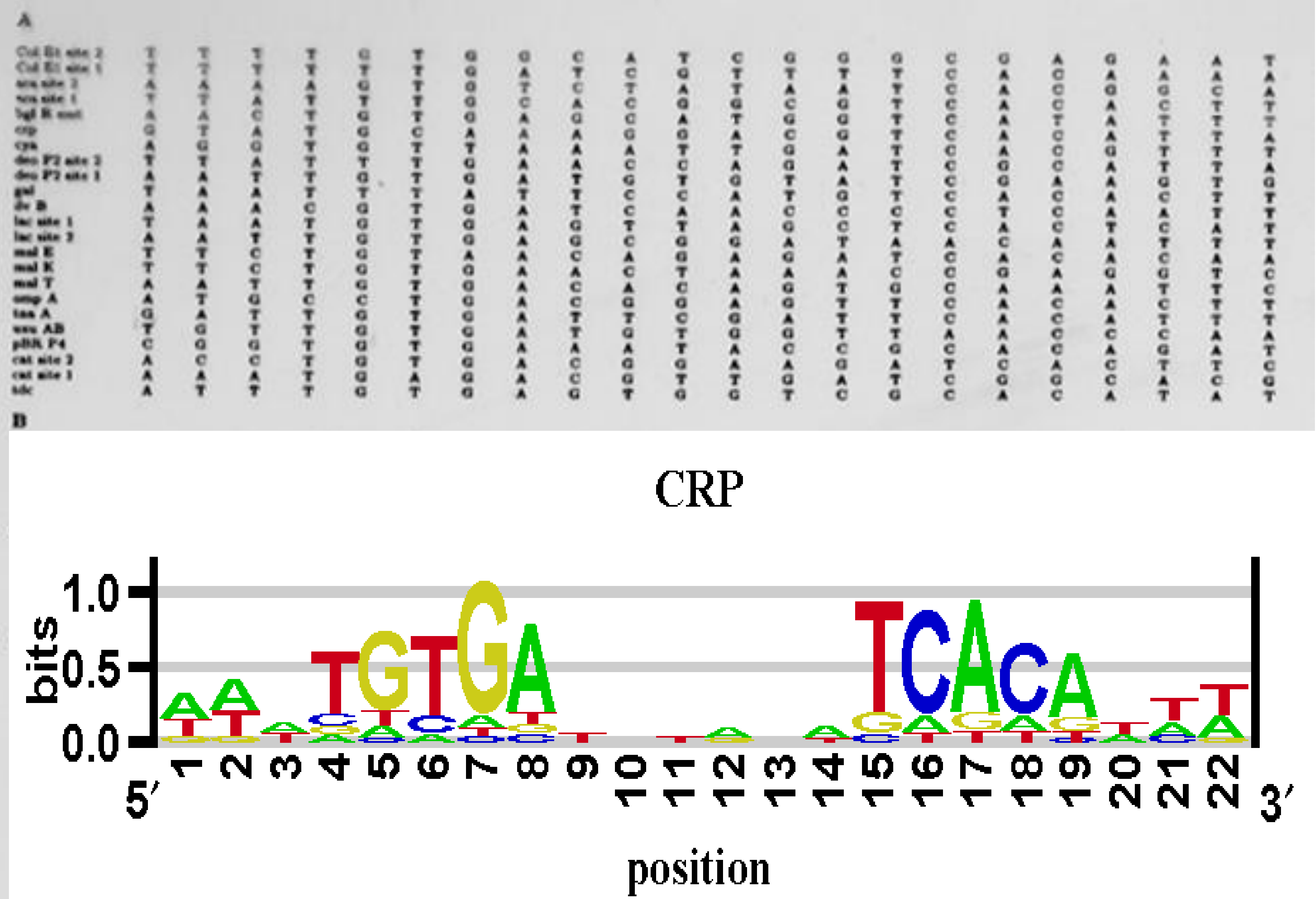
Real-world case

The Data Set: Sequences containing sites for cAMP receptor protein (CRP)

locus	sequence
colel	taatgtttgtgctggt**TTTTGTGGCATCGGGCGAGAAT**agcgcgtggtgaaagactgt**TTTTTTGATCGTTTTCACAAAA**atggaagtccacagtcttgacag
ecoarabop	gacaaaaacgcgtaac**AAAAGTGTCTATAATCACGGCA**gaaaagtccacattga**TTATTTGCACGGCGTCACACTT**tgctatgccatagcattttatccataag
ecobglrl	acaaatcccaataacttaattattgggatttgttatatataactttataaattcctaaaattacacaaagttaat**AACTGTGAGCATGGTCATATTT**ttatcaat
ecocrp	cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgtactgcat**GTATGCAAAGGACGTCACATTA**ccgtgcagtacagttgatagc
ecocya	acggtgctacacttgtatgtagcgcatctttctttacggtcaatcagca**AGGTGTTAAATTGATCACGTTT**tagaccatttttcgtcgtgaaactaaaaaacc
ecodeop	agtgaa**TTATTTGAACCAGATCGCATTA**cagtgatgcaaacttgtaagtagatttcctt**AATTGTGATGTGTATCGAAGTG**tgttgcggagtagatgttagaata
ecogale	gcgcataaaaaacggctaaattcttgtgtaaacgattccac**TAATTTATTCCATGTCACACTT**ttcgcatctttgttatgctatggttatttcataccataagcc
ecoilvbpr	gctccggcggggtttttttgttatctgcaattcagtaca**AAACGTGATCAICCCCTCAATT**ttccctttgctgaaaaattttccattgtctcccctgtaaagctgt
ecolac	aacgcaat**TAATGTGAGTTAGCTCACTCAT**taggcaccccaggctttacactttatgcttccggctcgtatgttgtgtgg**AATTGTGAGCGGATAACAATTT**cac
ecomale	acattaccgccaa**TTCTGTAACAGAGATCACACAA**agcgacggtggggcgtaggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgttta
ecomalk	ggaggaggcgggaggatgagaacacggc**TTCTGTGAACTAAACCGAGGTC**atgtaaggaatttcgtgatgttgcttgcaaaaatcgtggcgattttatgtgcgca
ecomalt	gatcagcgtcgttttaggtgagttgttaataaagatttgg**AATTGTGACACAGTGCAAATTC**agacacataaaaaaacgtcatcgcttgcattagaaaggtttct
ecoompa	gctgacaaaaaagattaaacataccttatacaagacttttttttcat**ATGCCTGACGGAGTTCACACTT**gtaagttttcaactacgttgtagactttacatcgcc
ecotnaa	tttttaaacattaaaattcttacgtaatttataatctttaaaaaaagcatttaatattgctccccgaac**GATTGTGATTCGATTCACATTT**aaacaatttcaga
ecouxul	cccatgagagtgaaat**TGTTGTGATGTGGTTAACCCAA**ttagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatcteatccgatgcaagc
pbr-p4	ctggcttaactatgcggcatcagagcagattgtactgagagtgcaccatatg**CGGTGTGAAATACCGCACAGAT**gcgtaaggagaaaataccgcatcaggcgctc
trn9cat	**CTGTGACGGAAGATCACTTC**gcagaataaataaatcctggtgtccctgttgataccgggaagccctgggccaacttttggcga**AAATGAGACGTTGATCGGCACG**
tdc	gatttttatactttaacttgttgatatttaaaggtatttaattgtaataacgatactctggaaagtattgaaagtta**ATTTGTGAGTGGTCGCACATAT**cctgtt

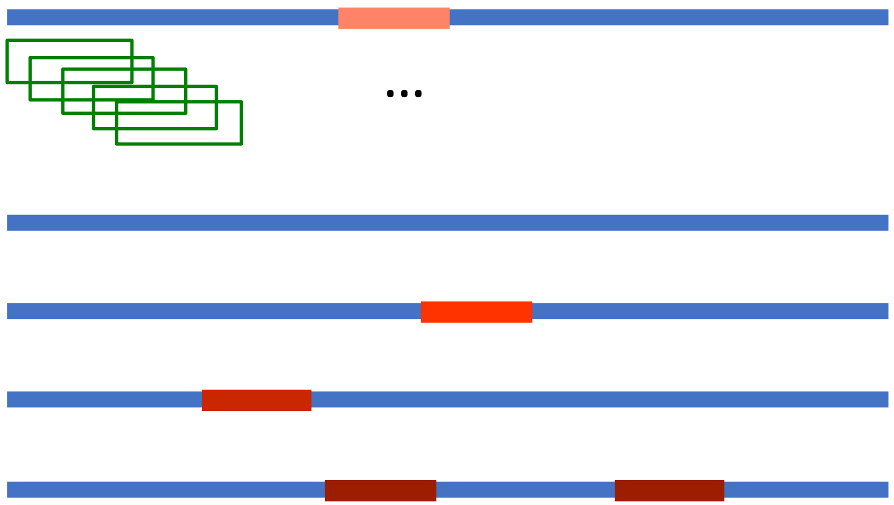
For this case, there are 18 sequences of length 105 bp and we are looking for a motif of width 20 bp. There are 86 different 20 bp subsequences per example and ~7x10^{34} alignments to check.

