2. Binding of transcription factors

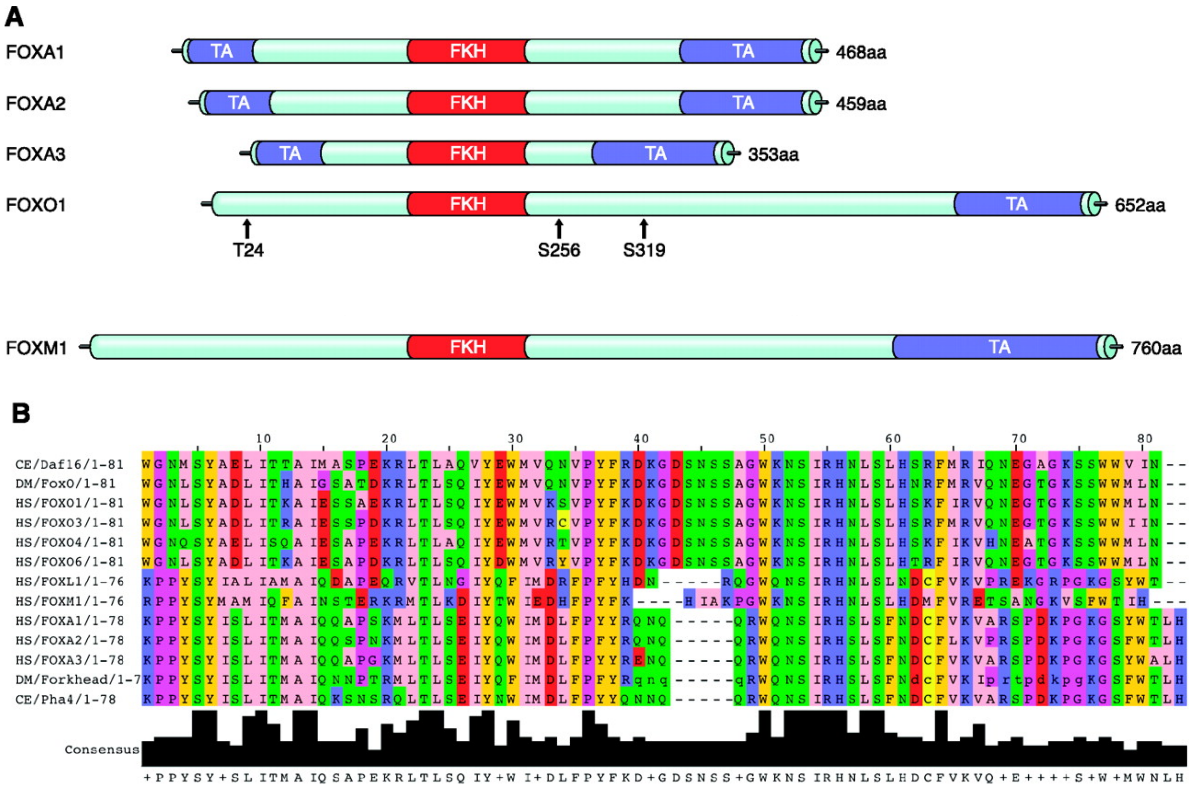
- protein-DNA interactions
- experimental approaches to determine binding sites (BS)
- representing binding specificity of TFs
- databases
- comparing binding profiles

IPMB
Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

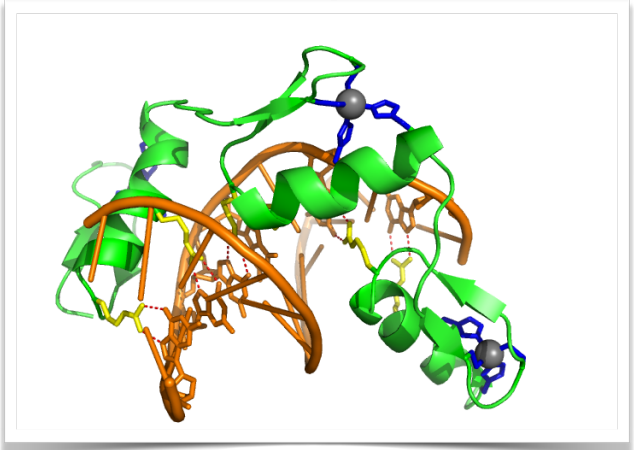
DNA binding domains

- Transcription factors contain a **DNA binding domain** (DBD) and a **transcriptional activator** (TA)
- Homologous TFs share similar DBDs (here: forkhead)



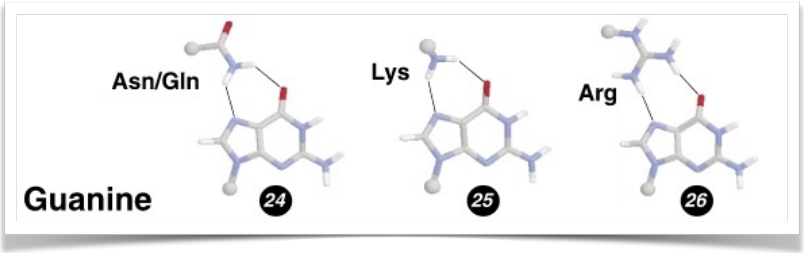
[Luscombe 2010]

Protein DNA interactions



Amino acids	Mode of interaction	Recognised base
Hydrogen bond		
[ARG, LYS]	Multiple-donor	G/complex
[HIS]	Multiple-donor (bifurcate)	G
[SER]	Multiple-donor (bifurcate)	G
	Acceptor+donor	complex
[ASN, GLN]	Acceptor+donor	A/complex
[ASP, GLU]	Multiple-acceptor	complex
van der Waals contacts		
[PHE, PRO]	Ring-stacking	A, T
[THR]	Methyl contact	T
[GLY, ALA, VAL, LEU, ISO, TYR]	-	many (non-specific)
No base contact		
[CYS, MET, TRP]	-	-

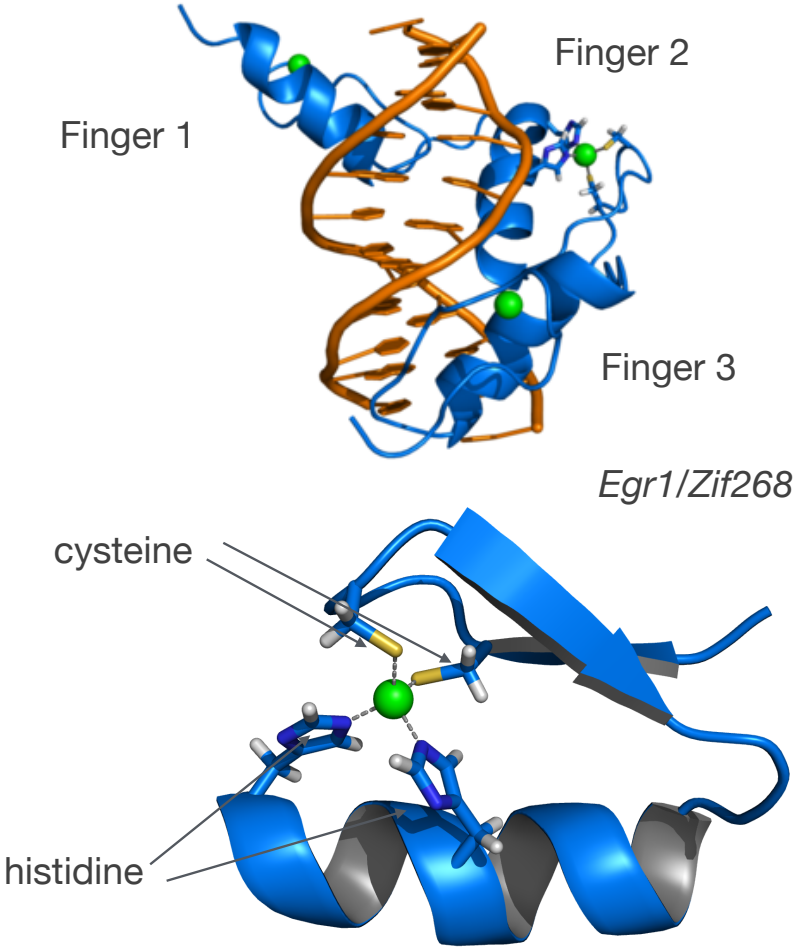
- majority of protein-DNA interactions for TF occur through a **alpha-helix** fitting into the major groove (=DNA binding domain)
- **hydrogen bonds** with specific bases
- stabilization of the protein-DNA complex is ensured by additional structures (helix, beta-sheet) via **van der Walls** interactions



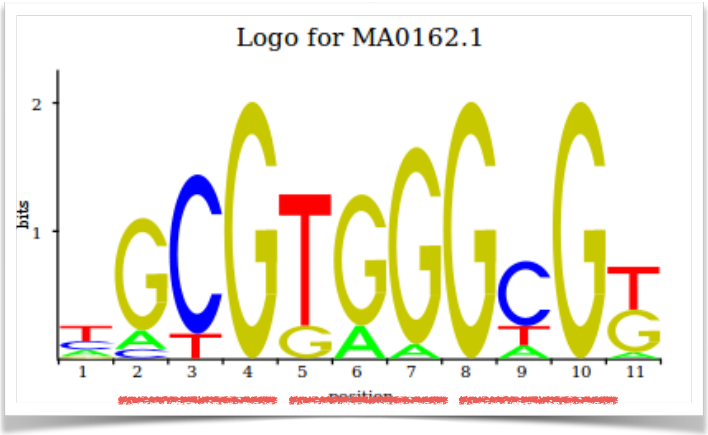
[...] while there appear to be favoured interactions, the specificity for the entire DNA sequence can rarely be explained by one-to-one correspondences between amino acids and bases.

