V9 – 7. Protein-DNA contacts

- Transcription factors (TFs)
- Transcription factor binding sites (TFBS)
- Experimental detection of TFBS
- Position-specific scoring matrices (PSSMs)
- Binding free energy models
- Cis-regulatory motifs

Tue, May 15, 2018

7. DNA-binding proteins

DNA binding proteins include:

- TFs that activate or repress gene expression

- Enzymes involved in DNA repair

- Enzymes that place chemical (epigenetic) modifications on DNA

- Proteins that pack or unpack the chromatin structure

- Proteins that help to unzip double-stranded DNA

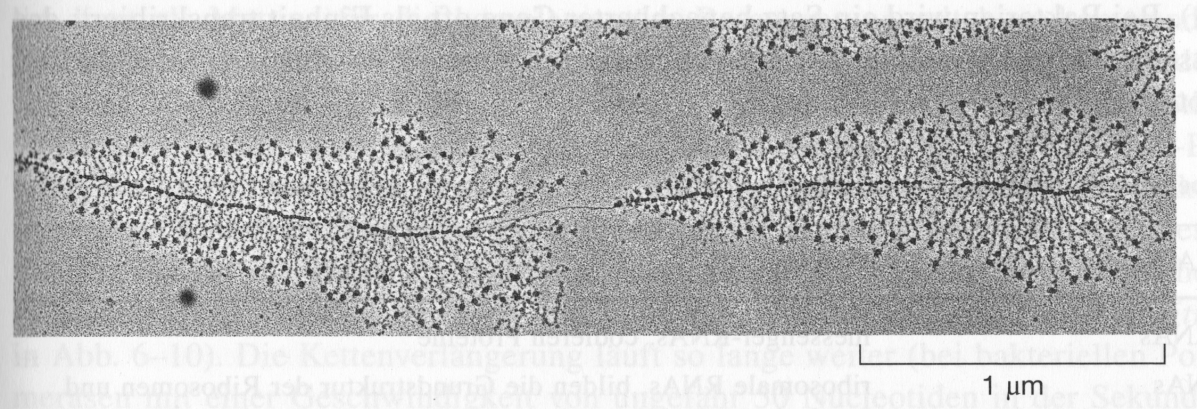
- DNA topoisomerases that are involved in DNA supercoiling etc.

From this long list, we will discuss here only TFs.

Transcription Initiation

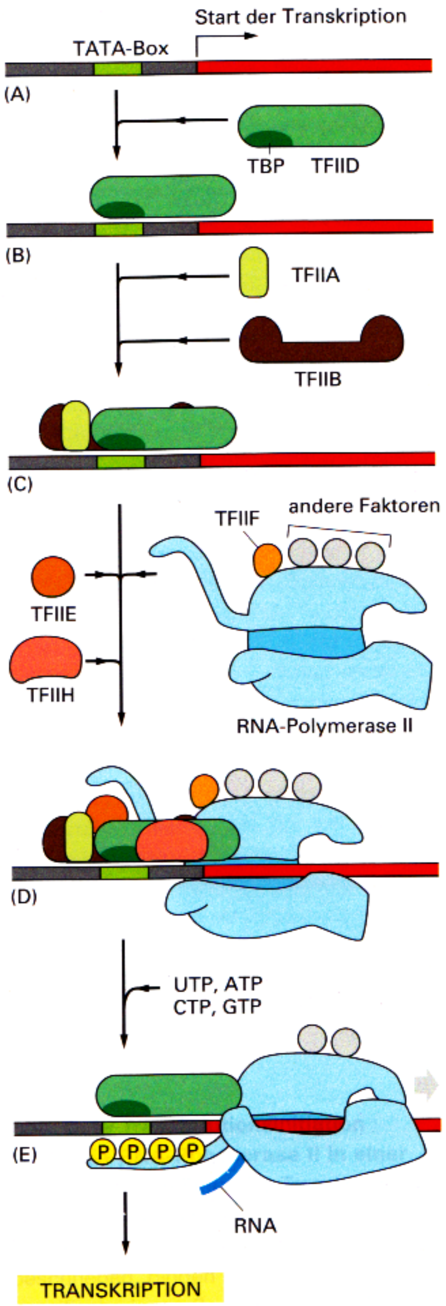
In eukaryotes:

- several **general** transcription factors **have** to bind to gene promoter
- **specific** enhancers or repressors **may** bind
- then the RNA polymerase binds
- and starts transcription



Shown here: many RNA polymerases read central DNA at different positions and produce ribosomal rRNAs (perpendicular arms). The large particles at their ends are likely ribosomes being assembled.

Alberts et al.
"Molekularbiologie der Zelle", 4.Aufl.



7. Binding forces

There is generally **electrostatic attraction** between the negatively charged phosphate groups of the DNA backbone and positively charged amino acids on the protein surface.

This interaction involves only the DNA-backbone and is thus mostly independent from the DNA sequence.

Attractive contribution:

specific polar and non-polar **interactions** between the nucleotide bases of particular DNA sequence motifs and their protein binding partner.

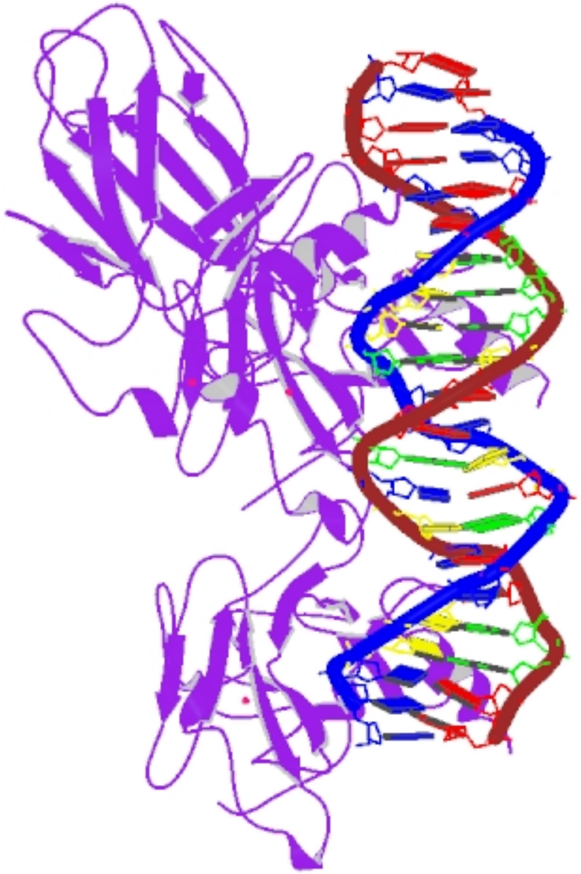
p53: example of a Protein-DNA-complex

PDB-structure 1TUP: tumor suppressor **p53**

Determined by X-ray crystallography

Purple (left): p53-protein (multiple copies)

Blue/red DNA double strand (right)



The protective action of the wild-type *p53* gene helps to suppress tumors in humans. The *p53* gene is the most commonly mutated gene in human cancer, and these mutations may actively promote tumor growth.

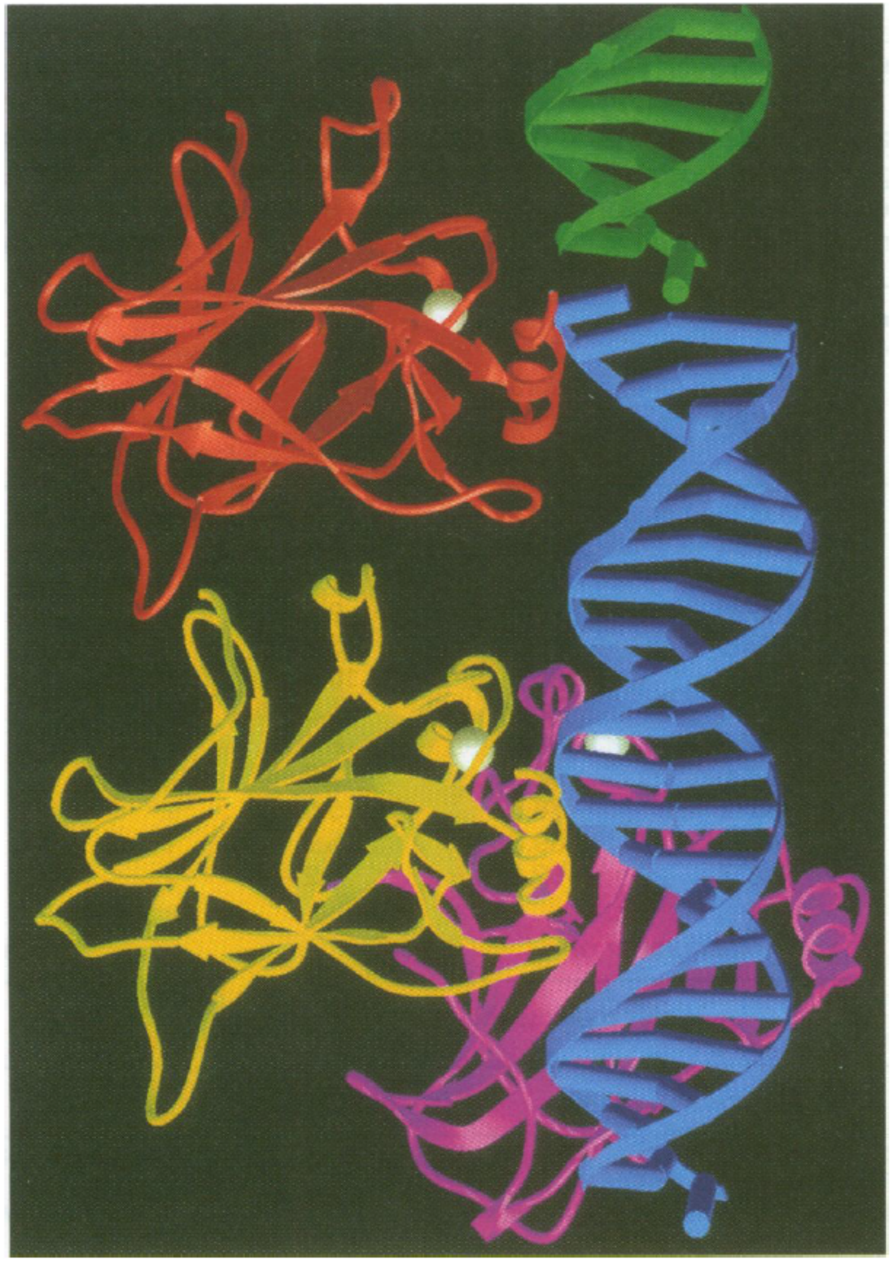
www.sciencemag.org (1993)

www.rcsb.org

Contacts establish specific binding mode

Nikola Pavletich,
Sloan Kettering
Cancer Center

Fig. 3. Schematic ribbon drawing of the asymmetric unit, which contains three p53 core domain molecules and one DNA duplex. Two of the core domains bind DNA (blue); one (yellow) interacts extensively with a consensus binding site, and the other (red) binds at a nonconsensus site at the interface of DNA fragments related by crystallographic symmetry (a portion of the symmetry-related DNA fragment is shown in green). The third core domain molecule (purple) does not bind DNA, but makes protein-protein contacts stabilizing crystal packing. The zinc atoms are shown as white spheres.