Stormo and Hartzell, Proc. Natl. Acad. Sci. (1989)



Col E1 site 2	T	T	T	T	G	T	G	G	C	A	T	C	G	G	G	C	G	A	G	A	A	T
Col E1 site 1	T	T	T	T	T	T	G	A	T	C	G	T	T	T	T	C	A	C	A	A	A	A
ara site 2	A	A	A	A	G	T	G	T	C	T	A	T	A	A	T	C	A	C	G	G	C	A
ara site 1	T	T	A	T	T	T	G	C	A	C	G	G	C	G	T	C	A	C	A	C	T	T
bgl R wt	A	A	C	T	G	T	G	A	G	C	A	T	G	G	T	C	A	T	A	T	T	T
crp	G	T	A	T	G	C	A	A	A	G	G	A	C	G	T	C	A	C	A	T	T	A
cya	A	G	G	T	G	T	T	A	A	A	T	T	G	A	T	C	A	C	G	T	T	T
deo P2 site 2	T	T	A	T	T	T	G	A	A	C	C	A	G	A	T	C	G	C	A	T	T	A
deo P2 site 1	A	A	T	T	G	T	G	A	T	G	T	G	T	A	T	C	G	A	A	G	T	G
gal	T	A	A	T	T	T	A	T	T	C	C	A	T	G	T	C	A	C	A	C	T	T
ilv B	A	A	A	C	G	T	G	A	T	C	A	A	C	C	C	C	T	C	A	A	T	T
lac site 1	T	A	A	T	G	T	G	A	G	T	T	A	G	C	T	C	A	C	T	C	A	T
lac site 2	A	A	T	T	G	T	G	A	G	C	G	G	A	T	A	A	C	A	A	T	T	T
mal E	T	T	C	T	G	T	A	A	C	A	G	A	G	A	T	C	A	C	A	C	A	A
mal K	T	T	C	T	G	T	G	A	A	C	T	A	A	A	C	C	G	A	G	G	T	C
mal T	A	A	T	T	G	T	G	A	C	A	C	A	G	T	G	C	A	A	A	T	T	C
omp A	A	T	G	C	C	T	G	A	C	G	G	A	G	T	T	C	A	C	A	C	T	T
tna A	G	A	T	T	G	T	G	A	T	T	C	G	A	T	T	C	A	C	A	T	T	T
uxu AB	T	G	T	T	G	T	G	A	T	G	T	G	G	T	T	A	A	C	C	C	A	A
pBR P4	C	G	G	T	G	T	G	A	A	A	T	A	C	C	G	C	A	C	A	G	A	T
cat site 2	A	C	C	T	G	T	G	A	C	G	G	A	A	G	A	T	C	A	C	T	T	C
cat site 1	A	A	A	T	G	A	G	A	C	G	T	T	G	A	T	C	G	G	C	A	C	G
tdc	A	T	T	T	G	T	G	A	G	T	G	G	T	C	G	C	A	C	A	T	A	T

An (intractable) solution



(Exhaustive algorithm)
Construct every possible combination of alignments and keep the one with the highest information content.

Given a motif of width w, and k sequences of length l, there are L = (l-w+1) possible locations in each sequence, and L^k alignments to check.

Greedy Algorithm (Consensus)

- Simple version: assume every sequence contains at least one true binding site
- Using each l-mer find best match to generate 2-seq alignments
- Using top K PWMs to search remaining sequences to include a new sequence
- Repeat until all seqs contribute
    - Or objective function is maximized (IC, p-value)

Expectation Maximization (MEME)

- Initial “seed” PWM (at random or empirically generated from the average over all potential sites)
- Use the current PWM to determine probability of all positions being sites
- Re-estimate PWM based on the full set of those probabilities
- Continue until convergence – always convergences to a **local maximum**
- EM is **deterministic**, meaning it is sensitive to initial seed and may not converge to the **global maximum**
- For this reason, EM should be run **multiple times with different seeds**

Gibbs Sampling

- Similar to EM, but some important differences:
    - Initial “seed” PWM
    - Use the current PWM to determine probability of all positions
    - At each iteration, **pick one site on each sequence**, chosen by its probability, to update the PWM (rather than updating using the full set of probabilities)
- Not guaranteed to converge, but tends to increase objective (IC) and plateau
- Can **escape local maxima**, and therefore is **not sensitive to seed**
    - Other MCMC algorithms
        - Metropolis
        - Simulated annealing

Gibbs’ Sampling Approach to Motif Discovery

Basic Idea:

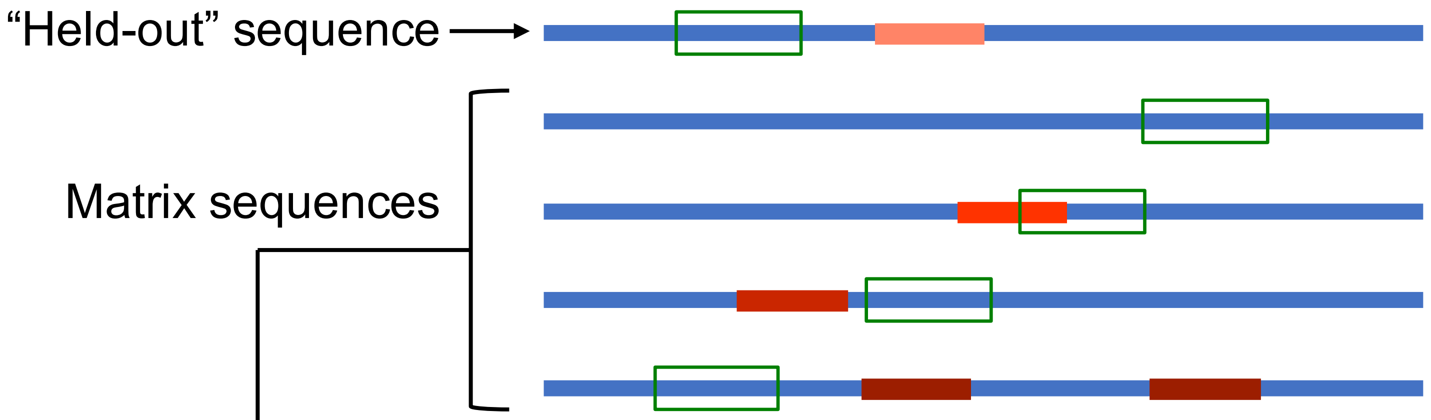
- Given “sites”, estimate pattern matrix
- Given “matrix”, pick likely sites according to their probability
- Iterate between those steps until “convergence”

Important details:

- Use “pseudocounts” to avoid prob. = 0
- Sample sites from estimated prob. distrib.

Gibbs Sampling

Initialization: Random assignment of motif locations a_1-a_k



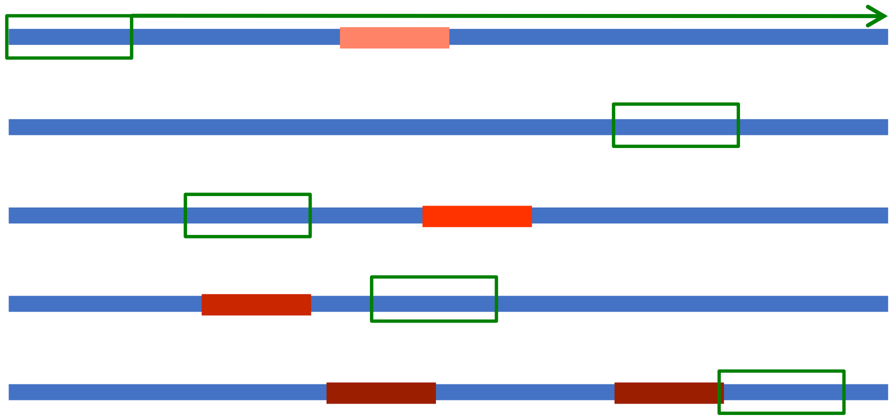
Construct initial matrix S from alignment of matrix sequences

A:	0.2	0.4	0.3	0.2	0.3	0.3
C:	0.2	0.2	0.2	0.3	0.3	0.1
G:	0.1	0.3	0.2	0.2	0.2	0.3
T:	0.5	0.1	0.3	0.3	0.2	0.3

Gibbs Sampling

Update:

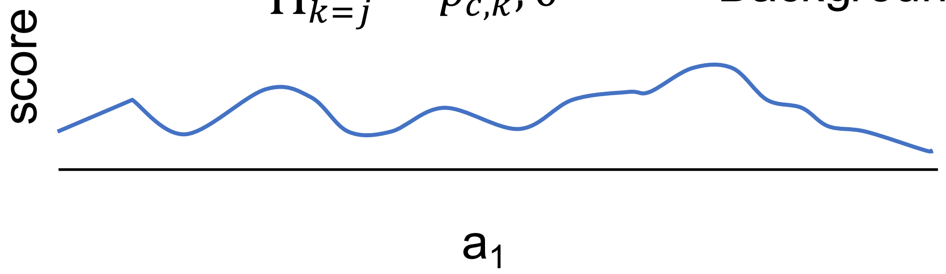
Score all possible motif locations of held-out sequence



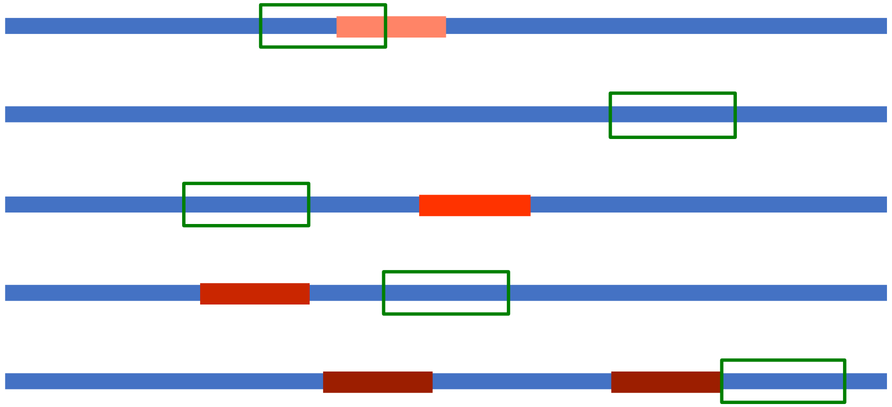
$$A_{i,j} = \frac{\prod_{k=j}^{j+W-1} p_{c,k}, k-j+1}{\prod_{k=j}^{j+W-1} p_{c,k}, 0}$$

← Updated PWM prob.

← Background PWM prob.

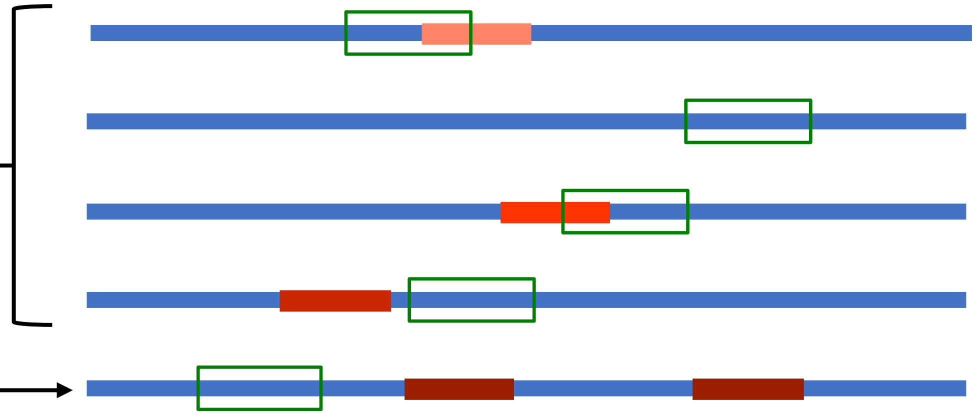


Select a new motif placement **randomly** based on probability distribution $A_{i,j}$



Gibbs Sampling

Iterate: Hold out new sequence



Matrix sequences

“Held-out” sequence

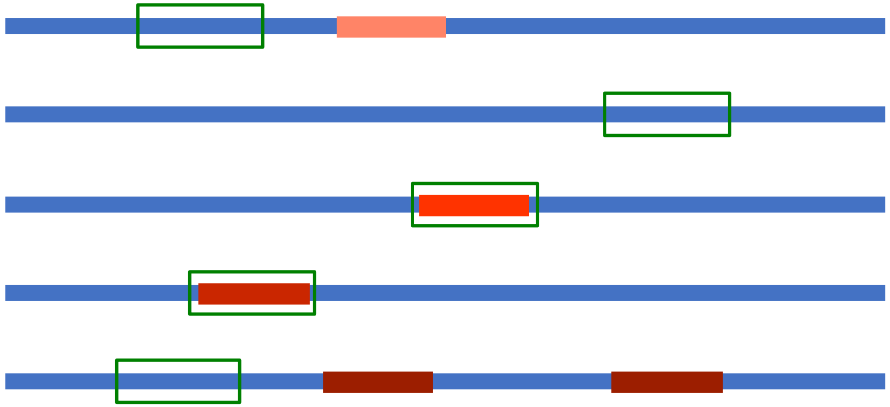
Construct new matrix S from alignment of matrix sequences

A:	0.2	0.4	**0.2**	0.2	**0.4**	**0.4**
C:	**0.1**	**0.1**	0.2	**0.4**	0.3	0.1
G:	0.1	**0.4**	0.2	0.2	0.2	**0.2**
T:	**0.6**	0.1	**0.4**	**0.2**	**0.1**	0.3

…and so on and so forth

Gibbs Sampling

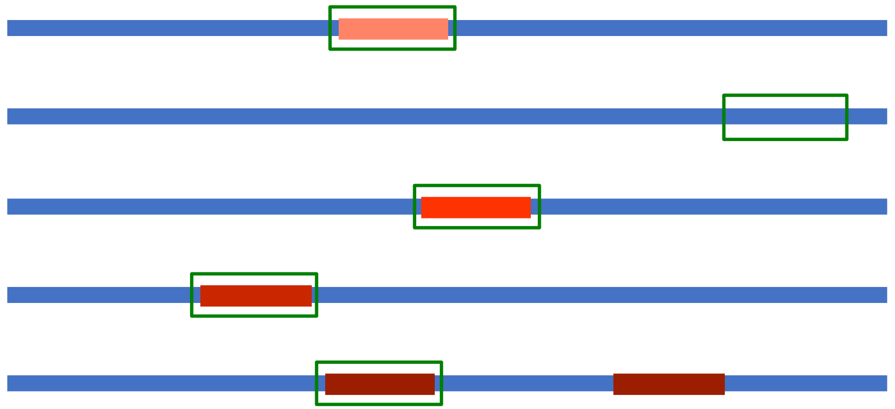
How does it end? Eventually you nucleate a few correct placements



The matrix has weak but sufficient scoring power

A:	**0.1**	0.4	0.2	0.2	0.4	**0.7**
C:	0.1	0.1	0.2	**0.6**	0.3	0.1
G:	0.1	0.4	0.2	**0.1**	0.2	**0.1**
T:	**0.7**	0.1	0.4	**0.1**	0.1	**0.1**

Gibbs Sampling



An approximately correct matrix rapidly converges, with the subsequent alignments possessing more information content and making better motif window placements

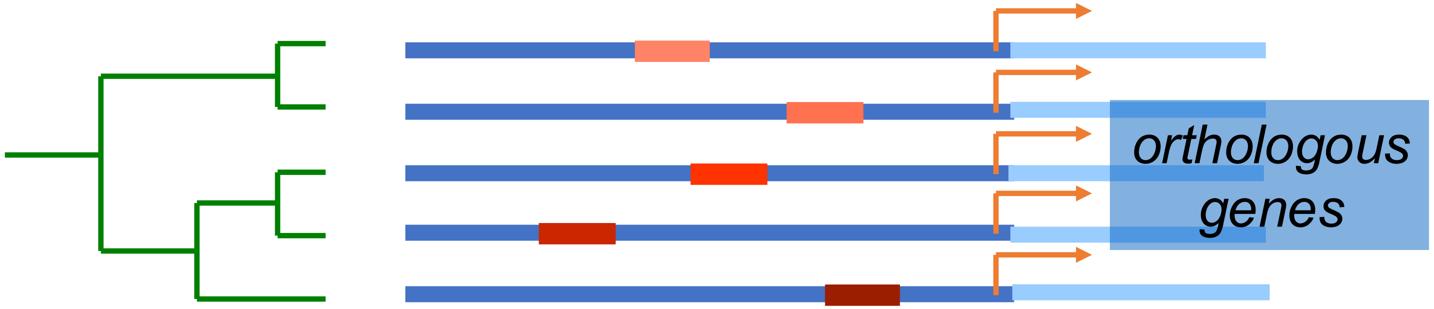
But notice two suboptimal results: we have one sequence with a placement but no genuine site, and one sequence with two sites but one placement. This is common enough to merit special treatment.