[Luscombe et al., NAR (2001)]

Structural family: Zinc coordinating



Cys2His2 Fold (“Zinc finger”)
→ one of the most common family
of transcription factors in mammalians

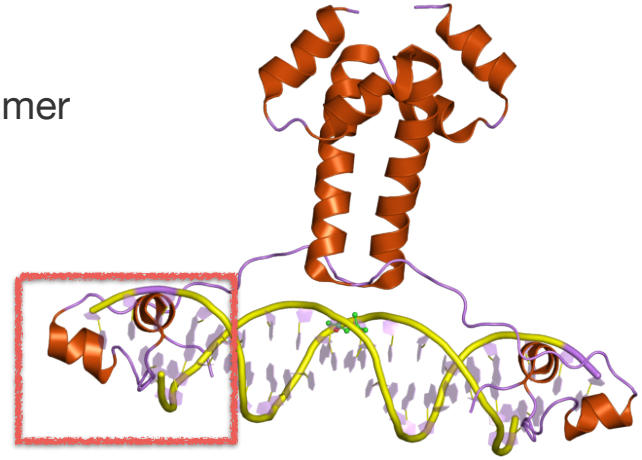


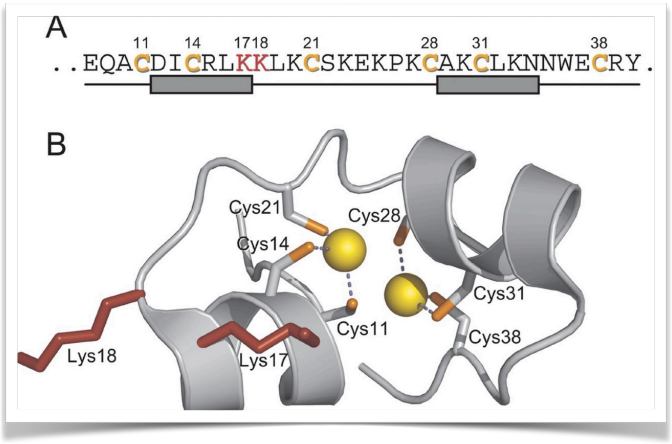
Finger 1 Finger 2 Finger 3

Structural family: Zinc coordinating

Gal4 *S. cerevisae*

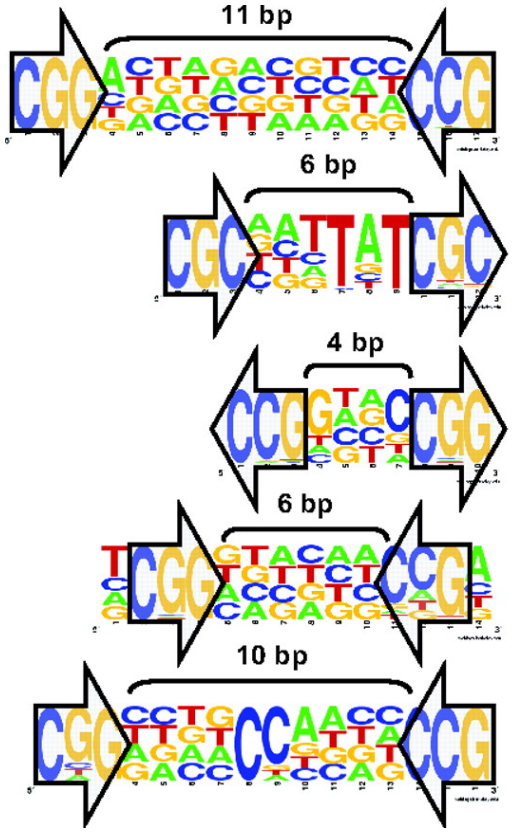
→ binds as homodimer





[Campbell (2008)]

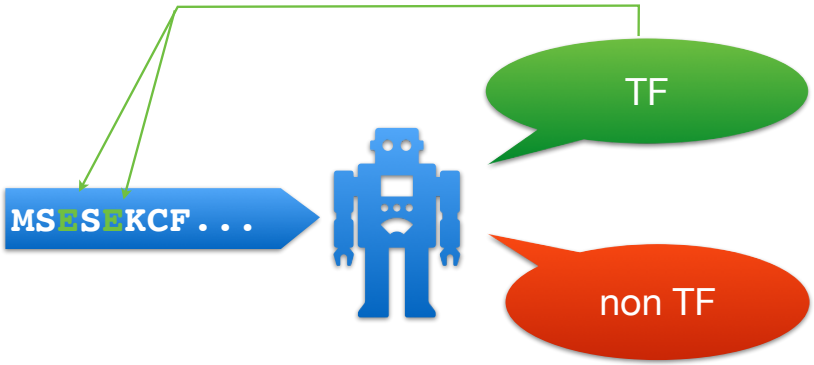
- 6 cysteins
- 2 Zinc atoms
= **Cys6Zn2 zinc-fingers**

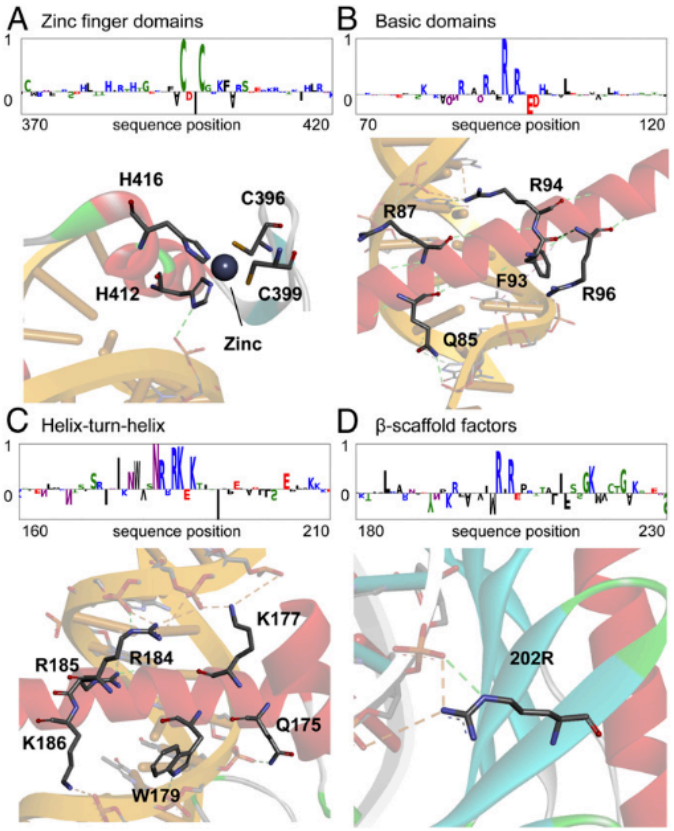


TF binding site is a **dyad** (short repeated motif with spacer)

Predicting TFs from protein sequence

- No prediction of binding site from protein sequence (yet...)!
- However, one can **predict if a protein is a TF** using only the protein sequence using deep learning approaches: **DeepTFactor**

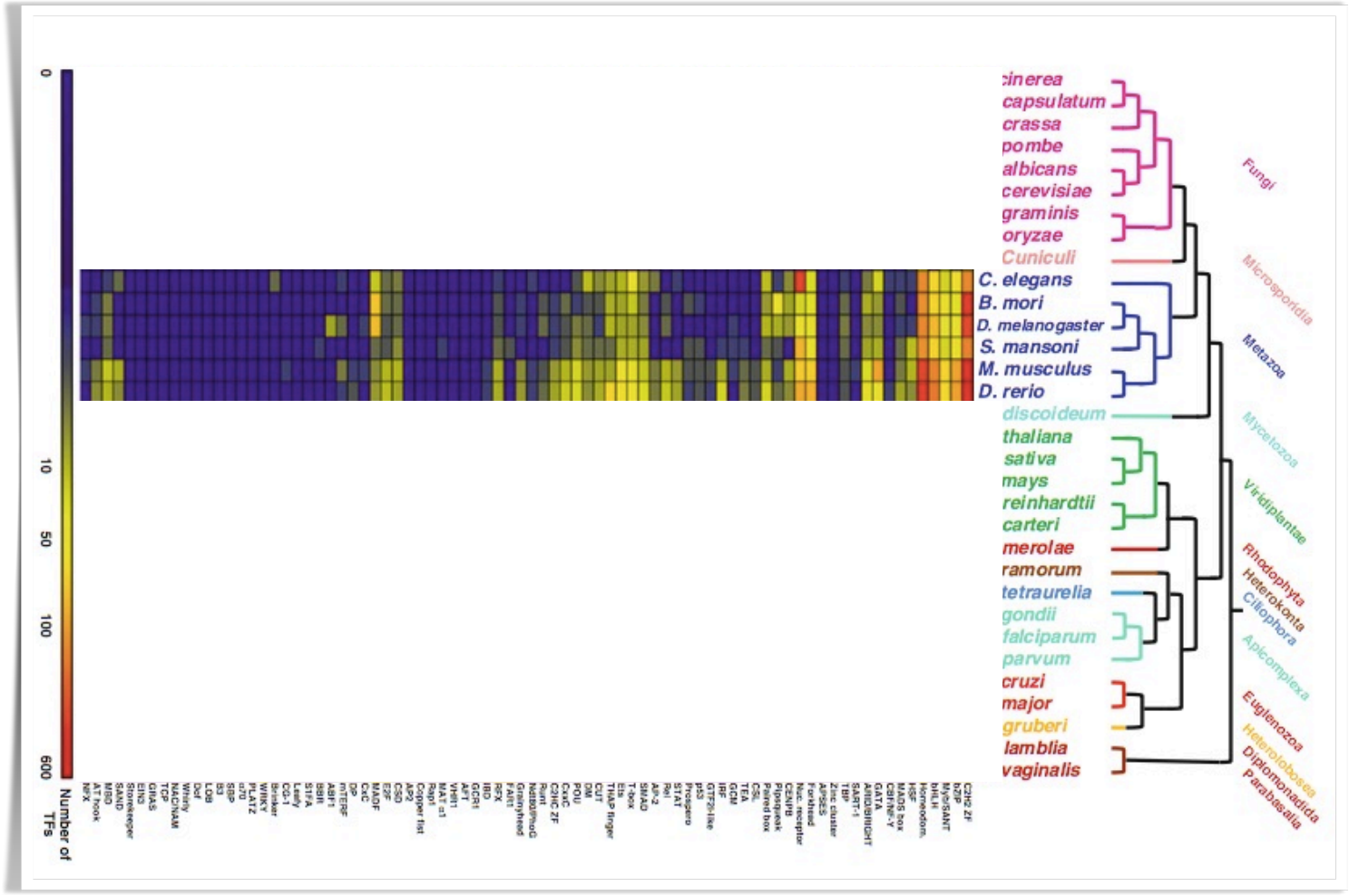




prediction of residues which are contributing highly to the classification of TFs