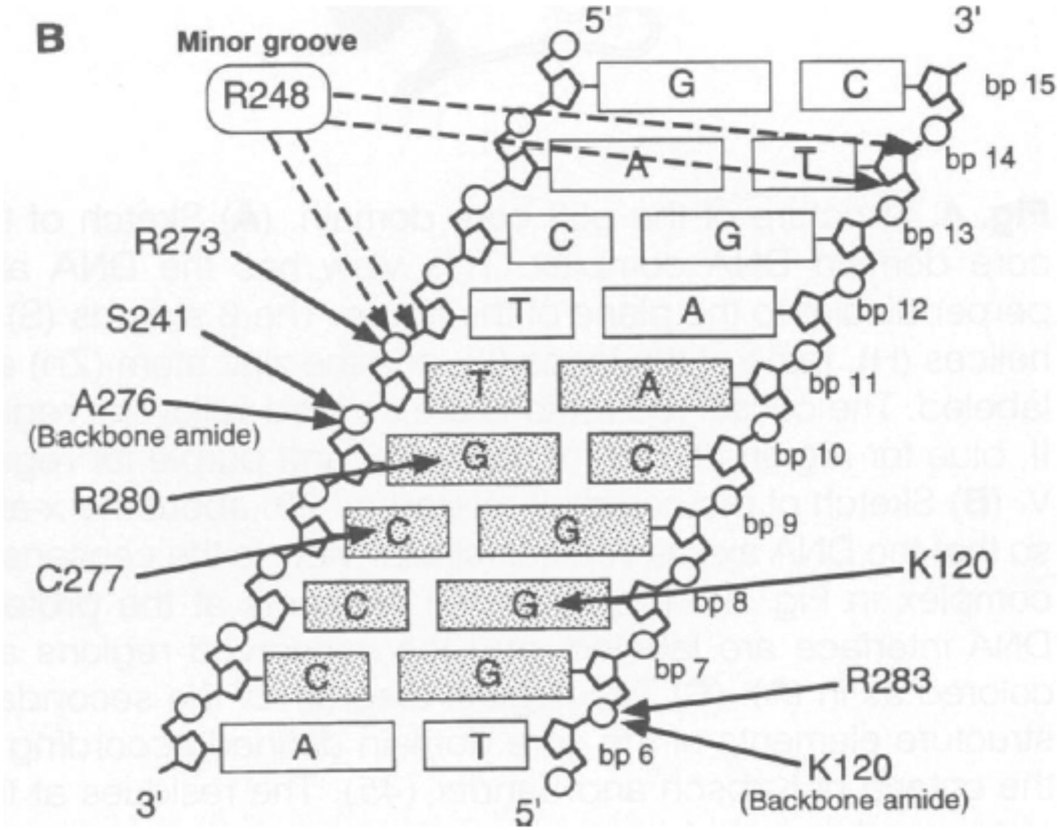
RESEARCH ARTICLE

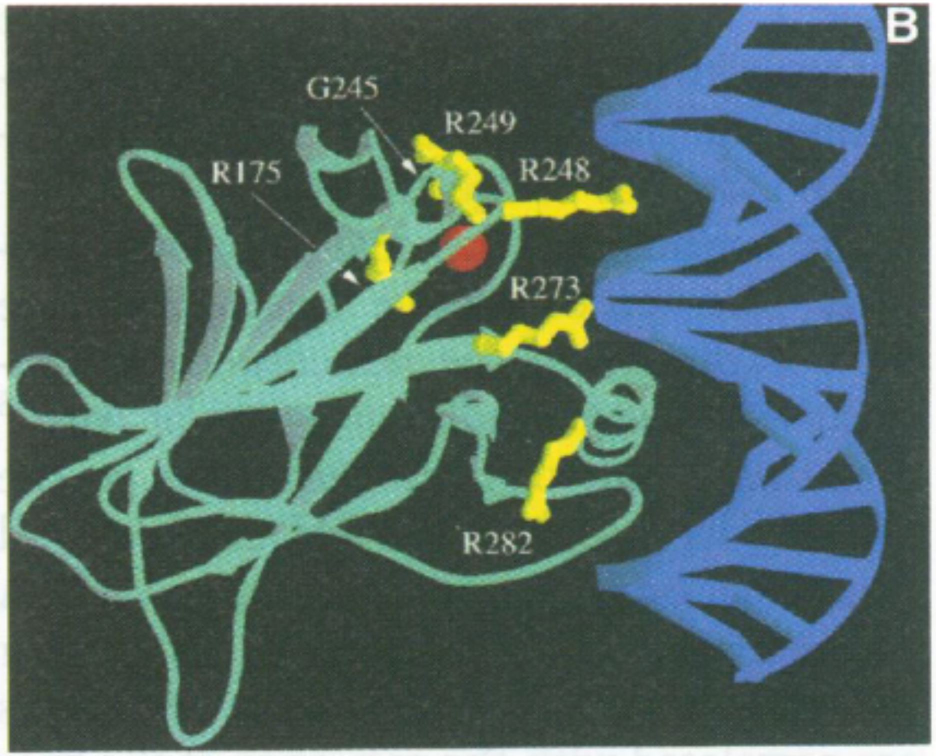
Crystal Structure of a p53 Tumor Suppressor–DNA Complex: Understanding Tumorigenic Mutations

Yunje Cho, Svetlana Gorina, Philip D. Jeffrey, Nikola P. Pavletich

Science 265, 346-355 (1994)

Contact residues





Left: Protein – DNA contacts involve many arginine (R) and lysine (K) residues

Right: the 6 most frequently mutated amino acids (yellow) in cancer.
5 of them are Arginines.

In p53 all 6 residues are located at the binding interface for DNA!

Science 265, 346-355 (1994)

Structural view at *E. coli* TFs

Determine homology between the domains and protein families of TFs and regulated genes and proteins of known 3D structure.

→ Determine uncharacterized *E.coli* proteins with DNA-binding domains (DBD)

→ identify large majority of *E.coli* TFs.

Sarah Teichmann
EBI

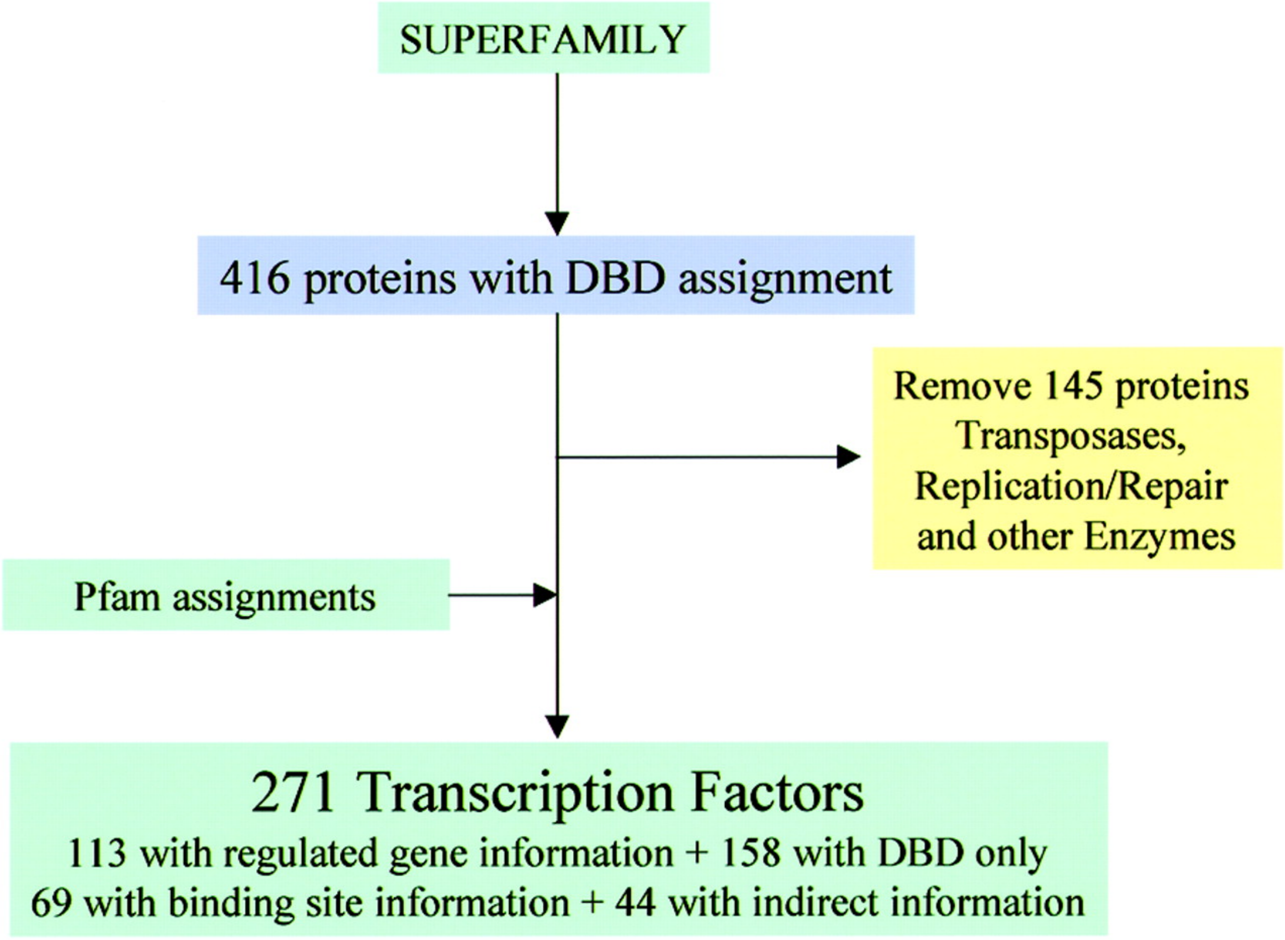
Madan Babu,
MRC

Babu, Teichmann, Nucl. Acid Res. 31, 1234 (2003)

Flow chart of method to identify TFs in E.coli

SUPERFAMILY database (C. Chothia) contains a library of HMM models based on the sequences of proteins in SCOP for predicted proteins of completely sequenced genomes.

Remove all DNA-binding proteins involved in replication/repair etc.