Summary

- The genome encodes much of its own regulation in protein binding sites
- A full description of the regulatory networks will require identifying these sites
- Compact descriptions of the DNA-binding preferences of TFs is afforded by weight matrices
- The information content of an alignment is a measure of specificity
- Weight matrix information for a TF is not enough to rule out false positives
- Multiple experimental techniques exist for identifying sequences harboring binding sites
- A variety of algorithms can be used to identify motifs in unaligned data

Motif Finding Algorithms Class II

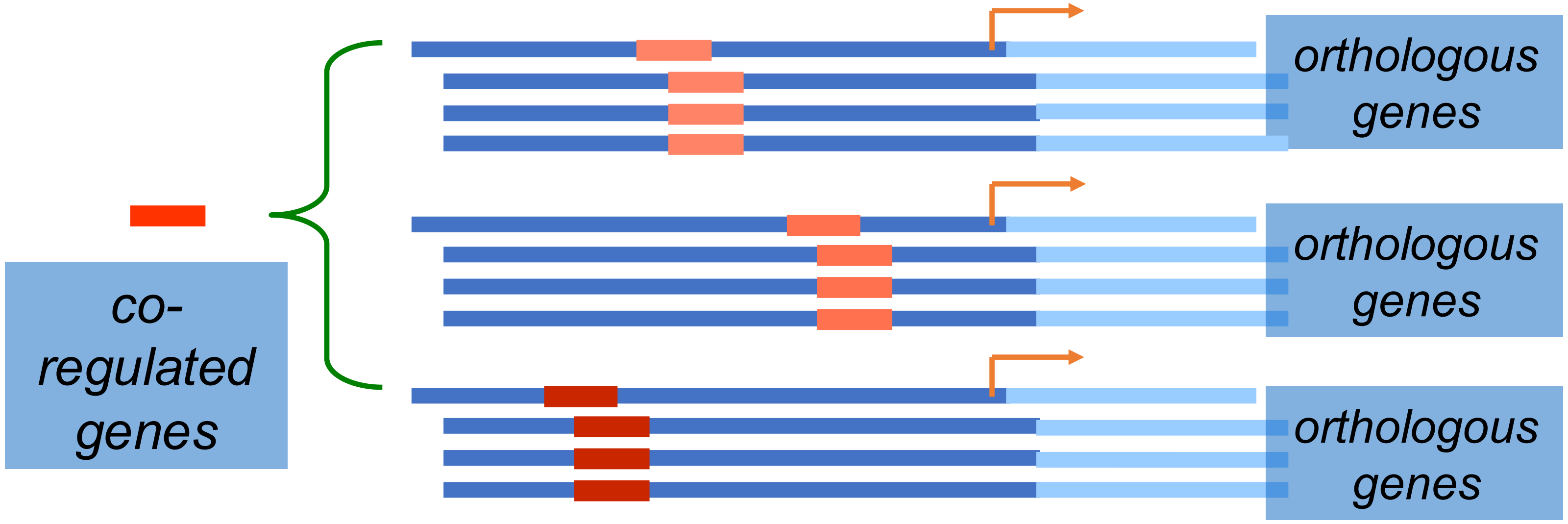
Single gene, multiple species (phylogenetic footprinting)



- orthologous background sequences
- sequences linked by a phylogenetic tree
- identify the “best conserved” motif that is under selective pressure

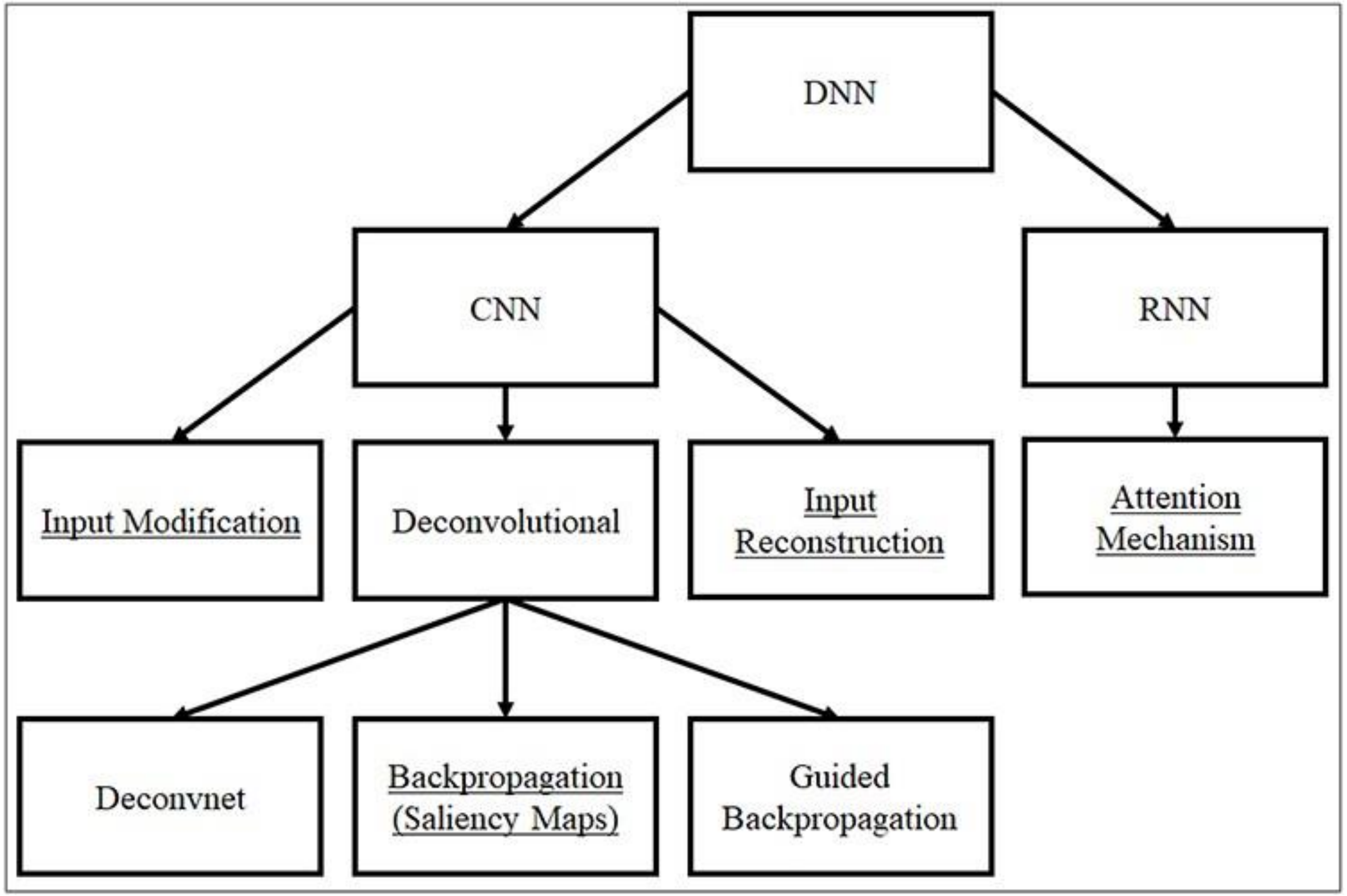
Motif Finding Algorithms Class III

Multiple genes, multiple species



- combination of phylogenetic data and gene regulation
- use phylogenetic data to reduce search space
- use correlation of motif occurences among orthologous genes to increase signal strength