[Kim et al., PNAS (2020)]

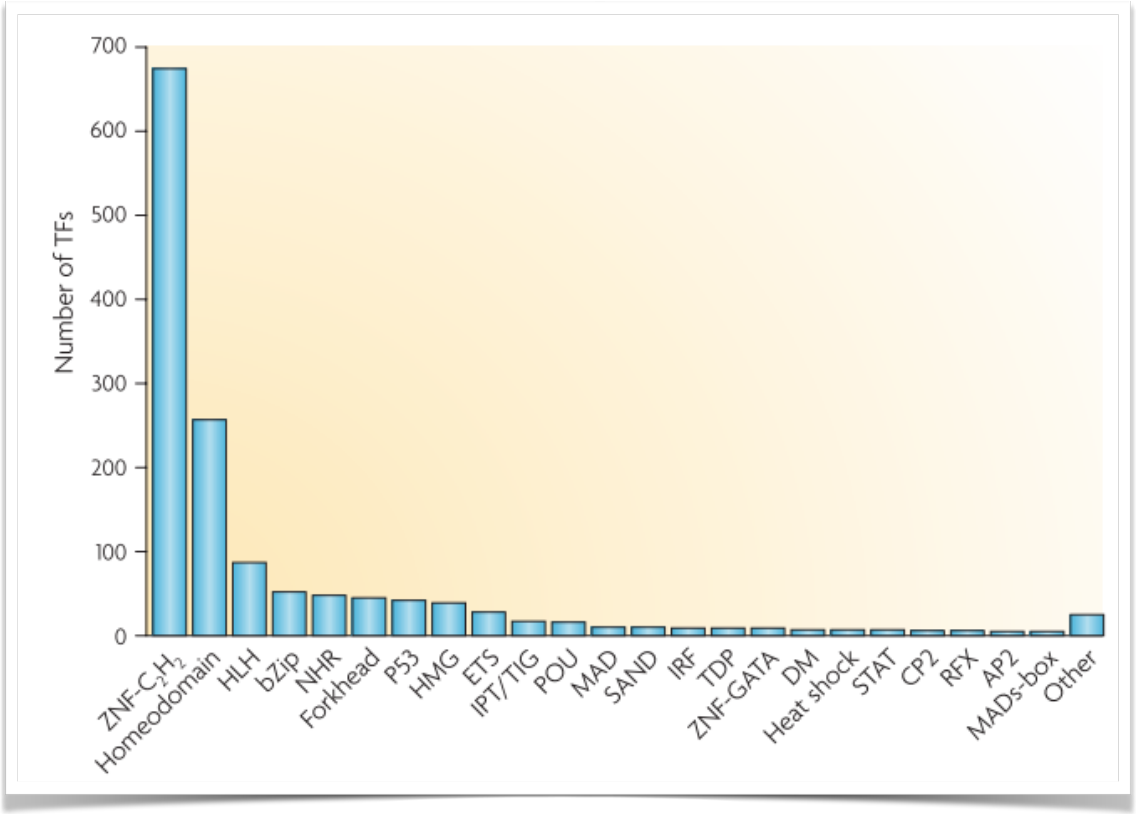
Structural families of TFs



[Weirauch, Hughes (2010)]

Structural families of TFs

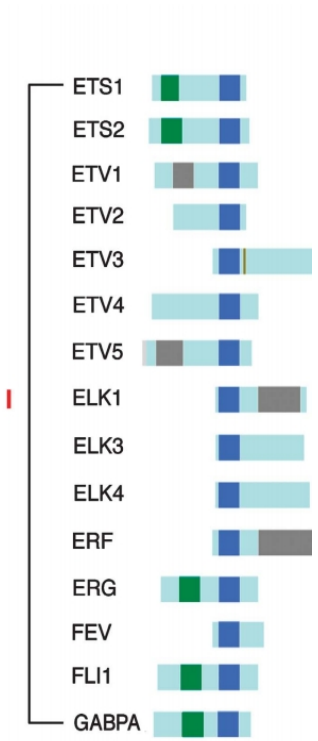
- in human, 3 classes account for > 80% of known transcription factors
  - Zinc-finger **C2H2**
    (→ Zinc finger nucleases)
  - **homeodomain**
    (e.g. Hox TFs)
  - **helix-loop-helix factors**
    (e.g. c-Myc, N-Myc,...)



[Vaquerizas et al. Nat.Rev.Gen (2009)]

From protein to DNA motif

- Transcription factors belonging to the same structural family have **similar binding profiles**
- Highly conserved accross species
- conserved **core motif**
- variable flanking regions
  → specificity



[Wei et al. EMBO Journal 2010]

2. Binding of transcription factors

- protein-DNA interactions
- experimental approaches to determine binding sites (BS)
- representing binding specificity of TFs
- databases
- comparing binding profiles

IPMB
Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Proc. Nat. Acad. Sci. USA
Vol. 70, No. 12, Part I, pp. 3581–3584, December 1973

The Nucleotide Sequence of the *lac* Operator

(regulation/protein–nucleic acid interaction/DNA–RNA sequencing/oligonucleotide priming)

WALTER GILBERT AND ALLAN MAXAM

Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138

Communicated by J. D. Watson, August 9, 1973

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:**

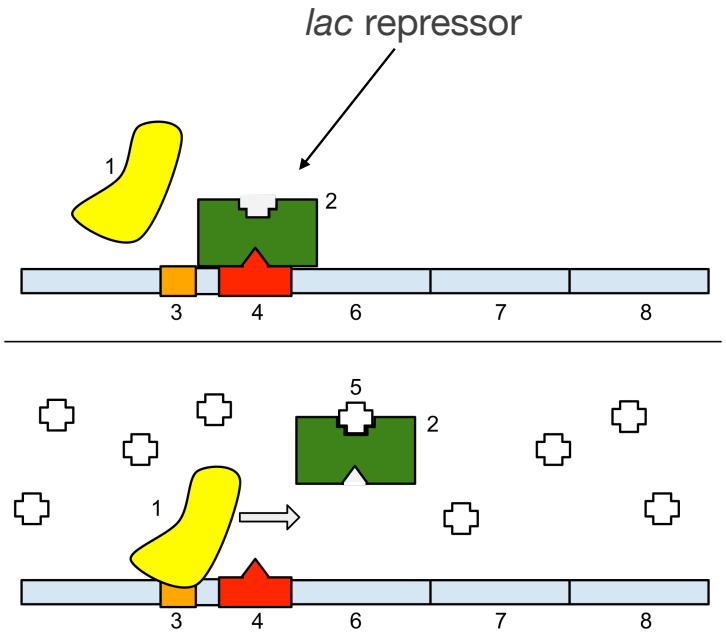
5′--T G G A A T T G T G A G C G G A T A A C A A T T 3′
3′--A C C T T A A C A C T C G C C T A T T G T T A A 5′

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

The lactose repressor selects one out of six million nucleotide sequences in the *Escherichia coli* genome and binds to it to prevent the expression of the genes for lactose metabolism. bind again to the repressor, and is about 27 base-pairs long. Here we shall describe its sequence.

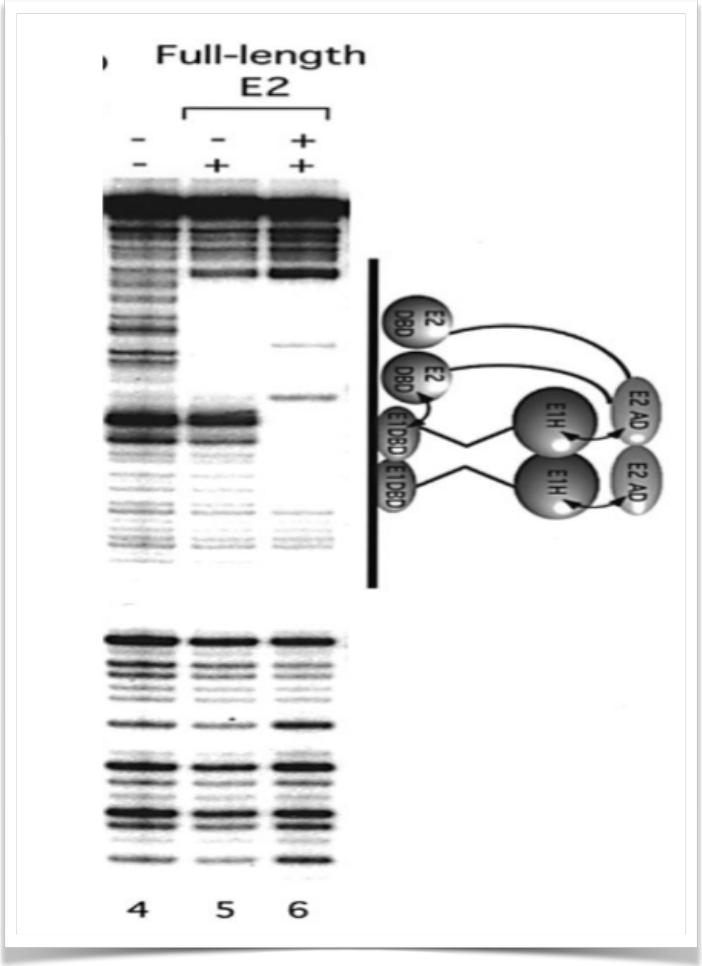
METHODS

Sonicated DNA Fragments. Sonicated [^{32}P]DNA fragments were made by growing a temperature-inducible lysogen of λc1857p*lac*5S7 at 34° in a glucose–50 mM Tris·HCl or TES (pH 7.4) medium in 3 mM phosphate, heating at 42° for 15 min at a cell density of 4×10^8/ml, then washing and resuspending the cells at a density of 8×10^8/ml in the same medium with 0.1 mM phosphate. 100 mCi of neutralized $H_3{}^{32}PO_4$ was added to 10 ml of cells, and the incorporation was continued for 2 hr at 34°. The cells were washed, suspended in 2