SUPERFAMILY
416 proteins with DBD assignment
Remove 145 proteins Transposases, Replication/Repair and other Enzymes
Pfam assignments
271 Transcription Factors
113 with regulated gene information + 158 with DBD only
69 with binding site information + 44 with indirect information

Babu, Teichmann, Nucl. Acid Res. 31, 1234 (2003)

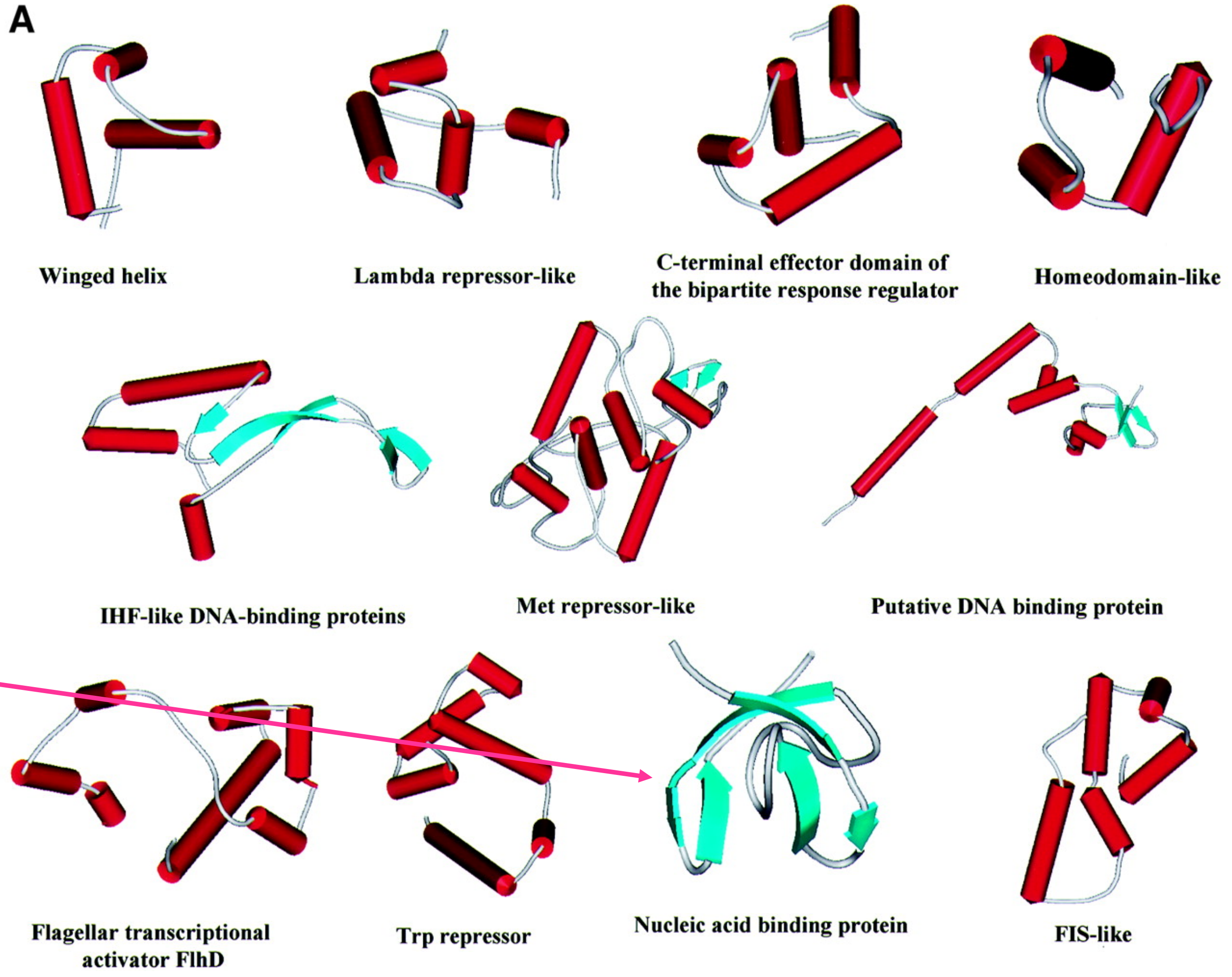
3D structures of putative (and real) TFs in *E.coli*

3D structures of the 11 DBD families seen in the 271 identified TFs in *E.coli*.

The **helix–turn–helix motif** is typical for DNA-binding proteins.

It occurs in all families except the nucleic acid binding family.

Still the scaffolds in which the motif occurs are very different.


A
Winged helix
Lambda repressor-like
C-terminal effector domain of the bipartite response regulator
Homeodomain-like
IHF-like DNA-binding proteins
Met repressor-like
Putative DNA binding protein
Flagellar transcriptional activator FlhD
Trp repressor
Nucleic acid binding protein
FIS-like

Babu, Teichmann, Nucl. Acid Res. 31, 1234 (2003)

Domain architectures of TFs

The 74 unique domain architectures of the 271 TFs.

The **DBDs** are represented as rectangles. ▭

The partner domains are represented as hexagons (**small molecule-binding domain**), ⬡

triangles (**enzyme** domains), △

circles (protein interaction domain), ○

diamonds (domains of unknown function). ◇

The receiver domain has a pentagonal shape.

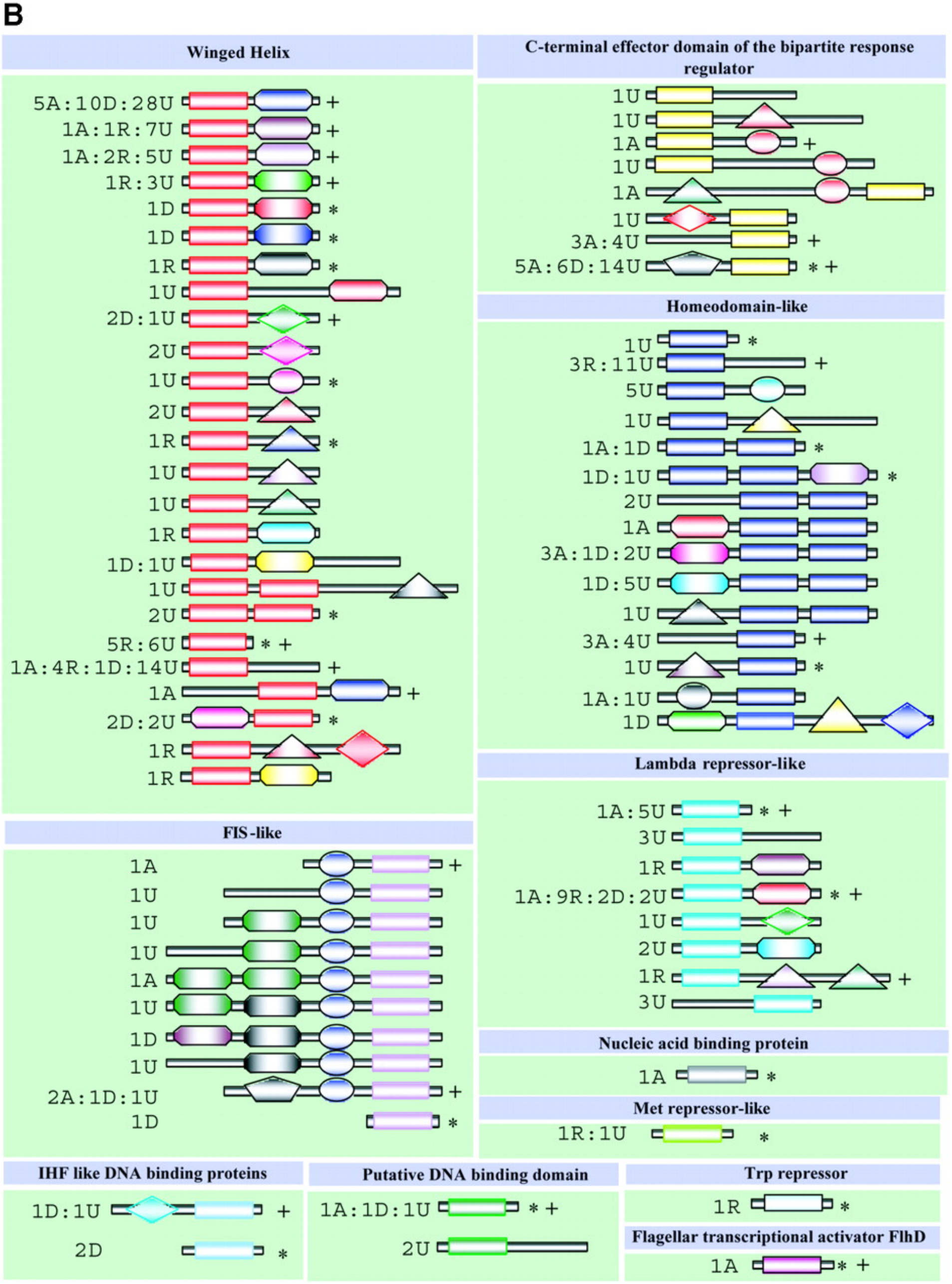
A, R, D and U stand for activators, repressors, dual regulators and TFs of unknown function.

The number of TFs of each type is given next to each domain architecture.

Architectures of known 3D structure are denoted by asterisks.

'+' are cases where the regulatory function of a TF has been inferred by indirect methods, so that the DNA-binding site is not known.

Babu, Teichmann, Nucl. Acid Res. 31, 1234 (2003)



Evolution of TFs

10%	1-domain proteins
75%	2-domain proteins
12%	3-domain proteins
3%	4-domain proteins

TFs have evolved by apparently extensive recombination of domains.

Proteins with the same sequential arrangement of domains
are likely to be direct duplicates of each other.

74 distinct domain architectures have duplicated to give rise to 271 TFs.