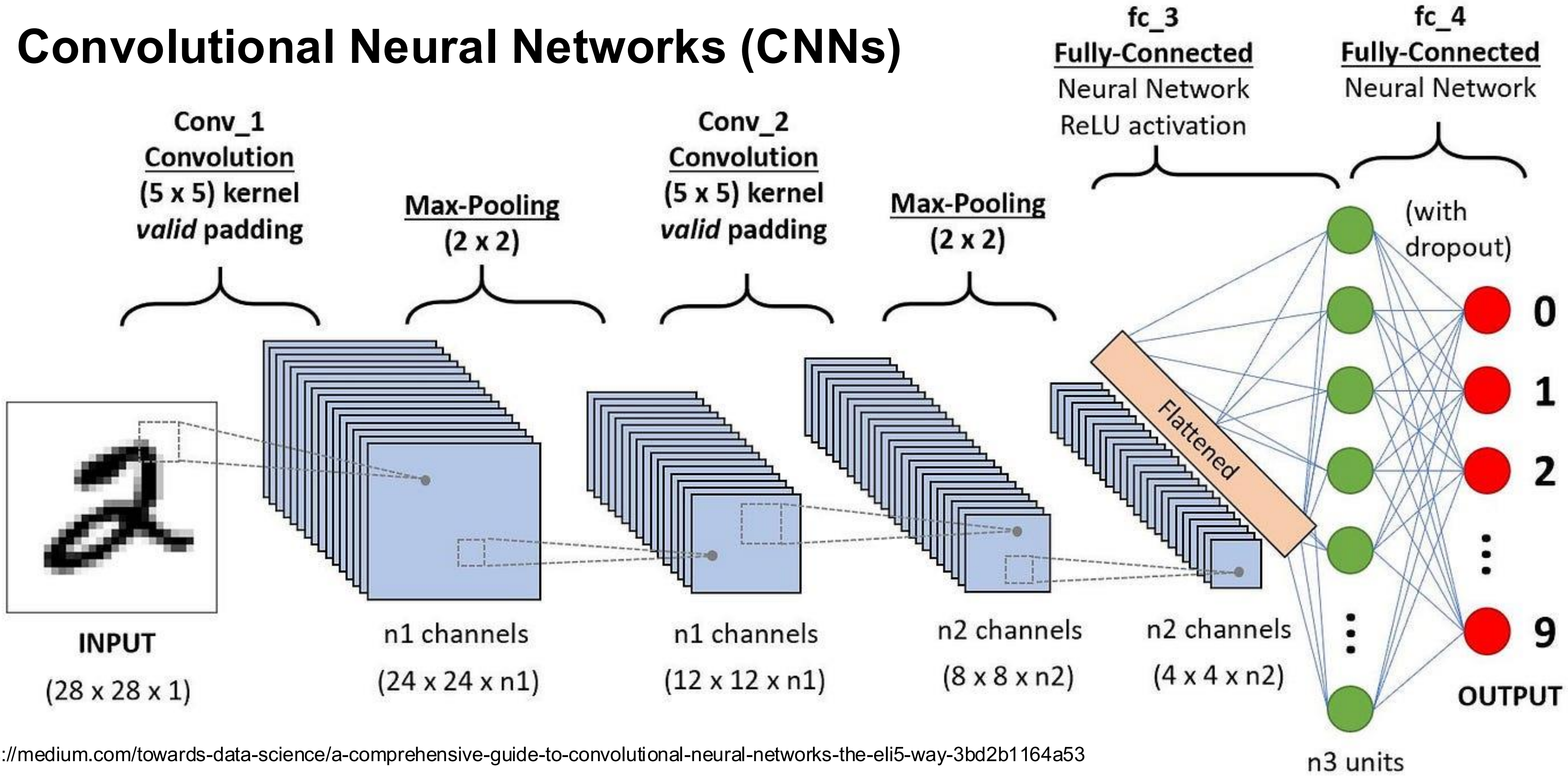
Deep learning approaches to motif discovery



Talukder et al. (2020)

Deep learning approaches to motif discovery

Convolutional Neural Networks (CNNs)



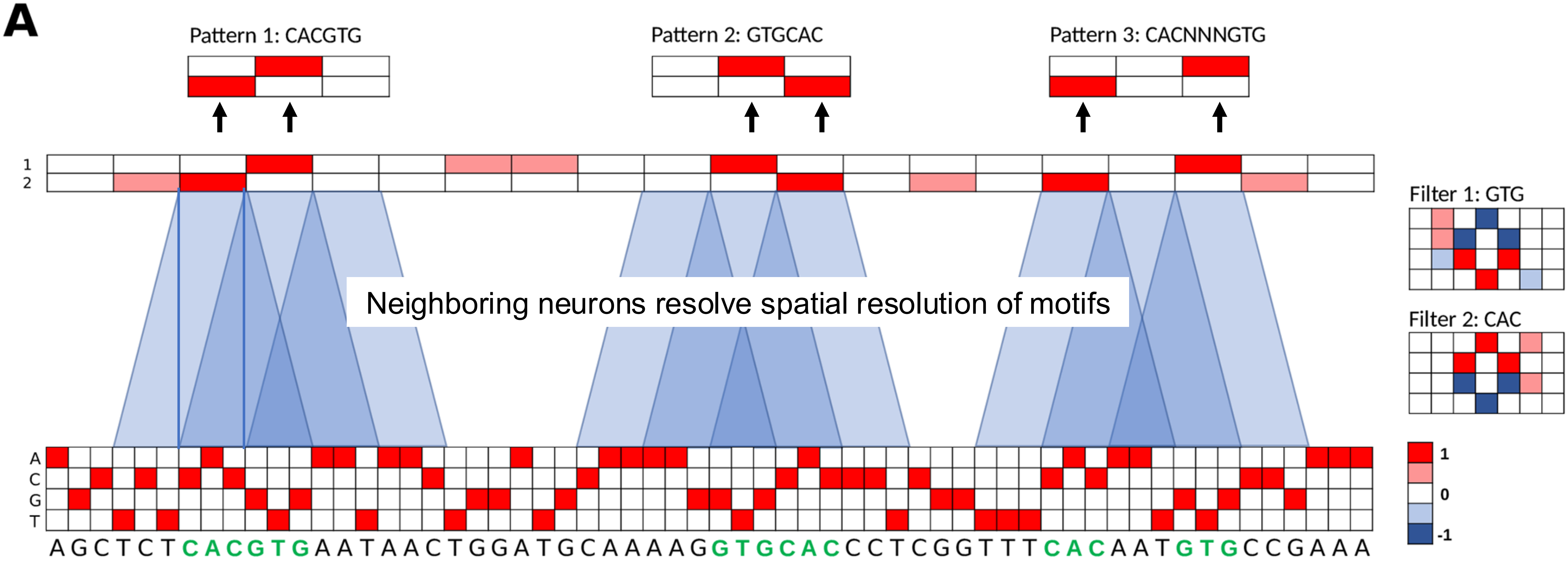
https://medium.com/towards-data-science/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53

Deep learning approaches to motif discovery

- Convolutional neural networks
  - Sequences are filtered through multiple **convolutional layers** (based on training sequences) and scored
  - Filtered sequence scores are **pooled** and max score retained
  - Many rounds of convolution->pooling can occur
  - Fully connected hidden layer used to score sequence
- Input data:
  - PBM/SELEX
  - Chromatin accessibility
  - ChIP-seq or CUT&RUN/Tag
  - Principle is to represent sequences that are biologically meaningful in training set

Deep learning approaches to motif discovery

Instance 1: Max pooling width = 3 nt; Receptive field = 9 nt

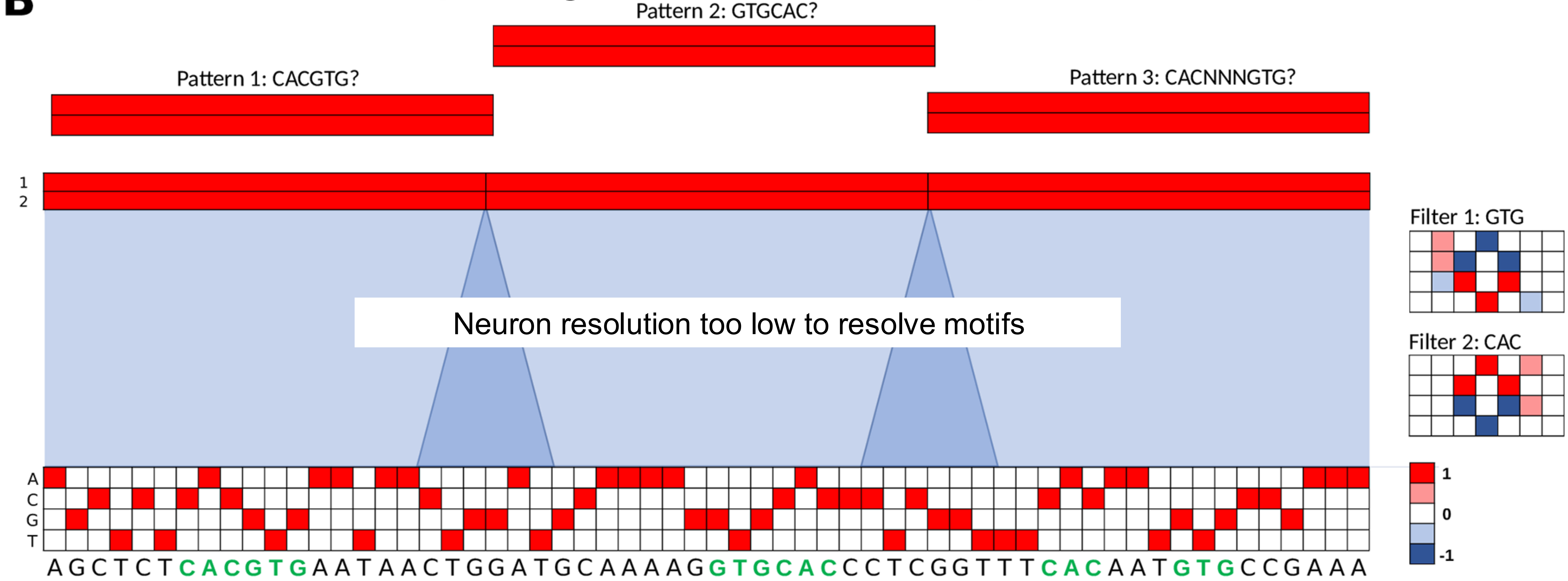

A
Pattern 1: CACGTG
Pattern 2: GTGCAC
Pattern 3: CACNNNGTG
1
2
Filter 1: GTG
Filter 2: CAC
1
0
-1
A
C
G
T
A G C T C T C A C G T G A A T A A C T G G A T G C A A A A G G T G C A C C C T C G G T T T C A C A A T G T G C C G A A A