Experimental identification of binding sites

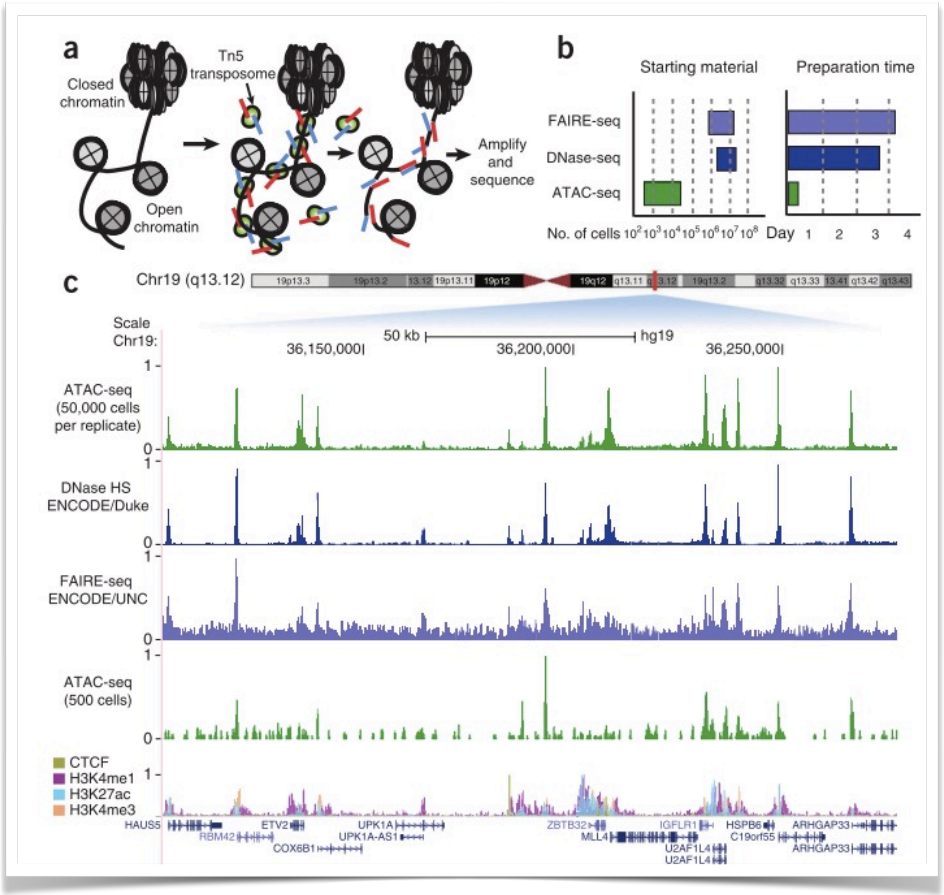
- **DNAse1 footprint:**
  protein-bound DNA is protected from digestion by the DNAse1 enzyme
- gel electrophoresis allows identification of the digested / undigested fragments
- if whole sequence is known, binding sites can be identified
- in vitro / low-throughput



[A.Stenlund The EMBO Journal (2003)]

Experimental identification of binding sites

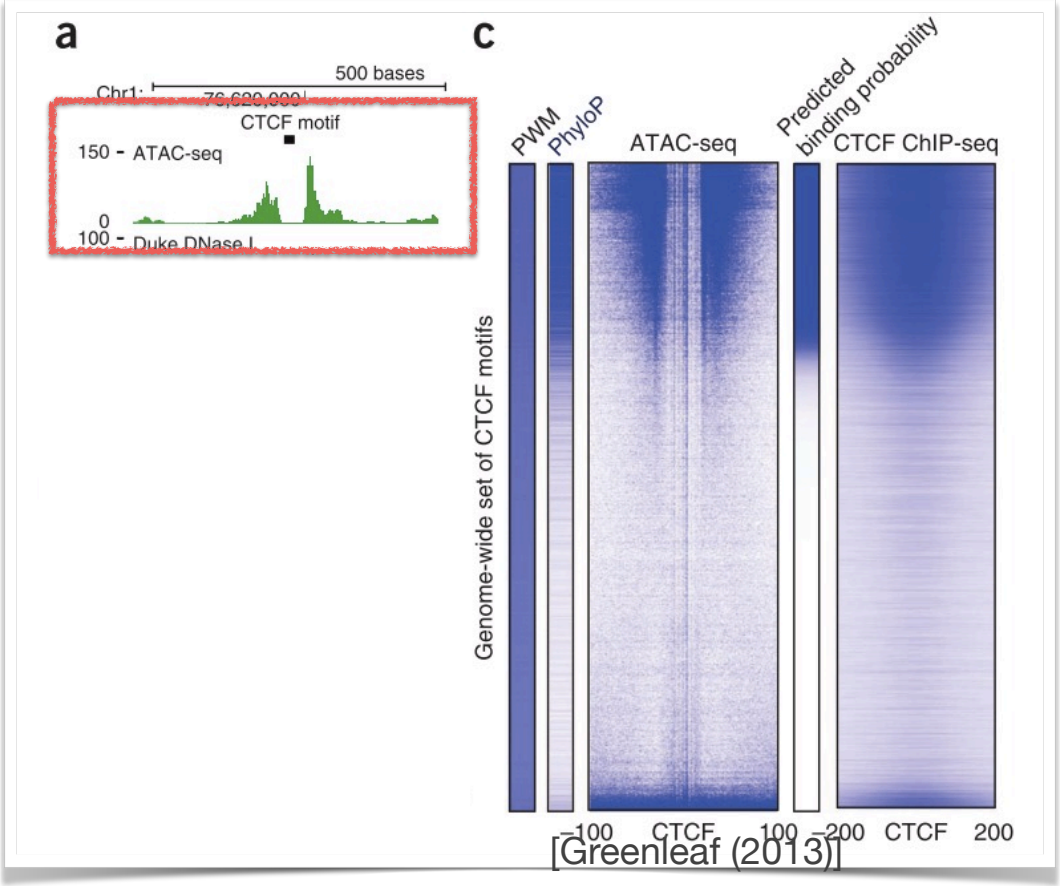
- **ATAC-seq**: using Tn5 transposase prepared with sequencing primers
- requires a small number of input material (~10,000 cells)
- identification of open chromatin regions (peaks)



[Greenleaf (2013)]

Experimental identification of binding sites

- From open regions to transcription factor binding sites
  → **footprinting**
- Zooming into the peaks (open regions) : valleys of undigested / un-transposed DNA
  → **TF binding sites (TFBS)**
- binding sequence can be identified with base-pair resolution

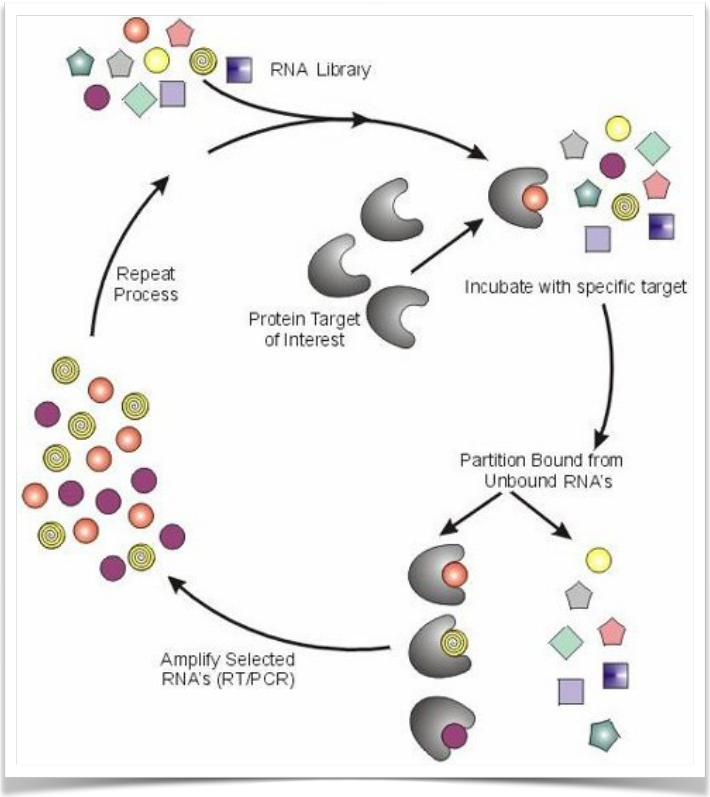


[Greenleaf (2013)]