Babu, Teichmann, Nucl. Acid Res. 31, 1234 (2003)

Evolution of the gene regulatory network

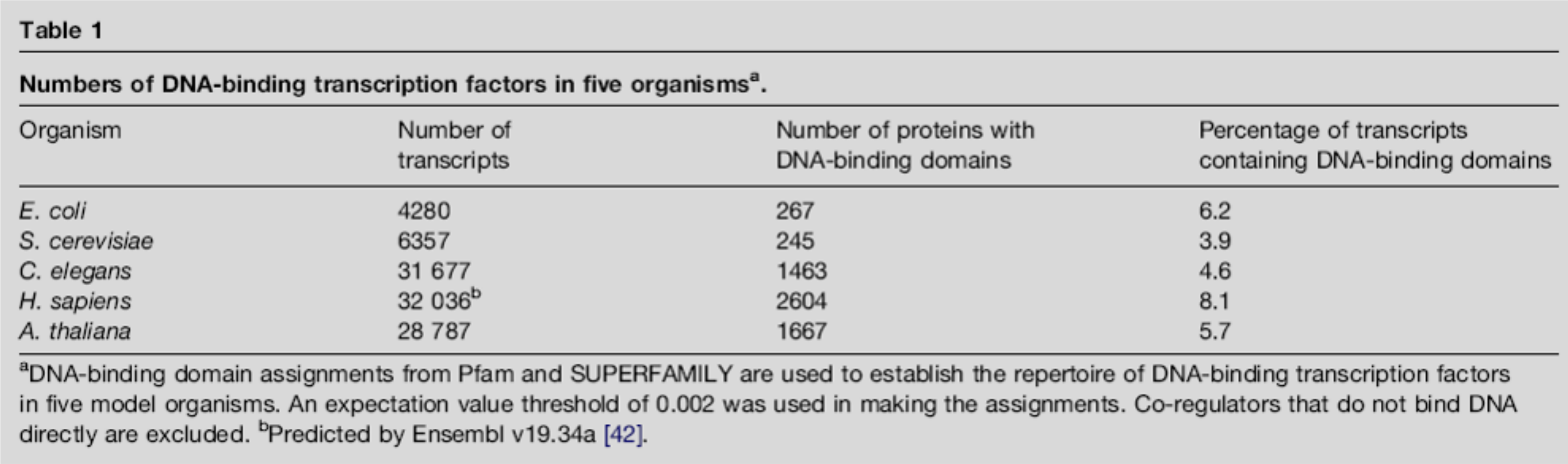


Table 1

Numbers of DNA-binding transcription factors in five organisms[a].

Organism	Number of transcripts	Number of proteins with DNA-binding domains	Percentage of transcripts containing DNA-binding domains
E. coli	4280	267	6.2
S. cerevisiae	6357	245	3.9
C. elegans	31 677	1463	4.6
H. sapiens	32 036[b]	2604	8.1
A. thaliana	28 787	1667	5.7

[a] DNA-binding domain assignments from Pfam and SUPERFAMILY are used to establish the repertoire of DNA-binding transcription factors in five model organisms. An expectation value threshold of 0.002 was used in making the assignments. Co-regulators that do not bind DNA directly are excluded. [b] Predicted by Ensembl v19.34a [42].

Most genomes contain hundreds up to a few thousands of TFs.

Larger genomes tend to have more TFs per gene.

Babu et al. Curr Opin Struct Biol. 14, 283 (2004)

Transcription factors in yeast *S. cereviseae*

Q: How can one define transcription factors?

Hughes & de Boer consider as TFs proteins that

(a) bind DNA directly and in a sequence-specific manner and

(b) function to regulate transcription nearby sequences they bind

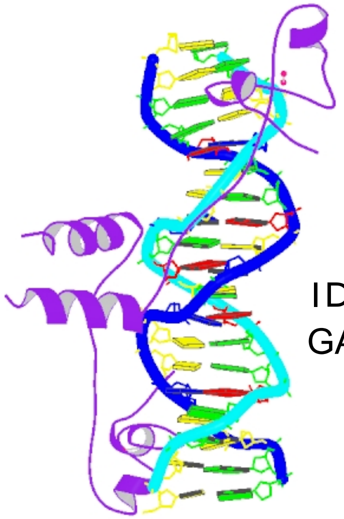
Q: Is this a good definition?

Yes. Only 8 of 545 human proteins that bind specific DNA sequences and regulate transcription lack a known DNA-binding domain (DBD).

Hughes, de Boer (2013) Genetics 195, 9-36

Transcription factors in yeast

Hughes and de Boer list 209 known and putative yeast TFs.
The vast majority of them contains a canonical DNA-binding domain.



1D66.pdb
GAL4 family

Most abundant:

- GAL4/zinc cluster domain (57 proteins),
  largely specific to fungi (e.g. yeast)

- zinc finger C2H2 domain (41 proteins),
  most common among all eukaryotes.

Other classes :

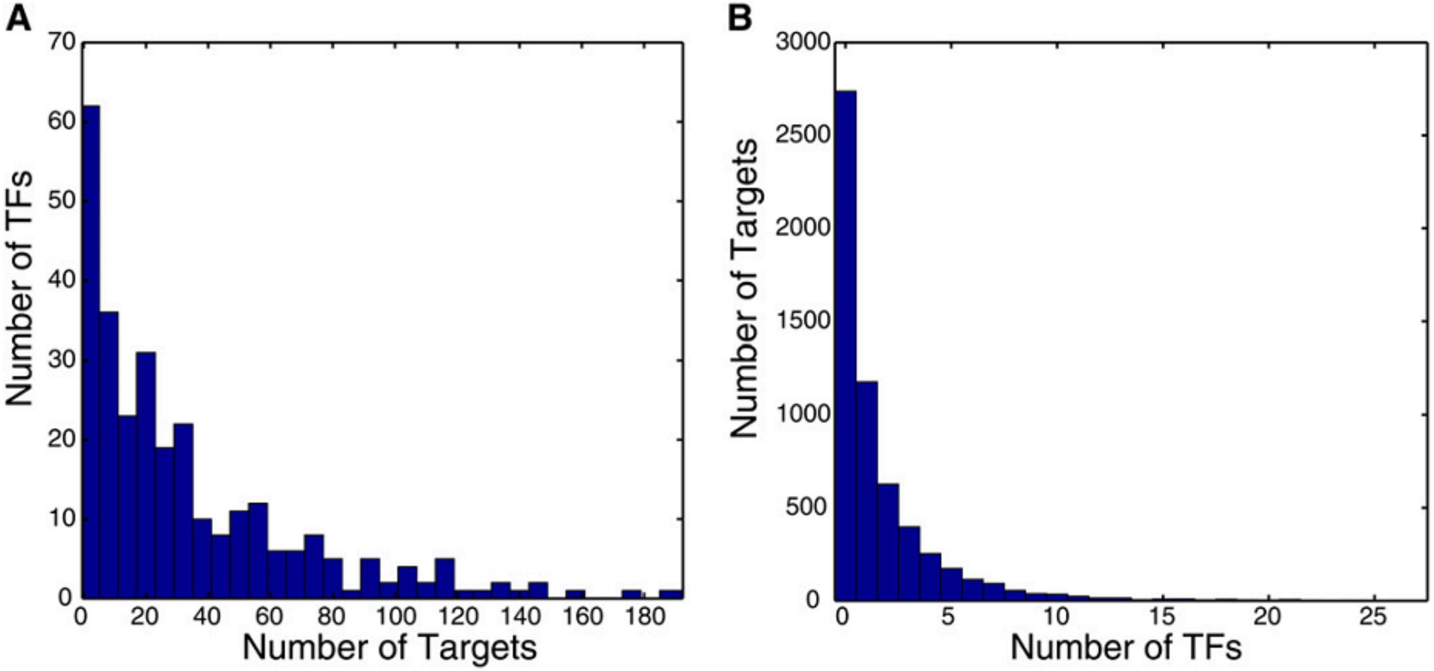
- bZIP (15),
- Homeodomain (12),
- GATA (10), and
- basic helix-loop-helix (bHLH) (8).

Hughes, de Boer (2013) Genetics 195, 9-36

TFs of *S. cereviseae*

(A) Most TFs tend to bind relatively few targets.
57 out of 155 unique proteins bind to ≤ 5 promoters in at least one condition.
17 did not significantly bind to any promoters under any condition tested.

In contrast, several TFs have hundreds of promoter targets.
These TFs include the general regulatory factors (GRFs), which play a global role in transcription under diverse conditions.

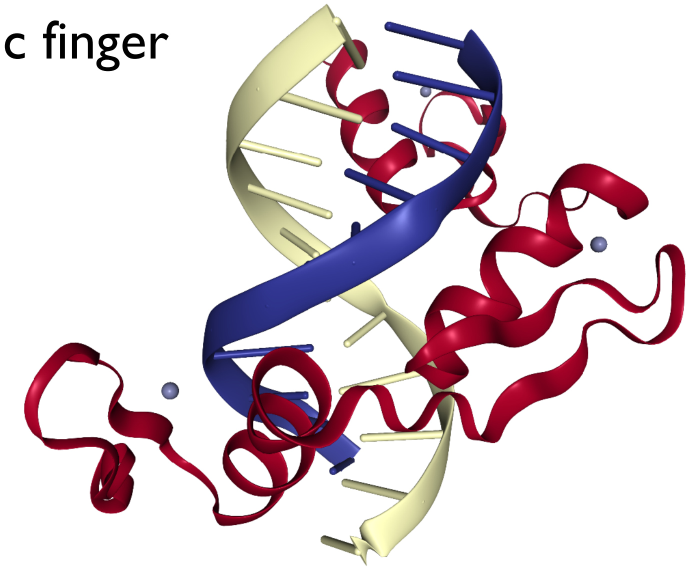


(B) # of TFs that bind to one promoter.

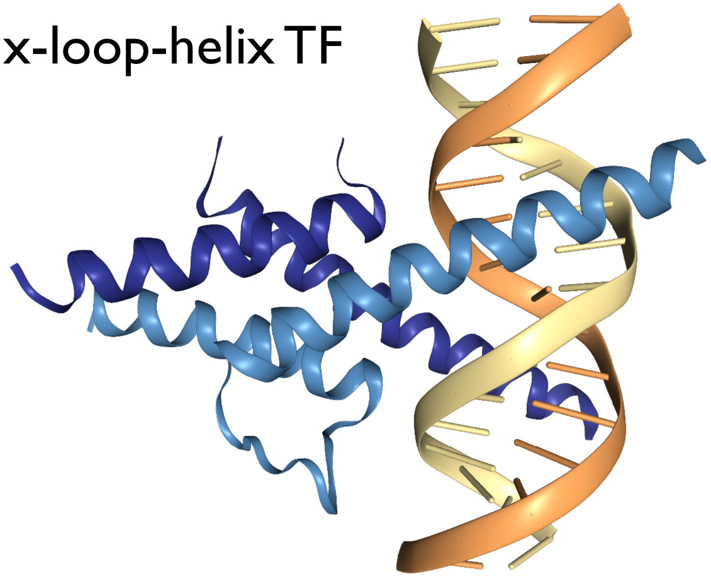
Hughes, de Boer (2013) Genetics 195, 9-36

7.1 Structural types of TFs

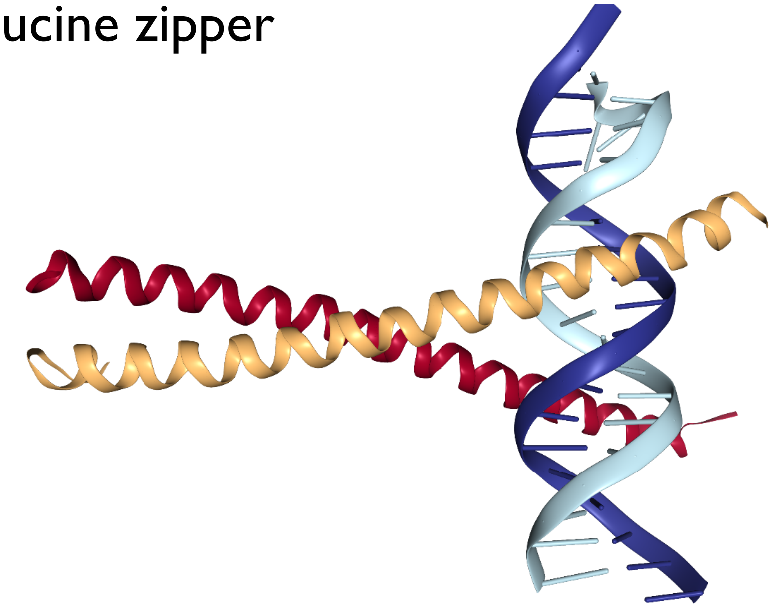
Zinc finger



Helix-loop-helix TF



Leucine zipper



High mobility group TF

7.2 Transcription factor binding sites (TFBSs)

TFBS: DNA region that forms a **specific physical contact** with a particular TF.