Neighboring neurons resolve spatial resolution of motifs

Koo and Eddy (2019)

Deep learning approaches to motif discovery

Instance 2: Max pooling width of 20 nt; Receptive field = 26 nt



Neuron resolution too low to resolve motifs

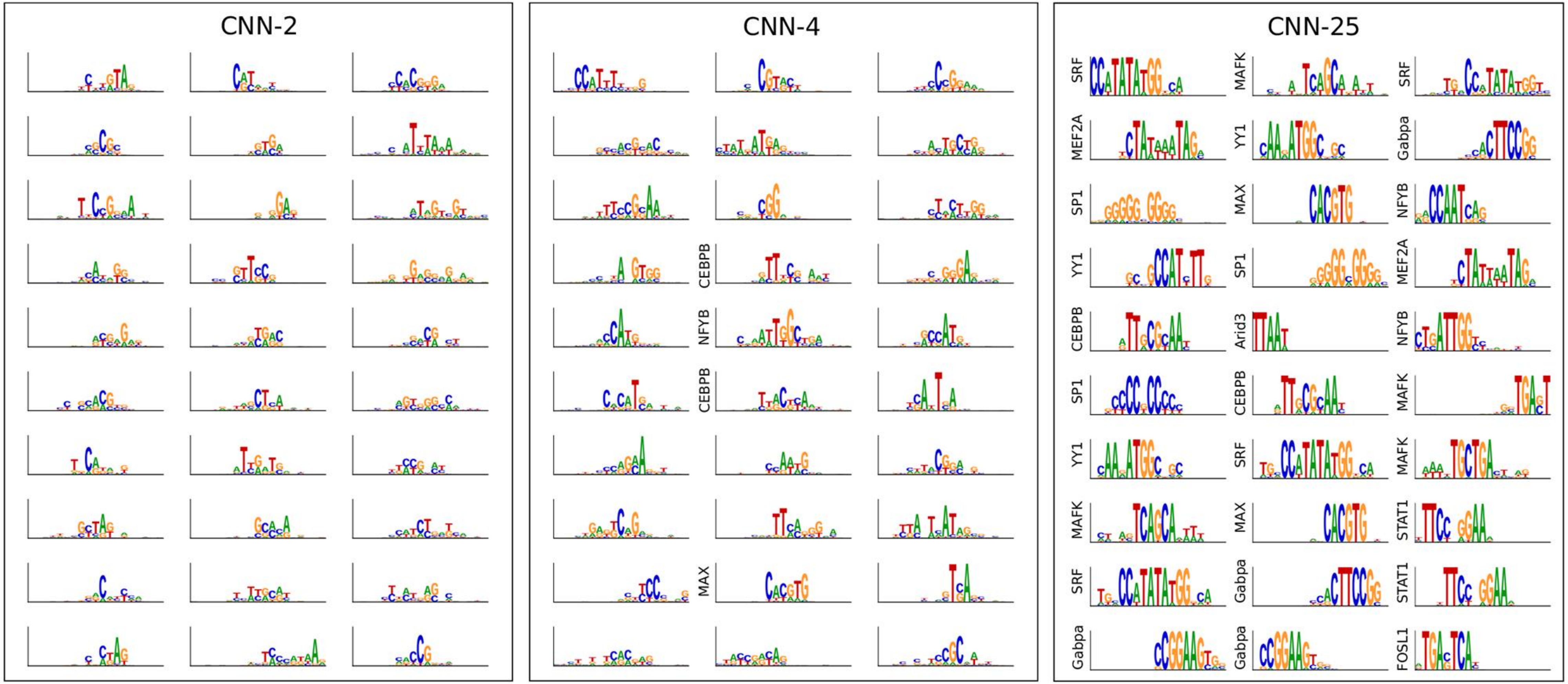
Koo and Eddy (2019)

Deep learning approaches to motif discovery

Small first filter

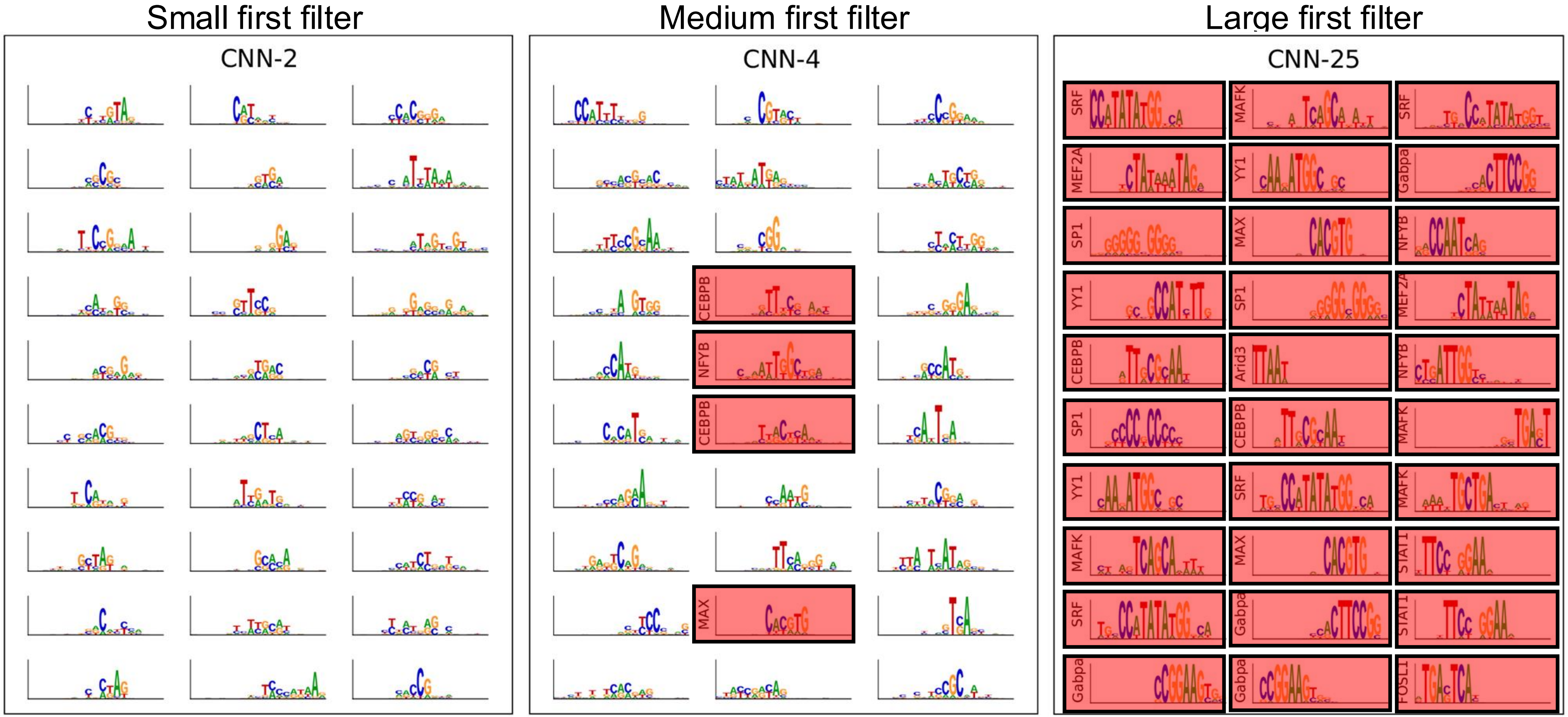
Medium first filter

Large first filter



Koo and Eddy (2019)