Experimental identification of binding sites

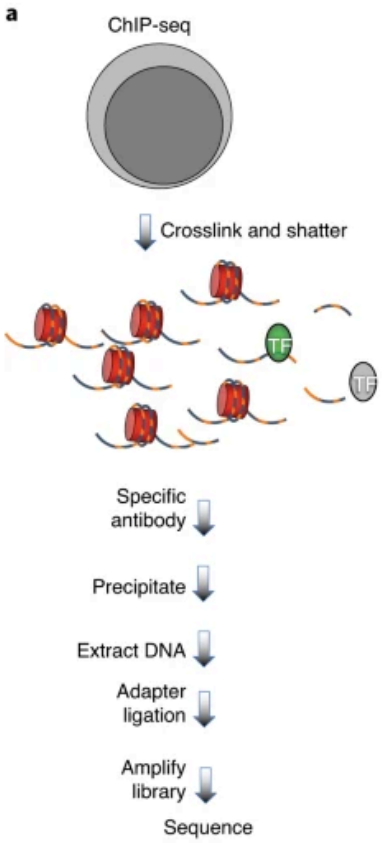
- **SELEX** (= *systematic evolution of ligands by exponential enrichment*)
  - random library of k-mers generated
  - incubated with TF of interest
  - elution to separate bound from unbound targets
  - PCR (=enrichment)
  - next cycle

- in-vitro / medium throughput
- increasing number of cycles leads to **over-select high affinity binding sites**



[A.Stenlund The EMBO Journal (2003)]

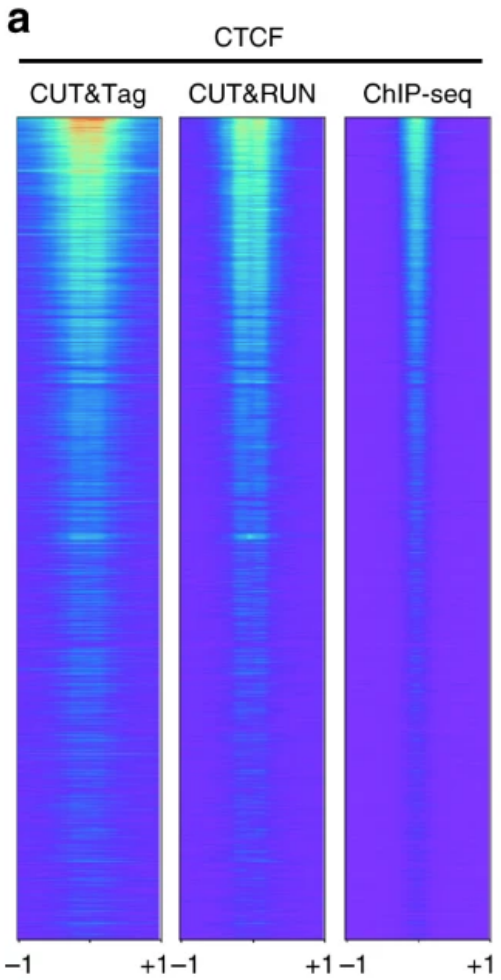
Antibody-based assays



[Kaya-Okur et al., Nat.Prot. (2020)]

Antibody-based assay

- Cut & Tag has **higher signal-to-noise** ratio compared to ChIP-seq
- Can be performed under **native conditions** → adaptable to single-cell
- **Greater resolution** to detect the actual binding site of the TF



[Kaya-Okur et al., Nat.Com. (2019)]

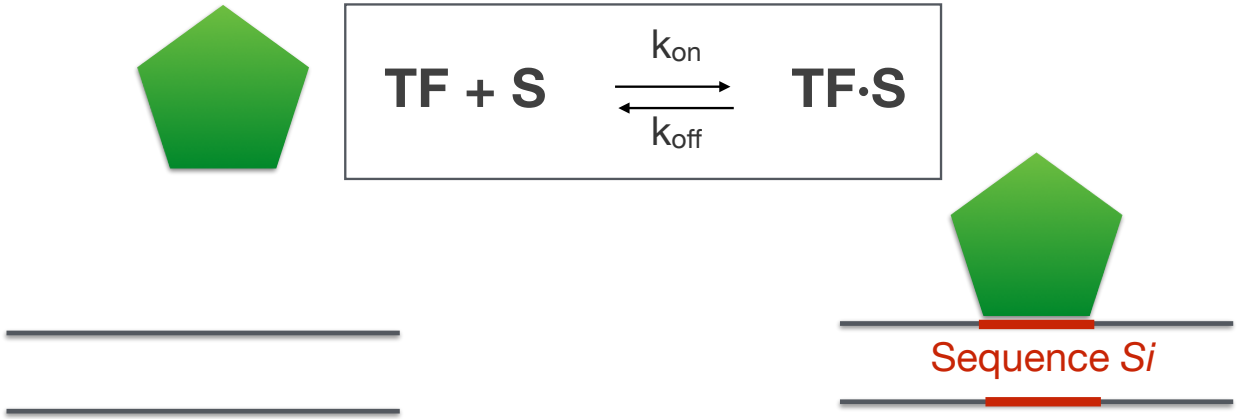
2. Binding of transcription factors

- protein-DNA interactions
- experimental approaches to determine binding sites (BS)
- representing binding specificity of TFs
- databases

IPMB
Institut für Pharmazie und Molekulare Biotechnologie

UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

DNA transcription factor interactions



Equilibrium constant

$$K_i = \frac{k_{on}}{k_{off}} = \frac{[TF \cdot S_i]}{[TF][S_i]}$$

Binding probability

$$P(s=1|S_i) = \frac{[TF \cdot S_i]}{[TF \cdot S_i] + [S_i]} = \frac{1}{1 + e^{E_i - \mu}}$$

$$E_i = -\ln K_i$$

free energy

$$\mu = \ln[TF]$$

chemical potential

DNA transcription factor interactions

- How many parameters are needed to describe the DNA-protein interaction ?

Naive approach

- if TF binds regions of size w, then if one K_i per possible sequence
  - 4^w parameters
  - w = 6 : 4096 parameters
  - w = 8 : 65536 parameters

Model :

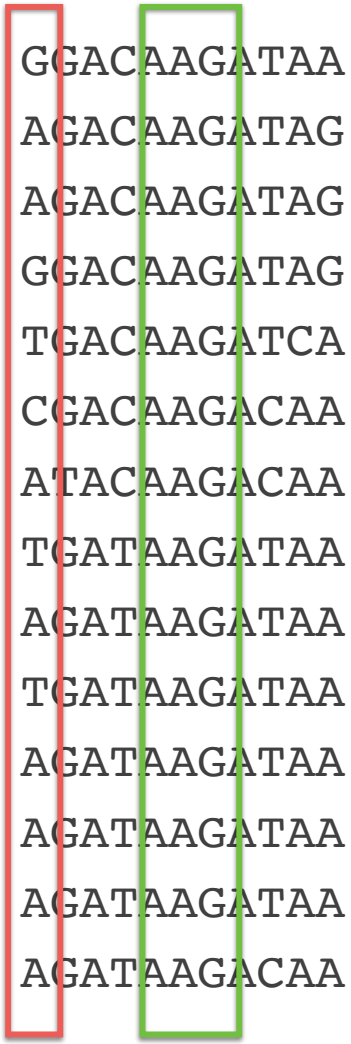
- Assume independent contribution of each position of S_i to the binding energy
- contribution proportional to **log(freq at position j)**
  → 3w parameters
- w = 6 : 18 parameters
- w = 8 : 24 parameters

[Berg, von Hippel (1987)]
[Stormo, Bioinformatics (2000)]

Characterizing binding affinities

- **Example** :
  - Evi-1 mouse transcription factor
  - 14 binding sites obtained using SELEX

- Binding sites of transcription factors are not fixed strings !



Some positions are very degenerate

Some positions are well conserved

Characterizing binding affinities

- **Solution 1:**
  represent binding sites as **consensus sequence**
- **IUPAC conventions**
  - `W = A or T ; R = A or G (puRines)`
  - `K = G or T ; S = C or G`
  - `Y = C or T (pYrimidine) ; M = A or C`
  - `B = C, G or T ; D = A,G or T`
  - `H = A,C or T ; V = A,C or G`
  - `N = any nucleotide`
- **Transfac conventions**
  - for a given position, order nucleotides by frequency : f1>f2>f3>f4
  - **single** nucleotide if f1 > 50% AND f1>2*f2
  - **double** if f1 + f2 > 75% AND f1<50 %; f2<50 %
  - **triple** if one nucleotide never appears and previous rules don't apply
  - N in all other cases

Characterizing binding affinities

- **Solution 1:**
  represent binding sites as consensus sequence
- **Weakness 1:**
  - column 1 has 8/14 A's
  - column 10 has 13/14 A's
  - both are represented by A
- **Weakness 2:**
  - other nucleotides in column 1 are not taken into account

Characterizing binding affinities

- **Solution 2 :**
- count frequencies of nucleotides at each position
- obtain a count matrix
- represent the **matrix** as a **logo**

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

a	8	0	14	0	14	14	0	14	0	13	11
c	1	0	0	7	0	0	0	0	3	1	0
g	2	13	0	0	0	0	14	0	0	0	3
t	3	1	0	7	0	0	0	0	11	0	0

Characterizing binding affinities

- **Solution 2 :**
- count frequencies of nucleotides at each position
- normalize to obtain **position frequency matrix (PFM)**

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

$f_{i,j}$

a|	0.57	0.00	1.00	0.00	1.00	1.00	0.00	1.00	0.00	0.93	0.79
c|	0.07	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.21	0.07	0.00
g|	0.14	0.93	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.21
t|	0.21	0.07	0.00	0.50	0.00	0.00	0.00	0.00	0.79	0.00	0.00

Characterizing binding affinities

- ***Beware of sampling effects !***
- Two different sets of BS for the same TF will give different PFM
- Example : Evi-1
  - set 1 : 14 BS (selex)
  - set 2 : 47 BS (selex)

a|	0.57	0.00	1.00	0.00	1.00	1.00	0.00	1.00	0.00	0.93	0.79
c|	0.07	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.21	0.07	0.00
g|	0.14	0.93	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.21
t|	0.21	0.07	0.00	0.50	0.00	0.00	0.00	0.00	0.79	0.00	0.00

a|	0.55	0.74	0.00	1.00	0.02	1.00	0.98	0.00	1.00	0.00	0.87	0.85	0.23	0.64
c|	0.09	0.04	0.02	0.00	0.43	0.00	0.00	0.00	0.00	0.13	0.04	0.00	0.23	0.21
g|	0.19	0.09	0.94	0.00	0.00	0.00	0.02	1.00	0.00	0.00	0.00	0.15	0.30	0.06
t|	0.17	0.13	0.04	0.00	0.55	0.00	0.00	0.00	0.00	0.87	0.09	0.00	0.23	0.09

Increasing the sample of binding sites makes some frequencies become > 0

What if we would further increase the dataset ?