TFBS are usually between 8 and 20 bp long
and contain a 5-8 bp long **core region** of well-conserved nucleotide bases.

Most TFs bind in the **major groove** of double-stranded DNA,
the others bind in the minor groove.

The **periodicity** of double-standed DNA is around 10 bp.
Thus, the core regions of TFBS are a bit longer than half a turn of dsDNA.

TFs may recognize DNA sequences that are similar,
but not identical, differing by a few nucleotides.

Sequence logos represent binding motifs

A **logo** represents each column of the alignment by a stack of letters.

The height of each letter is proportional to the **observed frequency** of the corresponding amino acid or nucleotide.

The overall height of each stack is proportional to the **sequence conservation** at that position.

Sequence conservation is defined as difference between the maximum possible entropy and the entropy of the observed symbol distribution:

$$R_{seq} = S_{max} - S_{obs} = \log_2 N - \left(- \sum_{n=1}^{N} p_n \log_2 p_n \right)$$

p_n : observed frequency of symbol n at a particular sequence position

N : number of distinct symbols for the given sequence type, either 4 for DNA/RNA or 20 for protein.

Crooks et al., Genome Research 14:1188–1190 (2004)

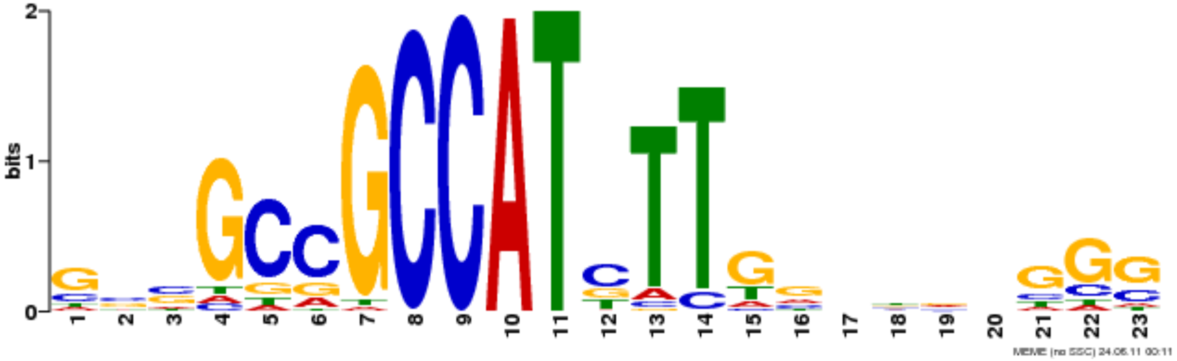
YY1 sequence logo

Sequence-logos are a convenient way to visualize the degree of degeneracy in the TFBS.

$$H_i = -\sum_{b=A}^{T} f_{b,i} \times log_2 f_{b,i}$$

$$R_i = log_2(4) - (H_i + e_n)$$

$$e_n = \frac{1}{ln2} \times \frac{s-1}{2n}$$



Sequence logo for the DNA binding motif that the TF YY1 (Yin Yang 1) binds to.

The motif was derived from the top 500 TF ChIP-seq peaks by the ENCODE consortium. For YY1, 468 out of 500 sequences contained this motif.

Figure from Factorbook repository (Wang *et al.* 2013).

H_i : uncertainty (Shannon entropy) of position *i*

R_i : information content (y-axis) of position *i*

e_n : small-sample correction, *s* = 4 for nucleotides, *n* : number of sequences

YY1 binding motifs

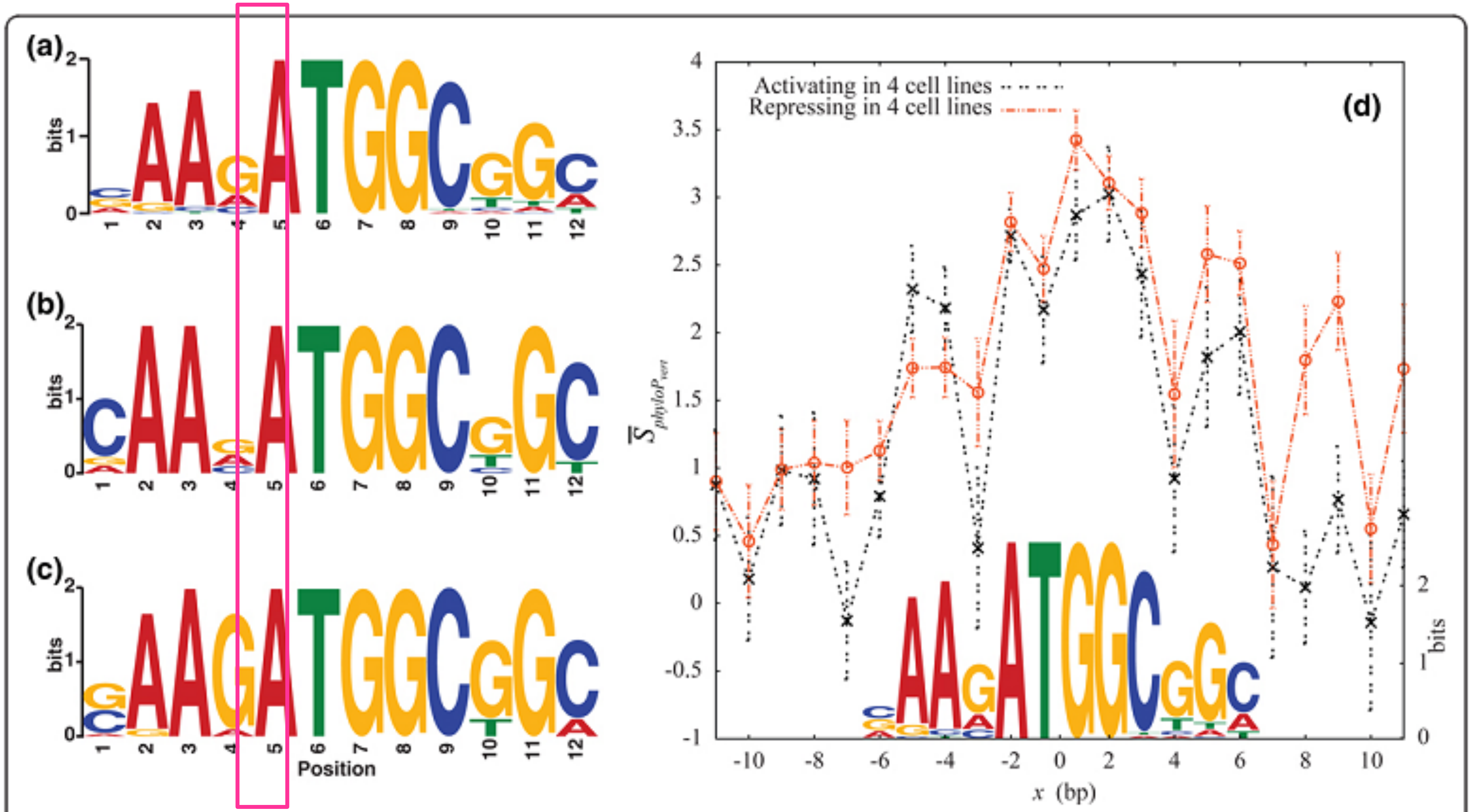
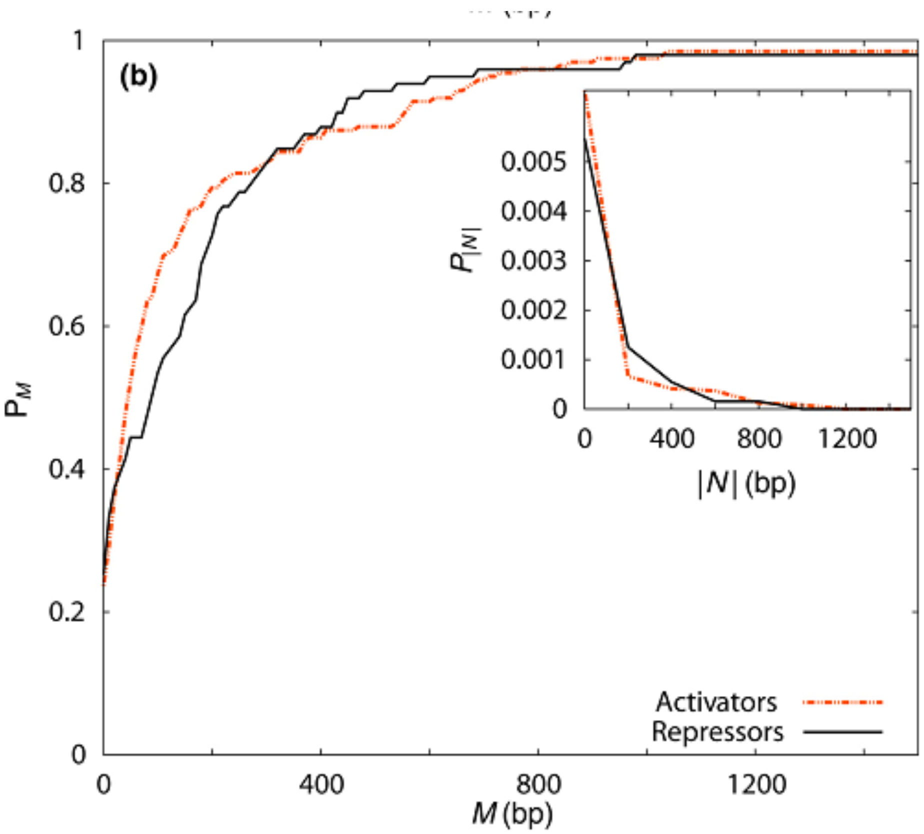


Figure 2 Characterization of functional YY1 binding sites. Sequence logo [102] for YY1 binding sites from **(a)** PWM and sites that are functionally **(b)** ubiquitously activating (9 BS) or **(c)** ubiquitously repressive (16 BS) in four human cell lines. In **(d)**, we plot the mean vertebrate phyloP conservation score [90] around functional YY1 binding sites. The mean score, $\bar{S}_{phyloP_{vert}}$, was computed at each base for sites where the binding event ubiquitously activated (black line) or repressed (red line) transcription in all four cell lines. The position weight matrix that was used to predict YY1 binding sites is shown (scale on the right axis).