Deep learning approaches to motif discovery


Small first filter

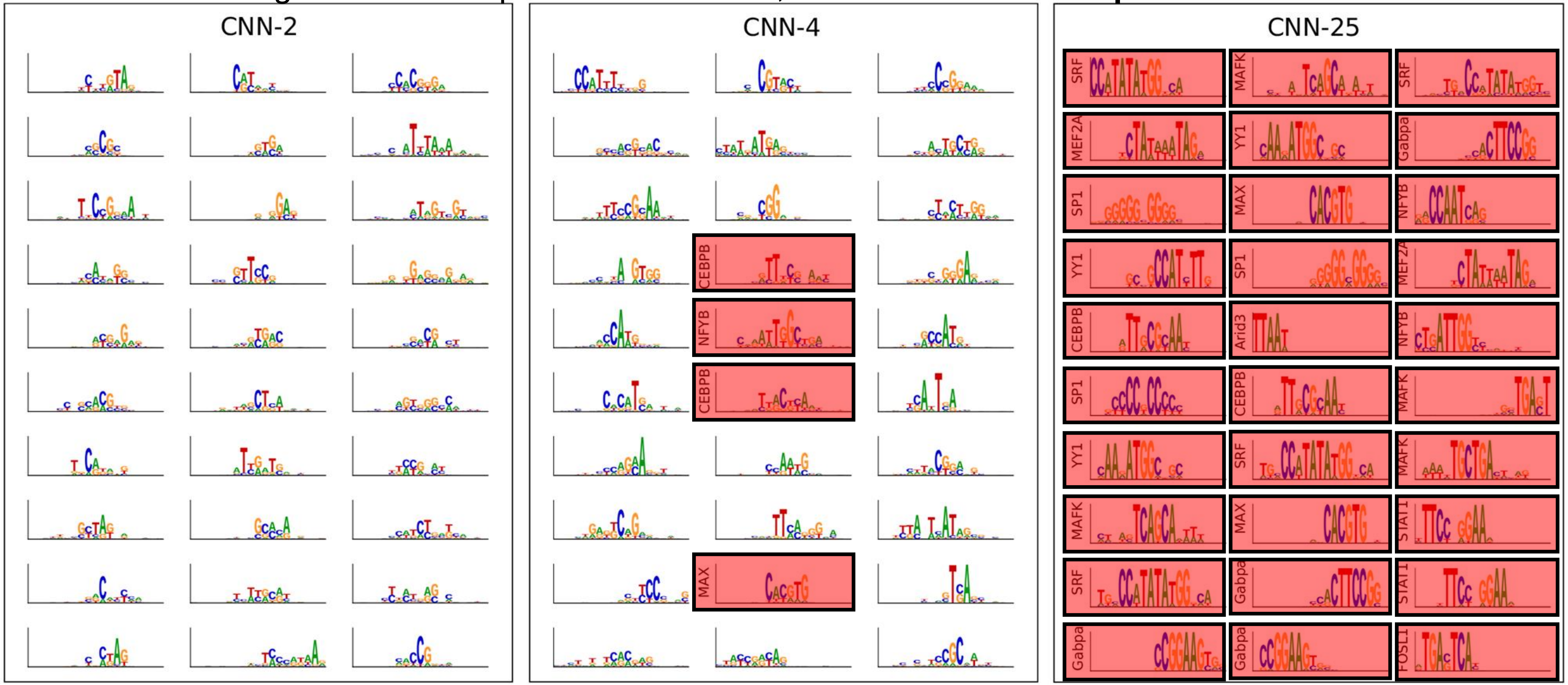
Medium first filter

Large first filter

Koo and Eddy (2019)

Deep learning approaches to motif discovery

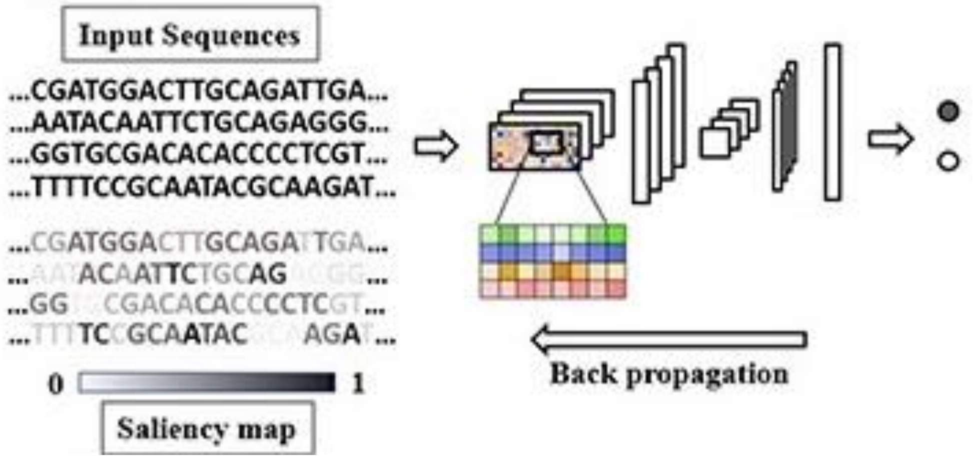
Large first filters represent **full motifs**; small first filters learn **partial motifs**



Koo and Eddy (2019)

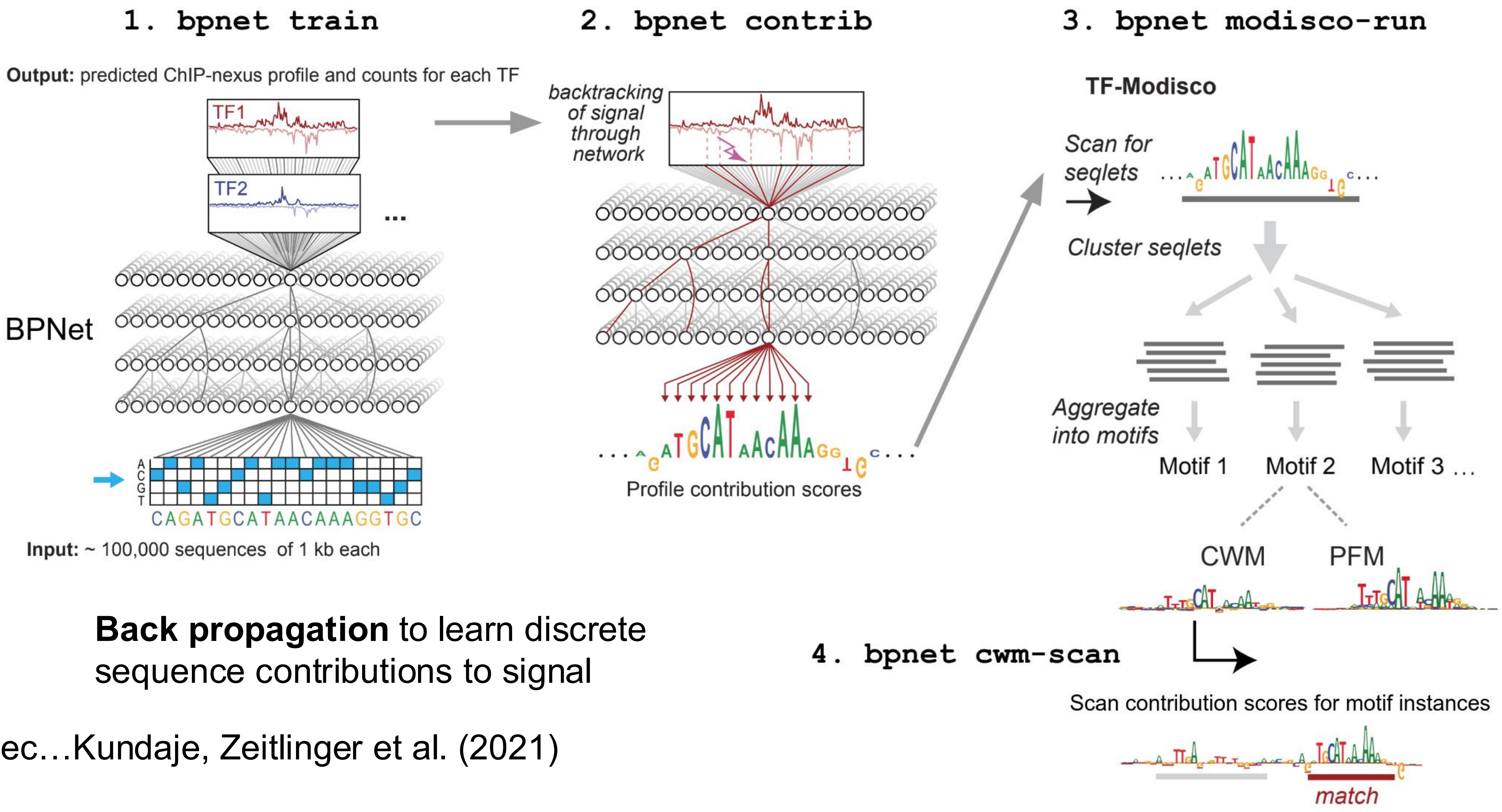
Deep learning approaches to motif discovery

- Back propagation: Determine the features being learned in early filter layers
- “Saliency maps” can be used to identify real features that the CNN deems important for prediction


Input Sequences
...CGATGGACTTGCAGATTGA...
...AATACAATTCTGCAGAGGG...
...GGTGCGACACACCCCTCGT...
...TTTTCCGCAATACGCAAGAT...
...CGATGGACTTGCAGATTGA...
...AATACAATTCTGCAGAGGG...
...GGTGCGACACACCCCTCGT...
...TTTTCCGCAATACGCAAGAT...
0
1
Saliency map
Back propagation

Talukder et al. (2020)

Deep learning approaches to motif discovery: BPNet



1. bpnet train

Output: predicted ChIP-nexus profile and counts for each TF

Input: ~ 100,000 sequences of 1 kb each

2. bpnet contrib

backtracking of signal through network

Profile contribution scores

3. bpnet modisco-run

TF-Modisco

Scan for seqlets

Cluster seqlets

Aggregate into motifs

Motif 1 Motif 2 Motif 3 ...