Characterizing binding affinities

- Sampling effects are attenuated by adding **pseudo-counts**
- at each position, a pseudo-count (usually 1) is distributed among the nucleotides
  - equal amount (0.25)
  - OR : fraction proportional to nucleotide prior frequency

a|	8	0	14	0	14	14	0	14	0	13	11
c|	1	0	0	7	0	0	0	0	3	1	0
g|	2	13	0	0	0	0	14	0	0	0	3
t|	3	1	0	7	0	0	0	0	11	0	0

$n_{i,j}$

add 1 using priors
a:t = 0.2 ; c:g = 0.3

a|	8.2	0.2	14.2	0.2	14.2	14.2	0.2	14.2	0.2	13.2	11.2
c|	1.3	0.3	0.3	7.3	0.3	0.3	0.3	0.3	3.3	1.3	0.3
g|	2.3	13.3	0.3	0.3	0.3	0.3	14.3	0.3	0.3	0.3	3.3
t|	3.2	1.2	0.2	7.2	0.2	0.2	0.2	0.2	11.2	0.2	0.2

$n'_{i,j}$

normalize to 1 by dividing by 14+1

a|	0.55	0.02	0.94	0.02	0.94	0.94	0.02	0.94	0.02	0.88	0.75
c|	0.08	0.02	0.02	0.48	0.02	0.02	0.02	0.02	0.22	0.08	0.02
g|	0.15	0.88	0.02	0.02	0.02	0.02	0.95	0.02	0.02	0.02	0.22
t|	0.22	0.08	0.02	0.48	0.02	0.02	0.02	0.02	0.75	0.02	0.02

$f'_{i,j}$

Polls and pseudo-counts

Polls

Résumé des intentions

	Ifop R 17-20 avril	BVA 18-19 avril	CSA 18-19 avril	Harris 18-19 avril	Ipsos 18-19 avril
Nathalie Arthaud	0,5	0	1	0,5	0
Philippe Poutou	1	1,5	1,5	1,5	1,5
Jean-Luc Mélenchon	13,5	14	14,5	12	14
François Hollande	**27**	**30**	**28**	**27,5**	**29**
Eva Joly	3	2	2	3	2
François Bayrou	10,5	10	10,5	11	10
Nicolas Sarkozy	**27**	**26,5**	**25**	**26,5**	**25,5**
Nicolas Dupont-Aignan	1,5	2	1,5	2	1,5
Marine Le Pen	16	14	16	16	16
Jacques Cheminade	0	0	0	0	0,5
	Détails	*Détails*	*Détails*	*Détails*	*Détails*

They should have added pseudo counts !

Final results

Characterizing binding affinities

- why is it so important to get rid of 0 frequencies ?
- compute the probability of a sequence to represent a binding site for Evi-1

a|	0.57	0.00	1.00	0.00	1.00	1.00	0.00	1.00	0.00	0.93	0.79
c|	0.07	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.21	0.07	0.00
g|	0.14	0.93	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.21
t|	0.21	0.07	0.00	0.50	0.00	0.00	0.00	0.00	0.79	0.00	0.00
	C	C	A	C	A	A	G	A	C	A	A

P = 0.07 * 0 * 1 * 0.5 * 1 * 1 * 1 * 1 * 1 * 0.21 * 0.93 * 0.79 = 0

a|	0.55	0.02	0.94	0.02	0.94	0.94	0.02	0.94	0.02	0.88	0.75
c|	0.08	0.02	0.02	0.48	0.02	0.02	0.02	0.02	0.22	0.08	0.02
g|	0.15	0.88	0.02	0.02	0.02	0.02	0.95	0.02	0.02	0.02	0.22
t|	0.22	0.08	0.02	0.48	0.02	0.02	0.02	0.02	0.75	0.02	0.02
	C	C	A	C	A	A	G	A	C	A	A

P = 0.08* 0.02* 0.94*0.48* 0.95* 0.95* 0.95*0.95*0.22 *0.88*0.75 = 8.6e-5

This sequence can NEVER be a BS → this sequence has non-0 probability of being a BS

Comparing two models

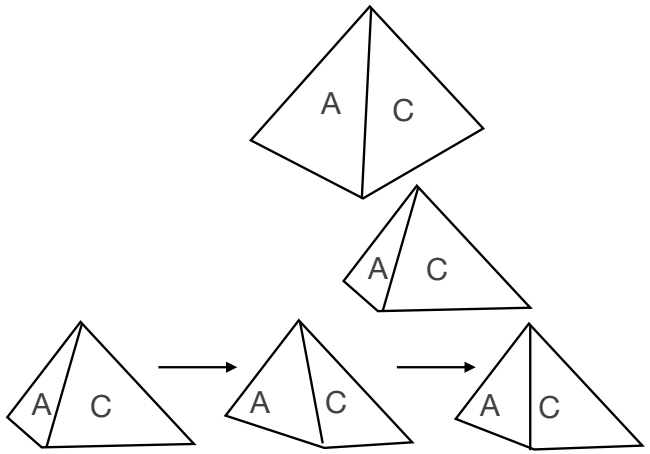
- We want to use the PFM model to find potential binding sites in a DNA sequence
- How well does the model discriminate between
  - True binding sites ("**Binding site model**")
  - Random background sequences ("**Background model**")
- Same as identikit procedure: how well does an identikit distinguish between
  - the real criminal searched for
  - a random person picked in the crowd

Not really: many average features

Probably more discriminative: many specific features!

Generating background sequences

- We often need to bioinformatically **generate sequences according a a specific model** *("rolling a dice")*
- this model gives the emission probabilities, i.e. the probability *P(X,j)* at each position *j* of the sequence to obtain one of the four bases A,C,G,T
- possible models
  - equiprobable model: $P(A)=P(C)=P(G)=P(T)=0.25$
  - more realistic model: $P(A)=P(T)=p_{A,T}$; $P(C)=P(G)=p_{C,G}$
  - position dependent model: $P(A,j) = f_{A,j}$



Information content of a matrix

Would I get the same PFM if I would consider 14 random sequences?

- Generate 14 sequences of length L=11 according to the emission probabilities of the PFM; probability of obtaining exactly the same count matrix:

$$P_{mat} = \prod_{j=1}^{L} \prod_{i \in A,C,G,T} C_{i,j} f_{i,j}'^{n'_{i,j}}$$

- Generate 14 **random sequences** on length L=11; probability of obtaining exactly the same count matrix:

$$P_{rand} = \prod_{j=1}^{L} \prod_{i \in A,C,G,T} C_{i,j}\ p_i^{n'_{i,j}}$$

$p_A = p_T = 0.2 \quad p_C = p_G = 0.3$

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

$n'_{i,j}$

a|	8.2	0.2	14.2	0.2	14.2	14.2	0.2	14.2	0.2	13.2	11.2
c|	1.3	0.3	0.3	7.3	0.3	0.3	0.3	0.3	3.3	1.3	0.3
g|	2.3	13.3	0.3	0.3	0.3	0.3	14.3	0.3	0.3	0.3	3.3
t|	3.2	1.2	0.2	7.2	0.2	0.2	0.2	0.2	11.2	0.2	0.2

p_i

a|	0.2
c|	0.3
g|	0.3
t|	0.2

$f'_{i,j}$

a|	0.55	0.02	0.94	0.02	0.94	0.94	0.02	0.94	0.02	0.88	0.75
c|	0.08	0.02	0.02	0.48	0.02	0.02	0.02	0.02	0.22	0.08	0.02
g|	0.15	0.88	0.02	0.02	0.02	0.02	0.95	0.02	0.02	0.02	0.22
t|	0.22	0.08	0.02	0.48	0.02	0.02	0.02	0.02	0.75	0.02	0.02

Information content of a matrix

- Log-likelihood ratio

$$LLR = \log_2\left(\frac{P_{mat}}{P_{rand}}\right) = \sum_{j=1}^{L}\sum_{i\in A,C,G,T} n'_{i,j}\log_2\frac{f'_{i,j}}{p_i}$$

- Information content (bits)

$$IC = \frac{1}{n}\log_2\left(\frac{P_{mat}}{P_{rand}}\right) = \sum_{j=1}^{L}\sum_{i\in A,C,G,T} f'_{i,j}\log_2\frac{f'_{i,j}}{p_i}$$

- Properties
  - if $f'_{ij} \rightarrow p_i$: IC → 0 : **zero information content** if the frequencies in the alignment are equal to the background frequencies ("random matrix")
  - IC is **maximal** if the least abundant nucleotide (smallest p_i) is the most represented in the matrix (largest f'_{ij})
  - information content is related to the **Shannon entropy**, mesuring the uncertainty : high information content → low Shannon entropy

[Hertz, Stormo 1999]