No noticeable difference in binding motifs of activated or repressed target genes.

Whitfield et al. Genome Biology 2012, 13:R50

Where are TFBS relative to the TSS?



Inset: probability to find binding site at position N from transcriptional start site (TSS)

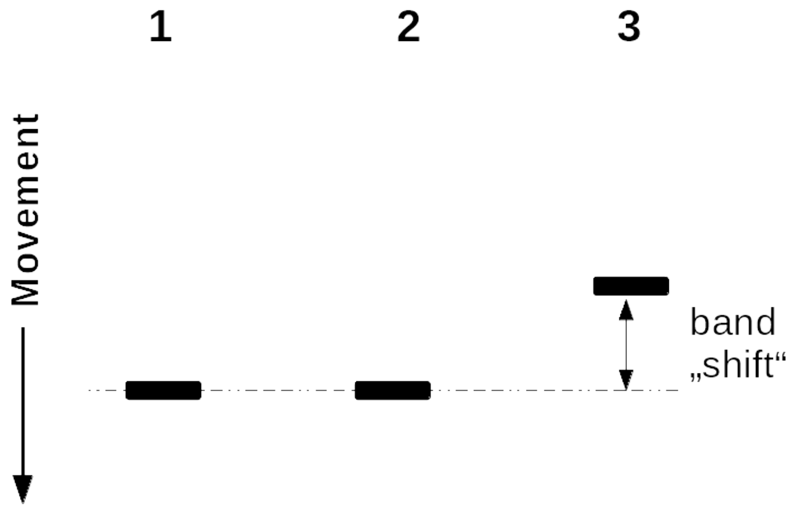
Main plot: cumulative distribution.

Activating TF binding sites are closer to the TSS than repressing TF binding sites ($p = 4.7\times10^{-2}$).

Whitfield et al. Genome Biology 2012, 13:R50

7.3 Experimental TFBS detection: EMSA shift assay

An **electrophoretic mobility shift assay** (EMSA) or **gel shift assay** is an affinity electrophoresis technique for identifying **specific binding** of a protein–DNA or protein–RNA pair **in vitro**.



The samples are electro-phoretically separated on a polyacrylamide or agarose gel.

The results are visualized by radioactive labelling of the DNA with ^{32}P or by tagging a fluorescent dye.

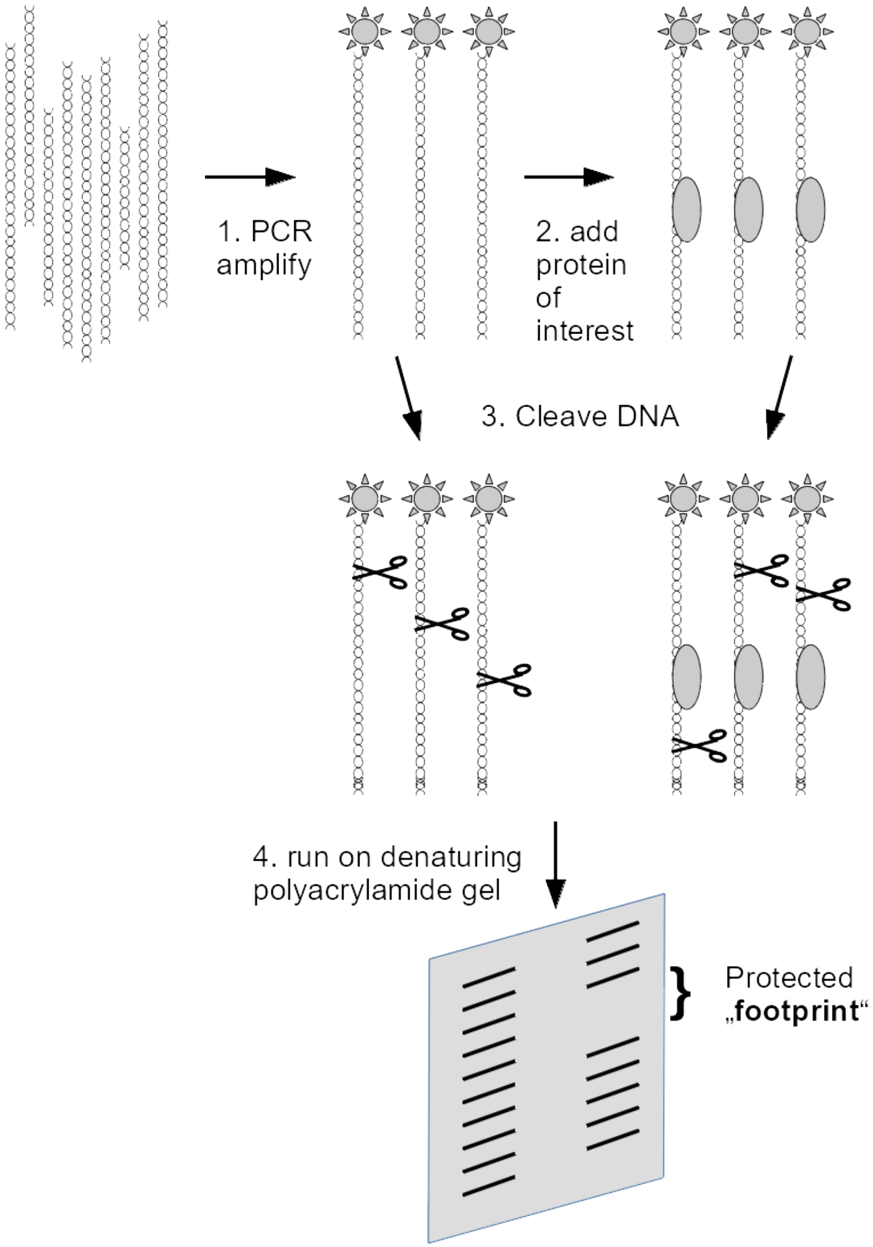
Control lane (1) contains DNA probe without protein. Obtained at the end of the experiment is a single band that corresponds to the unbound DNA.

Lanes (2) and (3) each contain a mixture of the DNA with a protein. If the protein actually binds to the DNA (3), this lane will show an up-shifted band relative to (1) which is due to the larger and less mobile protein:DNA complex.

7.3.2. DNAse footprinting

In DNAse footprinting, a **DNAse enzyme** is added to the sample that **cleaves** DNA non-specifically at many positions.
On a polyacrylamide gel, the cleaved DNA fragments of differing lengths will show up as different lanes (left figure).

In a second experiment, the protein of interest is added (right lane). If this protein binds specifically at a particular position of the DNA, it will prevent cleavage by DNAse at this position. Then, this DNA fragment cannot be found on the gel (bottom, right lane) and represents thus the specific binding motif in the investigated DNA sequence for the protein.


1. PCR amplify
2. add protein of interest
3. Cleave DNA
4. run on denaturing polyacrylamide gel
Protected „footprint“

7.3.3. High-throughput methods

There exist also several high-throughput *in vitro* methods to measure the TF-DNA binding affinity of large numbers of DNA variants.

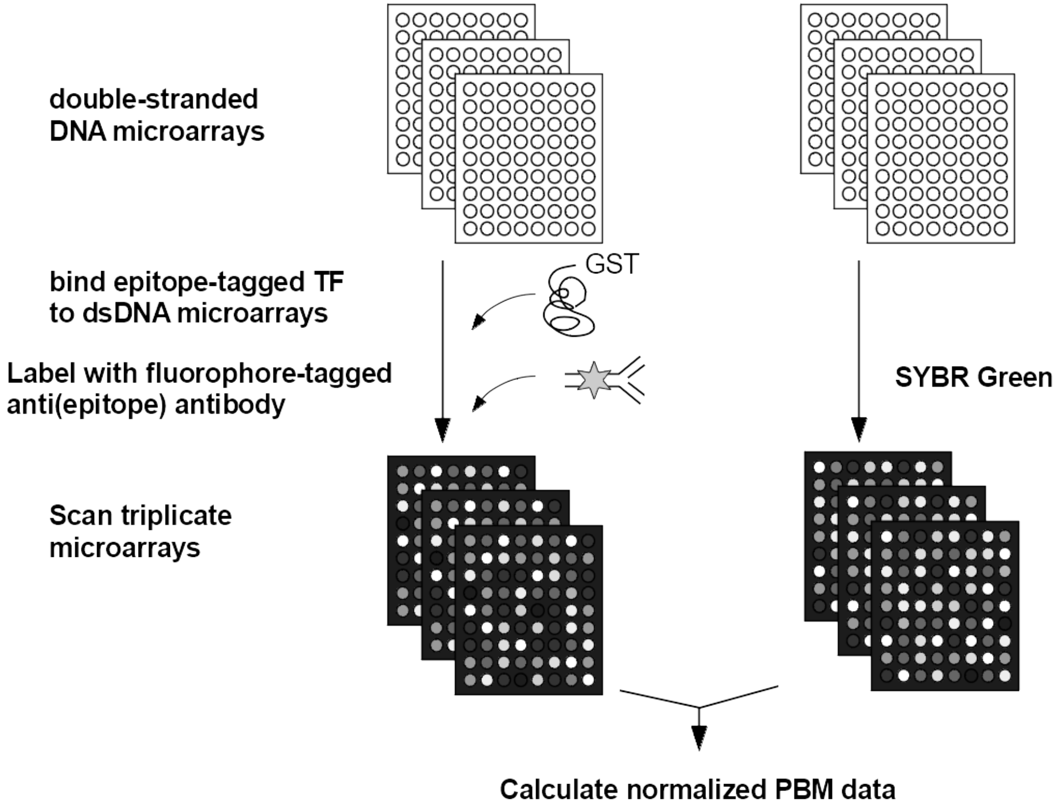
One of them is a DNA microarray-based method called protein binding microarray (PBM) (Berger and Bulyk, 2006).

With this technology, one can characterize the binding specificity of a single DNA binding protein *in vitro* by adding it to the wells of a microarray spotted with a large number of putative binding sites in double-stranded DNA.

7.3.3. Protein binding microarray

The protein of interest carrying an epitope tag is expressed and purified and then applied to the microarray.