CWM PFM

4. bpnet cwm-scan

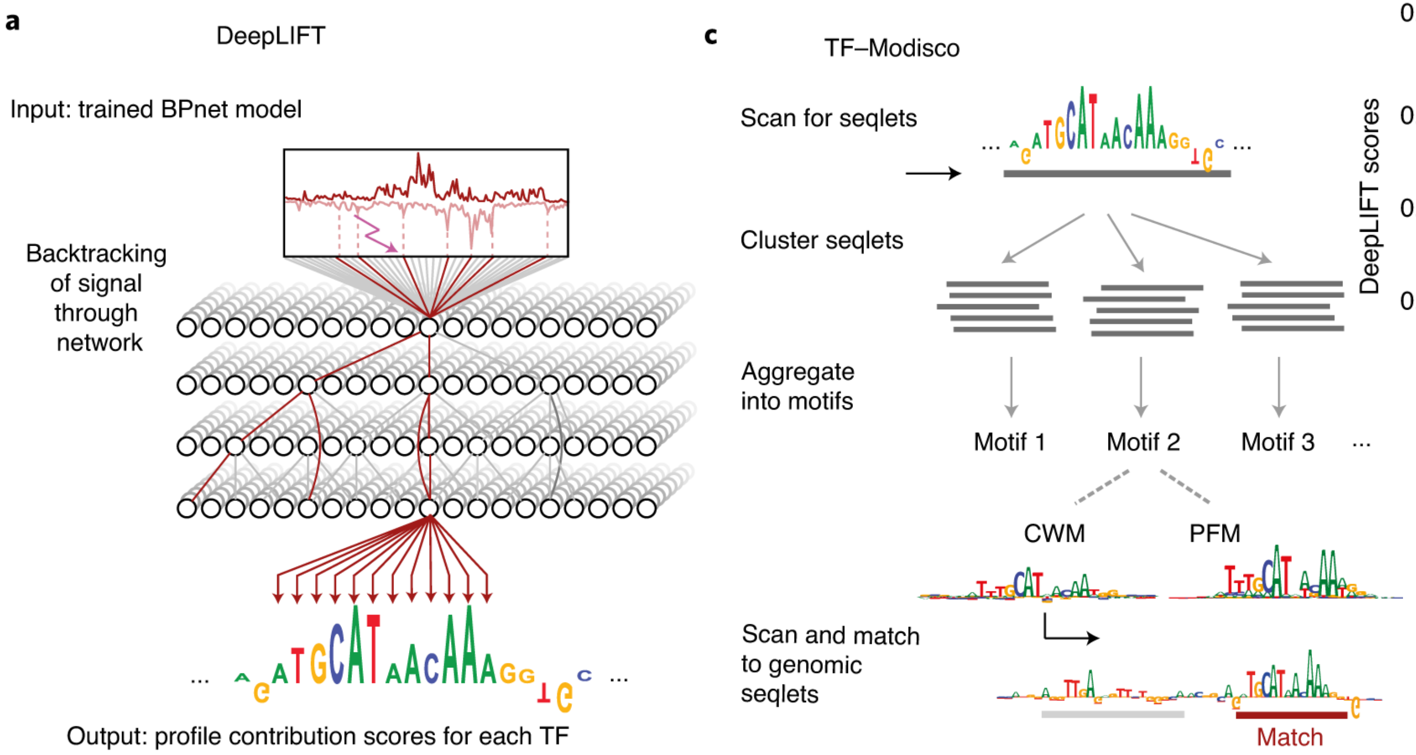
Scan contribution scores for motif instances

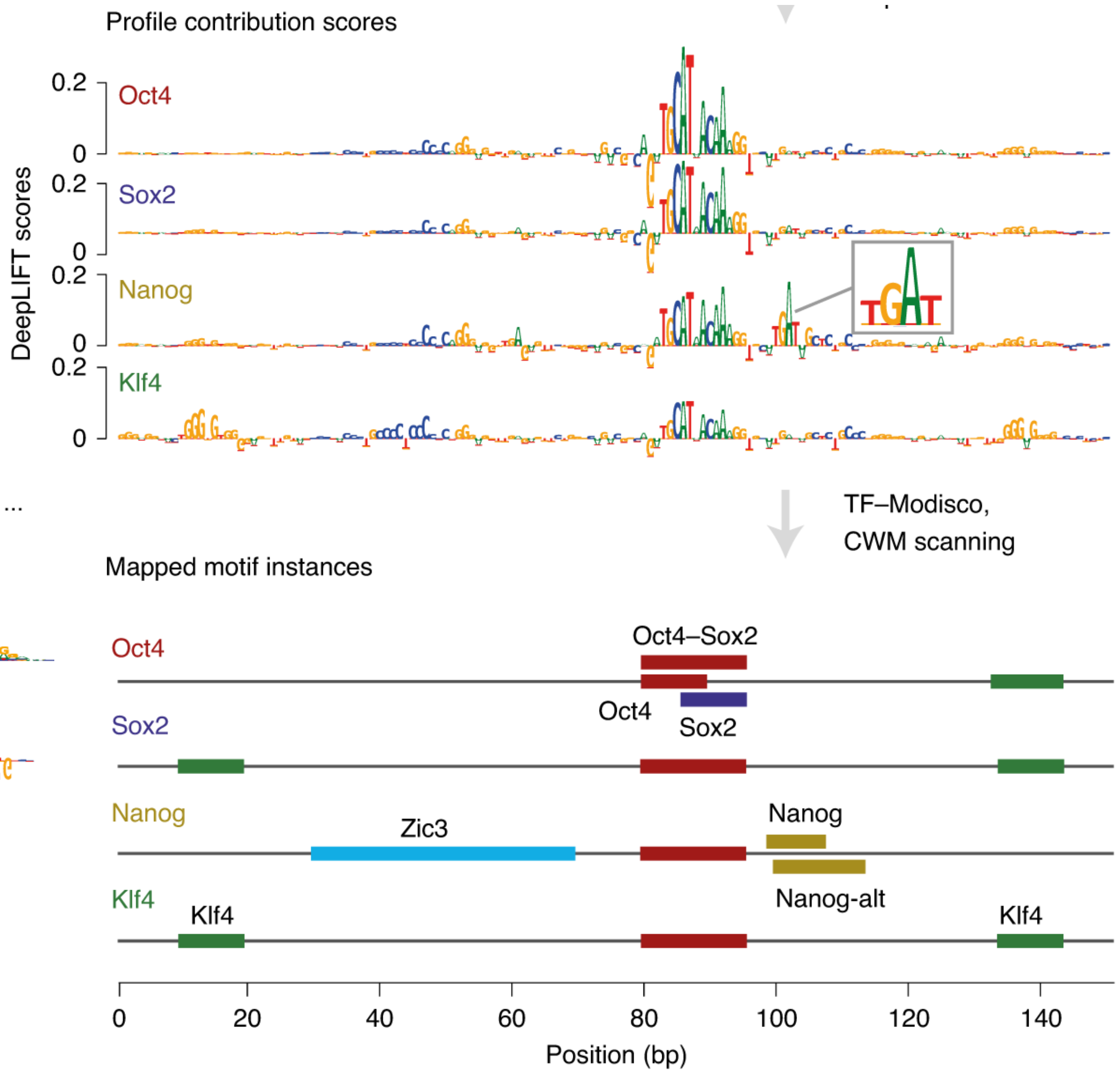
match

Back propagation to learn discrete sequence contributions to signal

Avsec…Kundaje, Zeitlinger et al. (2021)

Deep learning approaches to motif discovery: BPNet


a
DeepLIFT
Input: trained BPnet model
Backtracking of signal through network
Output: profile contribution scores for each TF
c
TF–Modisco
Scan for seqlets
Cluster seqlets
Aggregate into motifs
Motif 1
Motif 2
Motif 3
CWM
PFM
Scan and match to genomic seqlets
Match



Avsec…Kundaje, Zeitlinger et al. (2021)

Points of discussion

- Which motif finding algorithm is better?
  - Specific hypothesis
  - Binding site model
  - Search/optimization method and objective function
  - Some basic rules in practice
- Evolution of TF binding sites
  - Binding site turn over
  - Evolution by substitution, in/del, duplication, transposition
  - Co-evolution with TF
  - Impact on shaping regulatory networks
- Motif != binding != function
  - Sensitivity and specificity of wet/dry experiments
  - How to validate?
  - Biological function versus biochemical activity
- Species-specific regulation
- Beyond primary sequence conservation

Challenge of Specificity

- A 7-mer is expected to occur every 16,384 base pairs by chance
- In human, this means 3 X 10^9 / 16,384 ~ 180,000 sites in total
- TFBS are usually degenerative
- Total number of genes ~ 25,000
- Most of predicted binding sites are false positives!
- Need other restrictive information to reduce false positives