Matrix logos

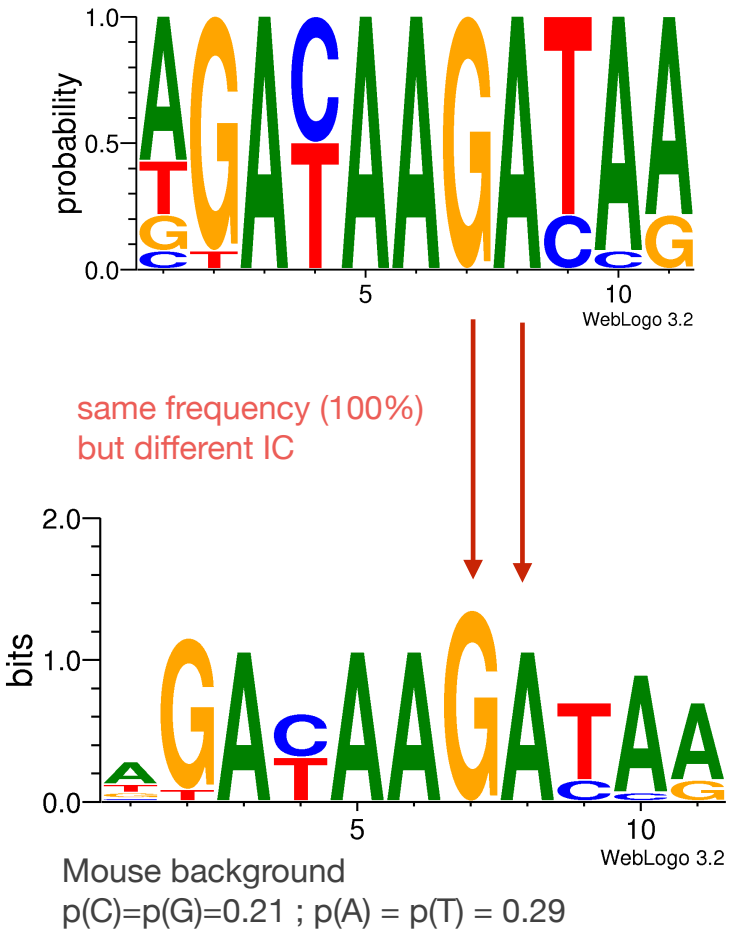
- Information content of the matrix:

$$IC = \sum_{j=1}^{L} \sum_{i \in A,C,G,T} f'_{i,j} \log_2 \frac{f'_{i,j}}{p_i}$$

- Information content of a column:

$$IC^j = \sum_{i \in A,C,G,T} f'_{i,j} \log_2 \frac{f'_{i,j}}{p_i}$$

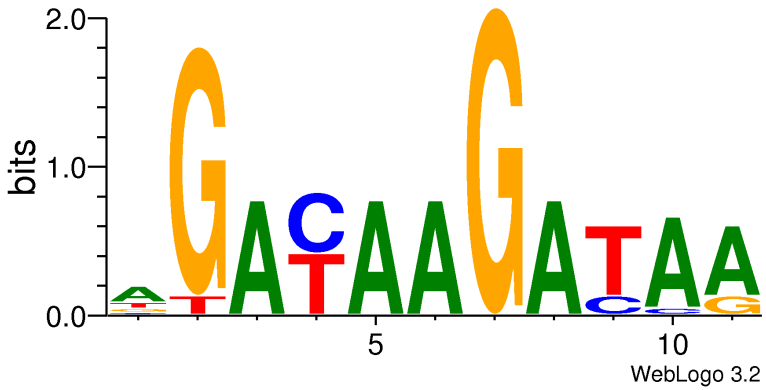
- Conventions:
  - height of column represents IC
  - relative sizes proportional to frequencies



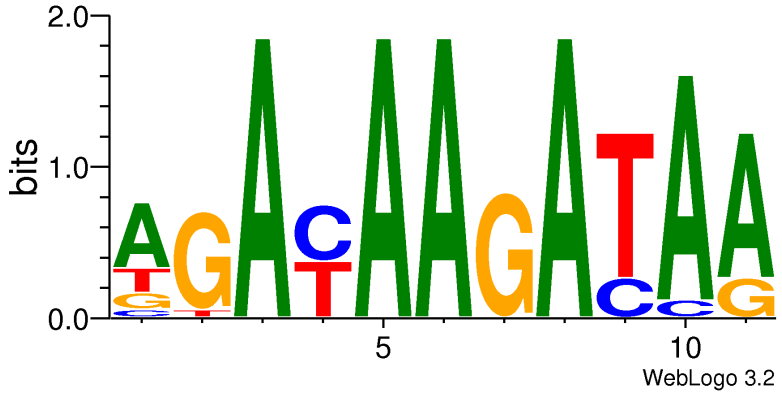
Mouse background
p(C)=p(G)=0.21 ; p(A) = p(T) = 0.29

Matrix logos

a	8	0	14	0	14	14	0	14	0	13	11
c	1	0	0	7	0	0	0	0	3	1	0
g	2	13	0	0	0	0	14	0	0	0	3
t	3	1	0	7	0	0	0	0	11	0	0



P. falciparum background
p(C)=p(G)=0.1 ; p(A) = p(T) = 0.4



Anaeromyxobacter background
p(C)=p(G)=0.37 ; p(A) = p(T) = 0.13

From frequencies (PFM) to weights (PWM)

- position frequency matrices do not contain information about the background distribution
- define **position-weight matrices (PWM)** as
  (or Position Specific Scoring Matrix PSSM)

$$PWM_{i,j} = \ln \frac{f'_{ij}}{p_i}$$

a|	8	0	14	0	14	14	0	14	0	13	11
c|	1	0	0	7	0	0	0	0	3	1	0
g|	2	13	0	0	0	0	14	0	0	0	3
t|	3	1	0	7	0	0	0	0	11	0	0

Count matrix

a|	0.55	0.02	0.94	0.02	0.95	0.95	0.02	0.95	0.02	0.88	0.75
c|	0.08	0.02	0.02	0.48	0.02	0.02	0.02	0.02	0.22	0.08	0.02
g|	0.15	0.88	0.02	0.02	0.02	0.02	0.95	0.02	0.02	0.02	0.22
t|	0.22	0.08	0.02	0.48	0.02	0.02	0.02	0.02	0.75	0.02	0.02

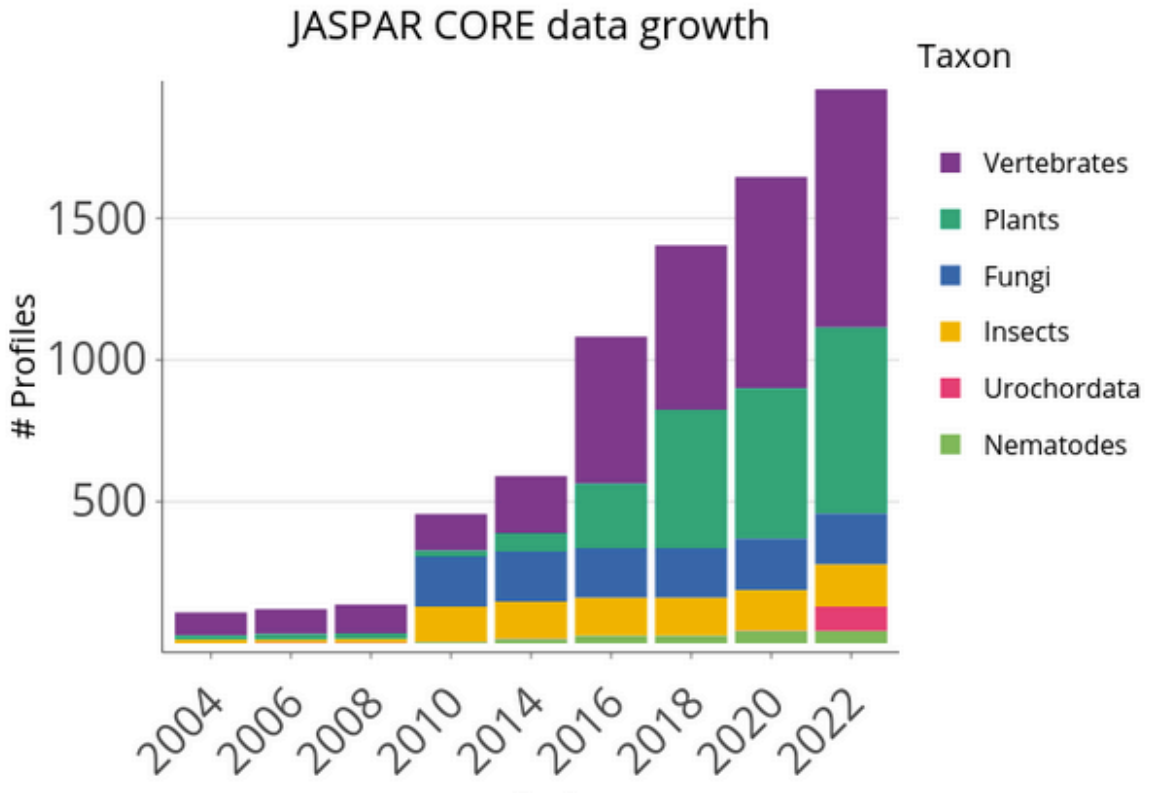
PFM

a|	0.75	-2.71	1.30	-2.71	1.30	1.30	-2.71	1.30	-2.71	1.22	1.06
c|	-1.04	-2.71	-2.71	0.73	-2.71	-2.71	-2.71	-2.71	-0.07	-1.04	-2.71
g|	-0.46	1.31	-2.71	-2.71	-2.71	-2.71	1.39	-2.71	-2.71	-2.71	-0.10
t|	-0.22	-1.16	-2.71	0.58	-2.71	-2.71	-2.71	-2.71	1.02	-2.71	-2.71