After removing nonspecifically bound protein by a washing step, the protein is detected in a labeling step where a fluorophore-conjugated antibody binds specifically to the epitope tag.


double-stranded DNA microarrays
GST
bind epitope-tagged TF to dsDNA microarrays
Label with fluorophore-tagged anti(epitope) antibody
SYBR Green
Scan triplicate microarrays
Calculate normalized PBM data

One identifies all spots carrying a significant amount of protein.

In the DNA sequences belonging to these spots, one identifies enriched DNA binding site motifs for the DNA binding protein of interest.

7.3.3. Problems of in vitro methods

Due to the short length of TFBS motifs and the relatively small number of invariant nucleotide positions in it, some motifs are found millions of times in the genome.

Thus, although any motif instance could potentially be bound *in vivo*, only about 1 in 500 are actually bound in organisms with large genomes.

As a specific example, the mouse genome contains ~8 million instances of a match to the binding site motif of **GATA-binding factor 1**, but only ~15,000 DNA segments are bound by this transcription factor in erythroid cells (Hardison and Taylor, 2012).

7.3.3. in vivo methods

To overcome the limitations of *in vitro* assays, new massively parallel methods such as ChIP-chip and ChIP-seq can identify TF binding sites *in vivo*.

These methods are based on DNA microarrays and new sequencing techniques, respectively.

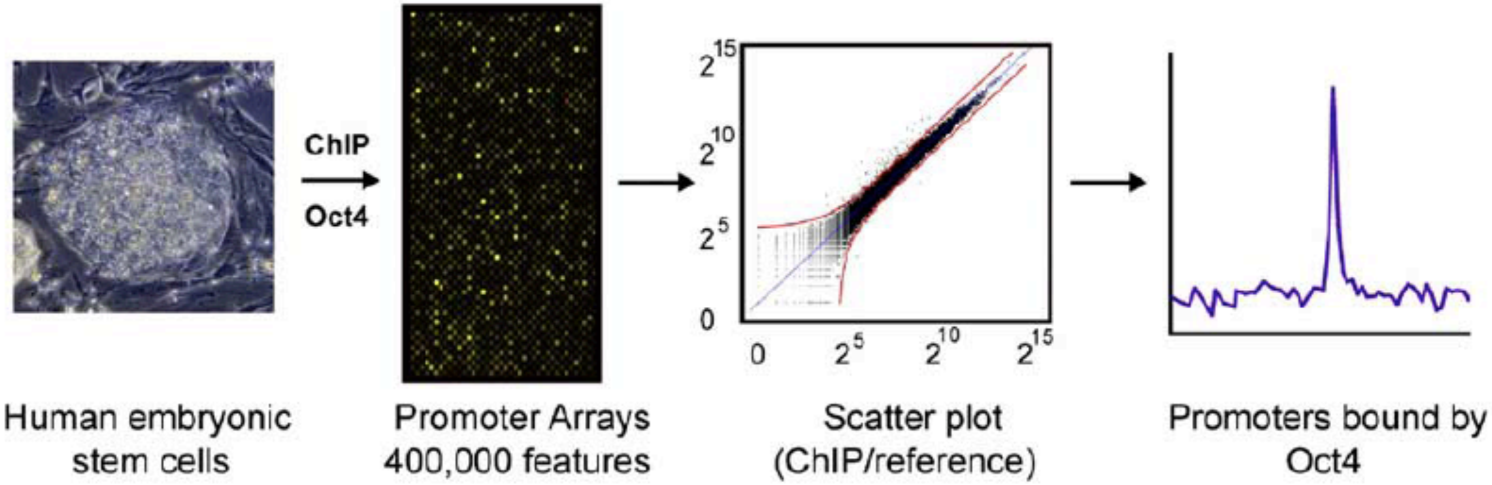
In **Chip-seq** experiments, a cellular extract is purified using an antibody against a particular TF.

Then, the DNA sequences bound to the TF are digested using a restriction enzyme. The remaining DNA can be considered as tightly bound to the TF.

This DNA is washed and sequenced.

All DNA reads correspond to DNA fragments that were bound to the TF before.

Which TF binds where?


ChIP
Oct4
2^{15}
2^{10}
2^{5}
0
0
2^{5}
2^{10}
2^{15}
Human embryonic stem cells
Promoter Arrays 400,000 features
Scatter plot (ChIP/reference)
Promoters bound by Oct4

Chromatin immuno precipitation: use e.g. antibody against Oct4

➔ ”fish“ all DNA fragments that bind Oct4

➔ sequence DNA fragments bound to Oct4

➔ align them + extract characteristic sequence features

➔ Oct4 binding motif

Boyer et al. Cell 122, 947 (2005)

7.4. Position-specific scoring matrix

PSSMs are used to represent motifs (patterns) in biological sequences.

	Position 1	Position 2	Position 3	Position 4
Sequence 1	A	C	A	T
Sequence 2	A	C	C	T
Sequence 3	A	G	G	G
Sequence 4	C	C	T	G
Sequence 5	A	T	A	G
Sequence 6	C	A	G	T

Toy example of six DNA sequences that are 4 bp long.

	Position 1	Position 2	Position 3	Position 4
Frequency A	4	1	2	0
Frequency C	2	3	1	0
Frequency G	0	1	2	3
Frequency T	0	1	1	3

Frequency n_i^j of nucleotide bases *(i)* at the 4 positions *(j)*.

Out of 6 × 4 = 24 nucleotides in the four sequences, 7 are adenine, 6 are cytosine, 6 are guanine, and 5 are thymine. Thus, the frequencies p_i of the four nucleotides are 0.29 (A), 0.25 (C and G), and 0.21 (T).

7.4. Position-specific scoring matrix

From the frequency matrix, one computes the score matrix using

$$s_i^j = ln \frac{\left(n_i^j + p_i\right)/(N+1)}{p_i},$$

where, N is the number of considered sequences (here, N = 6).

	Position 1	**Position 2**	**Position 3**	**Position 4**
score A	0.75	-0.45	0.12	-1.94
score C	0.25	0.62	-0.34	-1.94
score G	-1.94	-0.34	0.25	0.62
score T	-1.94	-0.19	-0.19	0.78

Adding the frequencies p_i in the denominator and dividing by N + 1 avoids problematic cases with $n_i^j = 0$ where the logarithm would not be defined otherwise.

Positions with score s_i^j = 0 occur at the frequency that is expected randomly, positive entries denote enriched nucleotides at this position, negative entries denote the opposite case.

7.5 Binding free energy models

The binding of a TF to single- or double-stranded DNA is an elementary biomolecular association reaction.

The binding free energy model of Djordjevic (2003) describes the reversible binding of a TF to a short piece of DNA with sequence S,

$$TF + DNA \underset{k_{diss}}{\overset{k_{bind}}{\rightleftarrows}} TF - DNA$$

with the sequence-dependent rate constants k_{bind} and k_{diss} for TF binding and dissociation, respectively.

In equilibrium, $[TF] \cdot [S] \cdot k_{bind}(S) = [TF{:}S] \cdot k_{diss}(S)$

The ratio of the bound and free forms thus equals the ratio of the two rate constants and is equal to $\frac{[TF:S]}{[TF][S]} = \frac{k_{bind}(S)}{k_{diss}(S)} = \frac{1}{K_D} = c \cdot e^{-\frac{\Delta G(S)}{kT}}$, where *c* is a constant and $\Delta G(S)$ is the (usually negative) binding free energy of the TF to its recognition sequence *S* on the DNA.

7.5 Binding free energy models

Let us consider the binding reaction of two molecules *L* and *M*:

$$L + M \overset{\rightarrow}{\underset{\leftarrow}{}} LM.$$

The dissociation equilibrium constant K_D is defined as:

$$K_D = \frac{[L][M]}{[LM]} = \frac{k_{diss}}{k_{bind}}$$

, where [*L*], [*M*], and [*LM*] are the molecular concentrations of *L* and *M* and of the complex *LM*.

In equilibrium, we may take *T* as the total concentration of molecule *L*

$$T = [L] + [LM].$$

y is the fraction of molecules *L* that have reacted (bound),

$$y = \frac{[LM]}{[LM] + [L]}$$

.

7.5 Binding free energy models

$$y = \frac{[LM]}{[LM] + [L]}$$

Substituting [*LM*] by [*L*] [*M*] / K_D gives

$$y = \frac{([L][M])/K_D}{([L][M])/K_D + [L]} = \frac{([M])/K_D}{([M])/K_D + 1}$$

.

When a solution contains both the DNA sequence and the TF with total concentration n_{tf}, the equilibrium probability that the DNA is bound to a TF molecule is (replace in upper eq. [M] by n_{tf}):

$$p(TF\ is\ bound\ to\ S) = \frac{\frac{1}{K_D} \cdot n_{tf}}{\frac{1}{K_D} \cdot n_{tf} + 1} = \frac{c \cdot e^{-\Delta G(S)/kT} \cdot n_{tf}}{c \cdot e^{-\Delta G(S)/kT} \cdot n_{tf} + 1}$$

We multiply this with $e^{+\Delta G(S)/kT}$ and divide by c·n_{tf}.

7.5 Binding free energy models

This gives: $P(TF\ is\ bound\ to\ S) = \frac{1}{1+\frac{e^{\Delta G(S_i)/kT}}{c \cdot n_{tf}}}$,

where ΔG(S_i) : free energy of the TF binding to S_i .

We set $c \cdot n_{tf} = e^{\frac{\mu}{kT}}$ or $\mu = kT \cdot ln(c \cdot n_{tf})$

μ : chemical potential set by the TF concentration. This gives

$$P(TF\ is\ bound\ to\ S) = \frac{1}{1 + e^{(\Delta G(S_i) - \mu)/kT}}$$

,

This is the so-called Fermi-Dirac form of binding probability.

A sequence having a binding free energy well below the chemical potential ($\Delta G(S_i) - \mu \ll 0$) is almost always bound to the TF.
($P(TF\ is\ bound\ to\ S) \to 1$ because the exponential term is very small.)

In cases when the binding free energy is well above the chemical potential, the sequence is rarely bound.

7.5 Binding free energy models

The binding energy model (BEM) uses a vector of (free) energy contributions, $\vec{E}$.

For any sequence S_i, the binding energy predicted by the BEM model is

$$E(S_i) = \vec{E} \cdot \vec{S}_i$$

where $\vec{S}_i$ is the vector encoding of sequence S_i that can include whatever features of the sequence are relevant to its binding energy.

If the only relevant features are which bases occur at each position within the binding site, then $\vec{E}$ will be a PSSM with the characteristic that each element is a (free) energy contribution.

7.5 Binding free energy models

When the (free) energy contributions of each position are independent, $\vec{E} \cdot \vec{S}_i$ can be written as:

$$E(S_i) = \sum_{b=A}^{T} \sum_{m=1}^{L} \epsilon(b,m) S_i(b,m)$$

where *L* : length of the binding site, *ε*(*b*, *m*) : (free) energy contributions of base *b* at position *m*, and $S_i(b, m)$: indicator variable with $S_i(b, m)$ = 1 if base *b* occurs at position *m* of sequence S_i and $S_i(b, m)$ = 0 otherwise.