PWM

Motif databases

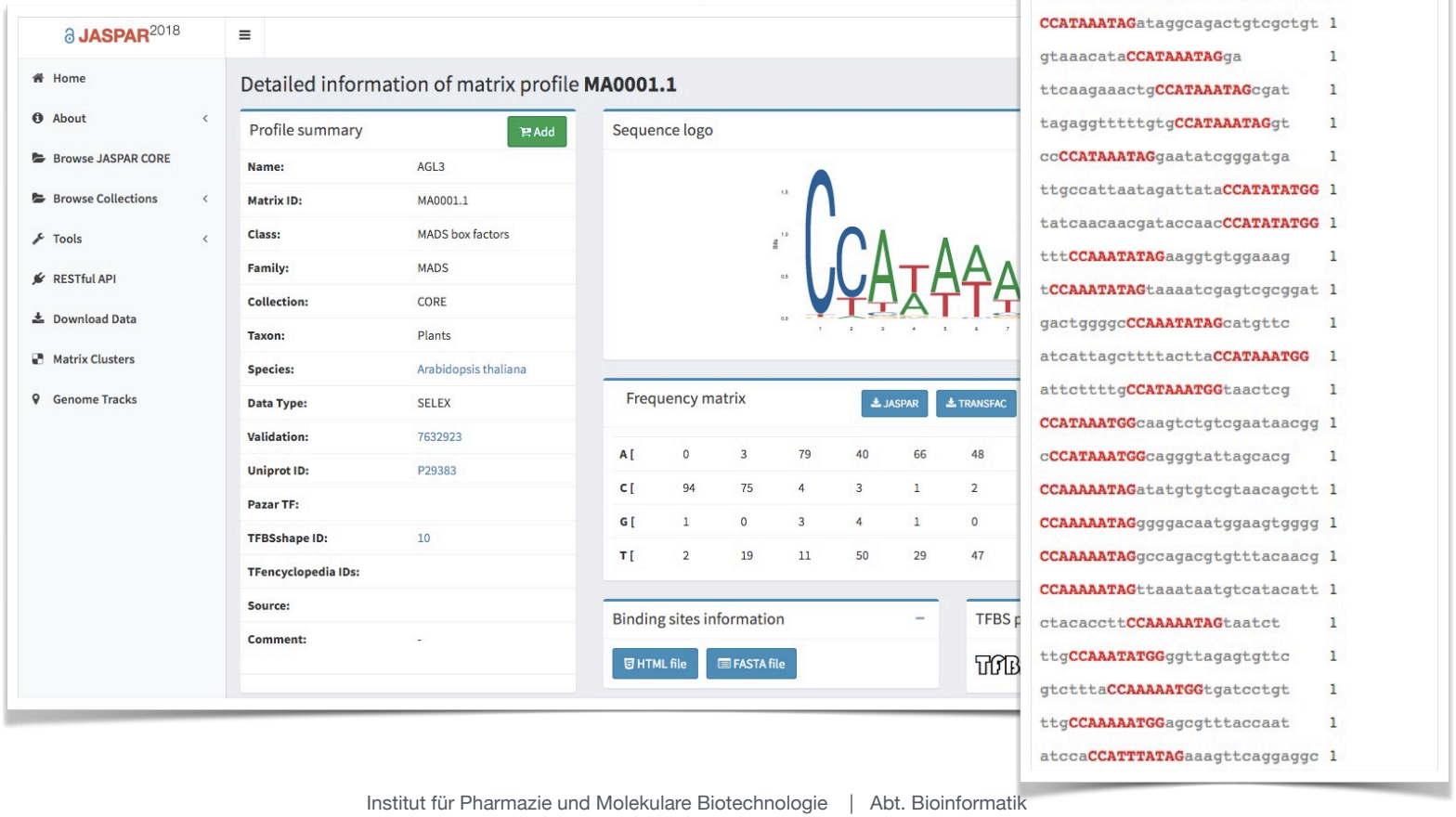
- JASPAR is an open database of binding profiles



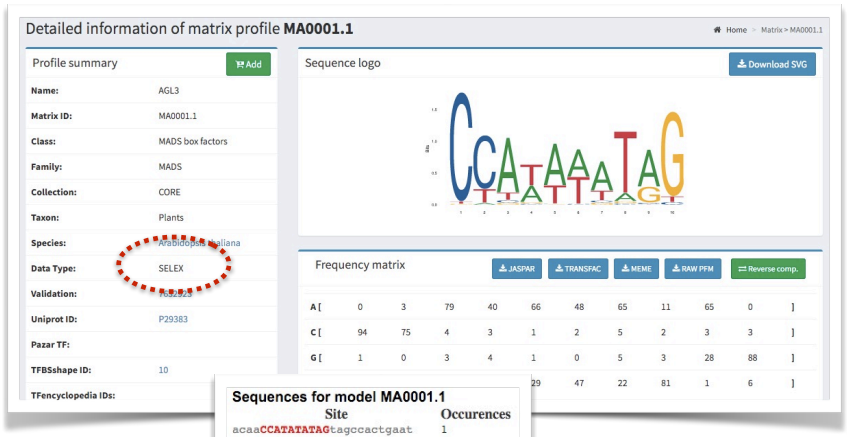
http://jaspar.genereg.net/

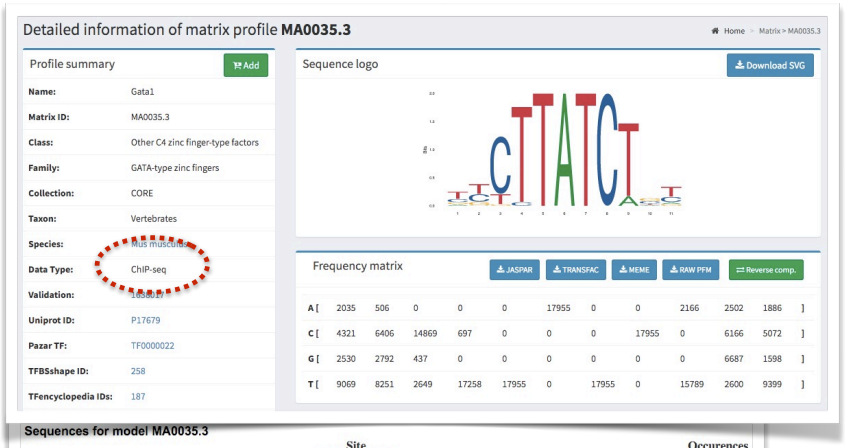
Motif databases

- JASPAR is an open database of binding profiles



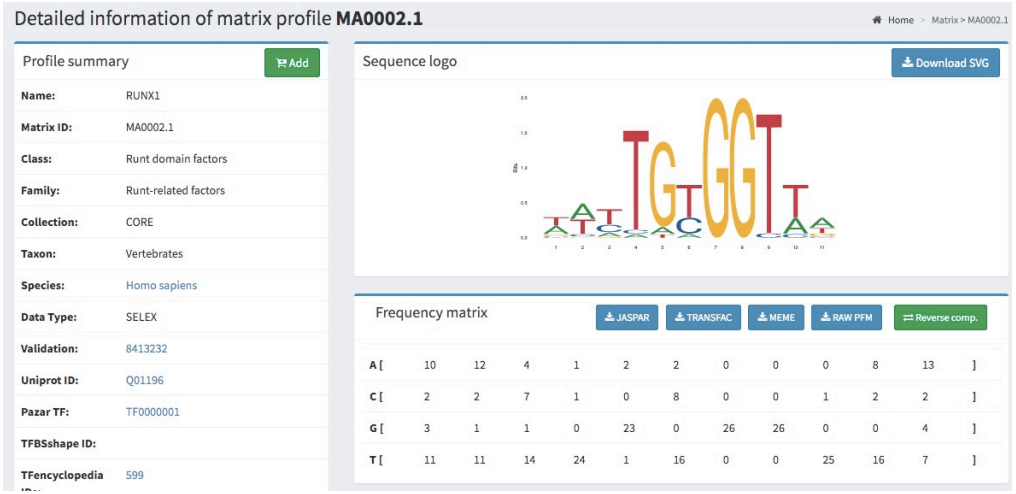
Different sources



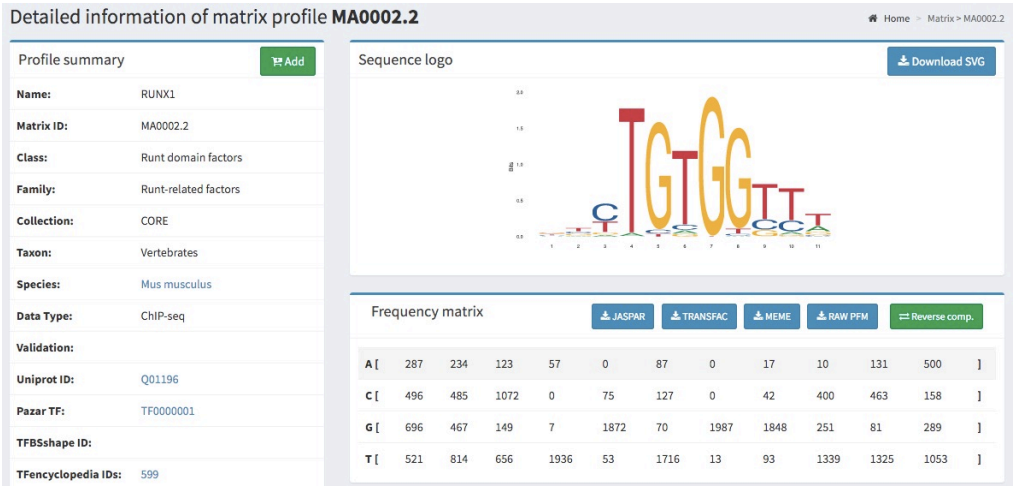


ChIP-seq: real binding site is hidden in much longer sequence
→ lower resolution

Redundant profiles: RUNX1



SELEX
26 binding sites



ChIP-seq
2000 binding sites