If the positions are not independent, one can include pairwise interactions between adjacent positions *m* and *n* by adding **interaction terms** to the energy function such that $\vec{E} \cdot \vec{S}_i$ is

$$E(S_i) = \sum_{b=A}^{T} \sum_{m=1}^{L} \epsilon(b,m) S_i(b,m) + \sum_{m=1}^{L-1} \sum_{n=m+1}^{L} \sum_{b=A}^{T} \sum_{c=A}^{T} \epsilon(b,m,c,n) S_i(b,m,c,n)$$

where *ε*(*b*, *m*, *c*, *n*) : energy contribution of having base *b* at position *m* and base *c* at position *n*.

7.6 Cis-regulatory motifs

Although hundreds of TFs are present in a typical eukaryotic cell, the complex expression patterns of thousands of genes can only be implemented by a regulatory machinery involving combinations of TFs.

Thus, prokaryotic and eukaryotic gene promoters often bind multiple TFs simultaneously.

These TFs may also make structural contacts to eachother and thus affect their mutual binding affinities in a cooperative manner.

In that case, for steric reasons, the distance between TFBSs of contacting TFs is constrained to a certain range.

All such combinatorial and cooperative effects are difficult to capture in a quantitative manner by a PSSM-based approach.

7.6 Cis-regulatory motifs

A **cluster** of TFBSs is termed a **cis-regulatory module** (CRM).
The existence of such a CRM is a footprint of a **TF complex**.

For metazoans, a typical CRM may be more than 500 bp long and is made up of 10 to 50 TFBSs to which between three and 15 different sequence-specific TFs bind.

If there exist multiple similar binding sites, this

- enhances the sensitivity for a TF,
- results in a more robust transcriptional response and
- affects how morphogen TFs are activated when the local TF concentration is low,

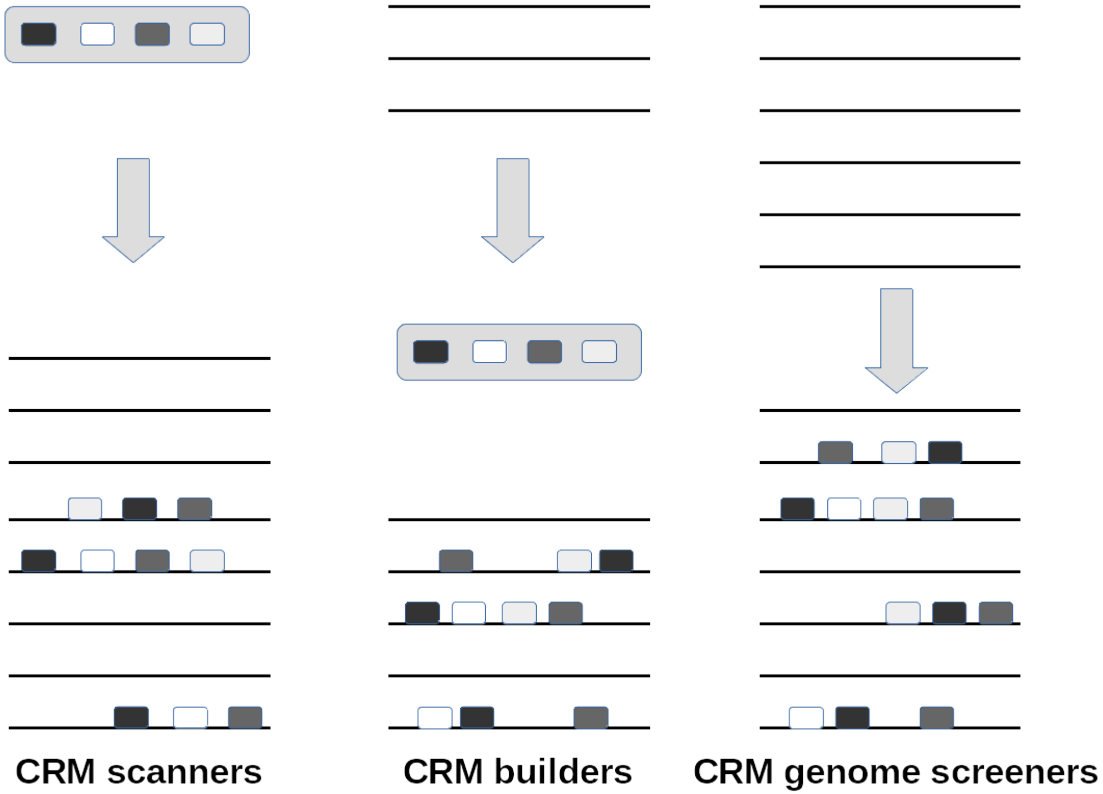
or they may simply favor the binding of a homo-oligomeric TF (e.g. p53, or NF-κB).

Some transcription factors such as the TF pair Oct4 and Sox2 have well known interaction partners.

7.6 identify Cis-regulatory motifs

(left) CRM scanners require user-defined motif combinations as input to search for putative regulatory regions.

(middle) CRM builders analyze a set of co-regulated genes as input and produce candidate motif combinations, as well as similar target regions.



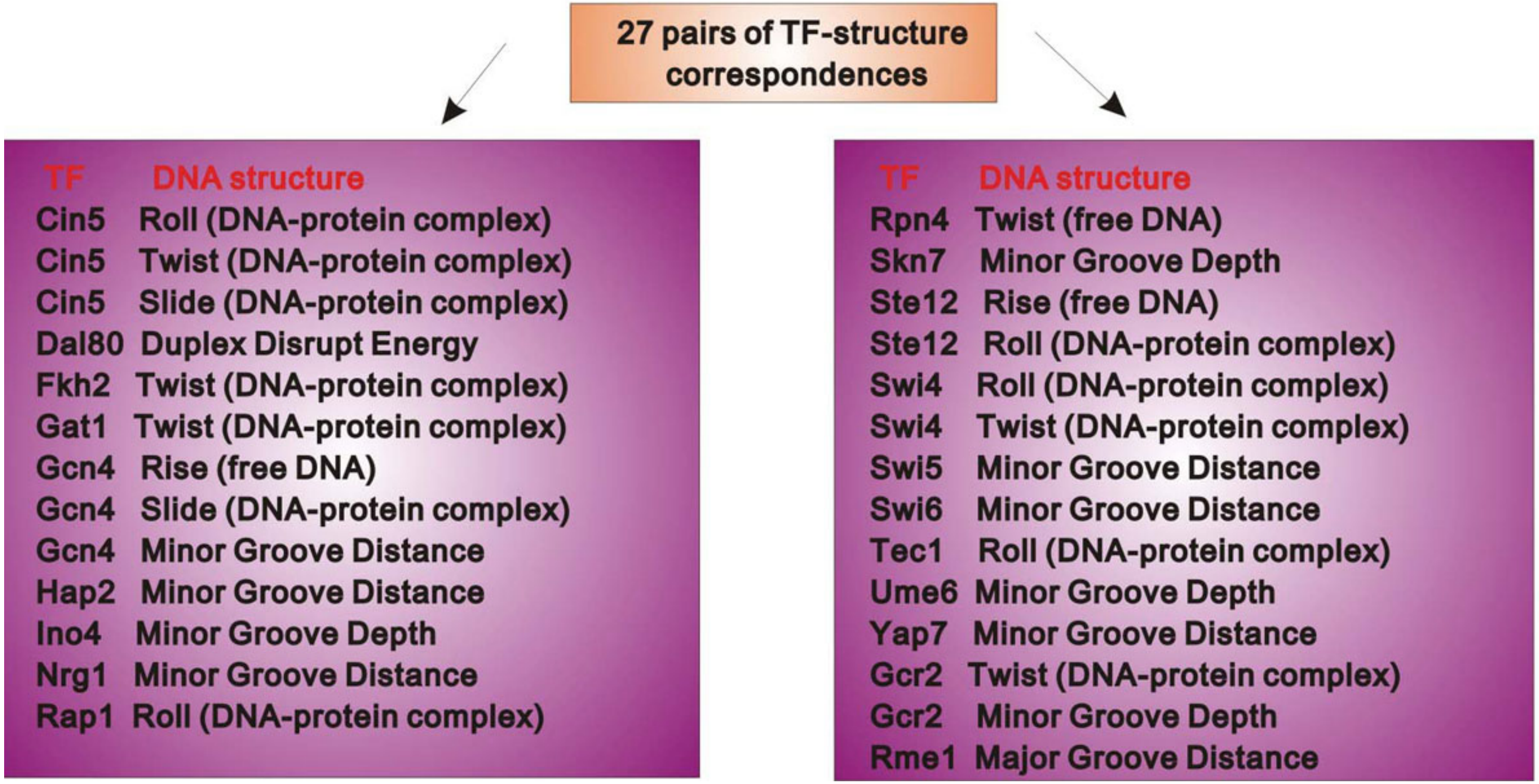
(right) CRM genome screeners search for homotypic or heterotypic motif clusters without making assumptions about the involved TFs.

What do TFs recognize?

(1) Amino acids of TFs make specific contacts (e.g. hydrogen bonds) with DNA base pairs

(2) DNA conformation depends on its sequence

→ Some TFs „measure“ different aspects of the DNA conformation



27 pairs of TF-structure correspondences

TF	DNA structure
Cin5	Roll (DNA-protein complex)
Cin5	Twist (DNA-protein complex)
Cin5	Slide (DNA-protein complex)
Dal80	Duplex Disrupt Energy
Fkh2	Twist (DNA-protein complex)
Gat1	Twist (DNA-protein complex)
Gcn4	Rise (free DNA)
Gcn4	Slide (DNA-protein complex)
Gcn4	Minor Groove Distance
Hap2	Minor Groove Distance
Ino4	Minor Groove Depth
Nrg1	Minor Groove Distance
Rap1	Roll (DNA-protein complex)

TF	DNA structure
Rpn4	Twist (free DNA)
Skn7	Minor Groove Depth
Ste12	Rise (free DNA)
Ste12	Roll (DNA-protein complex)
Swi4	Roll (DNA-protein complex)
Swi4	Twist (DNA-protein complex)
Swi5	Minor Groove Distance
Swi6	Minor Groove Distance
Tec1	Roll (DNA-protein complex)
Ume6	Minor Groove Depth
Yap7	Minor Groove Distance
Gcr2	Twist (DNA-protein complex)
Gcr2	Minor Groove Depth
Rme1	Major Groove Distance

Dai et al. *BMC Genomics* 2015, **16**(Suppl 3):S8

Co-expression of TFs and target genes?

Overexpression of a TF often leads to induction or repression of target genes.

This suggests that many target genes can be regulated simply by the abundance (expression levels) of the TF.

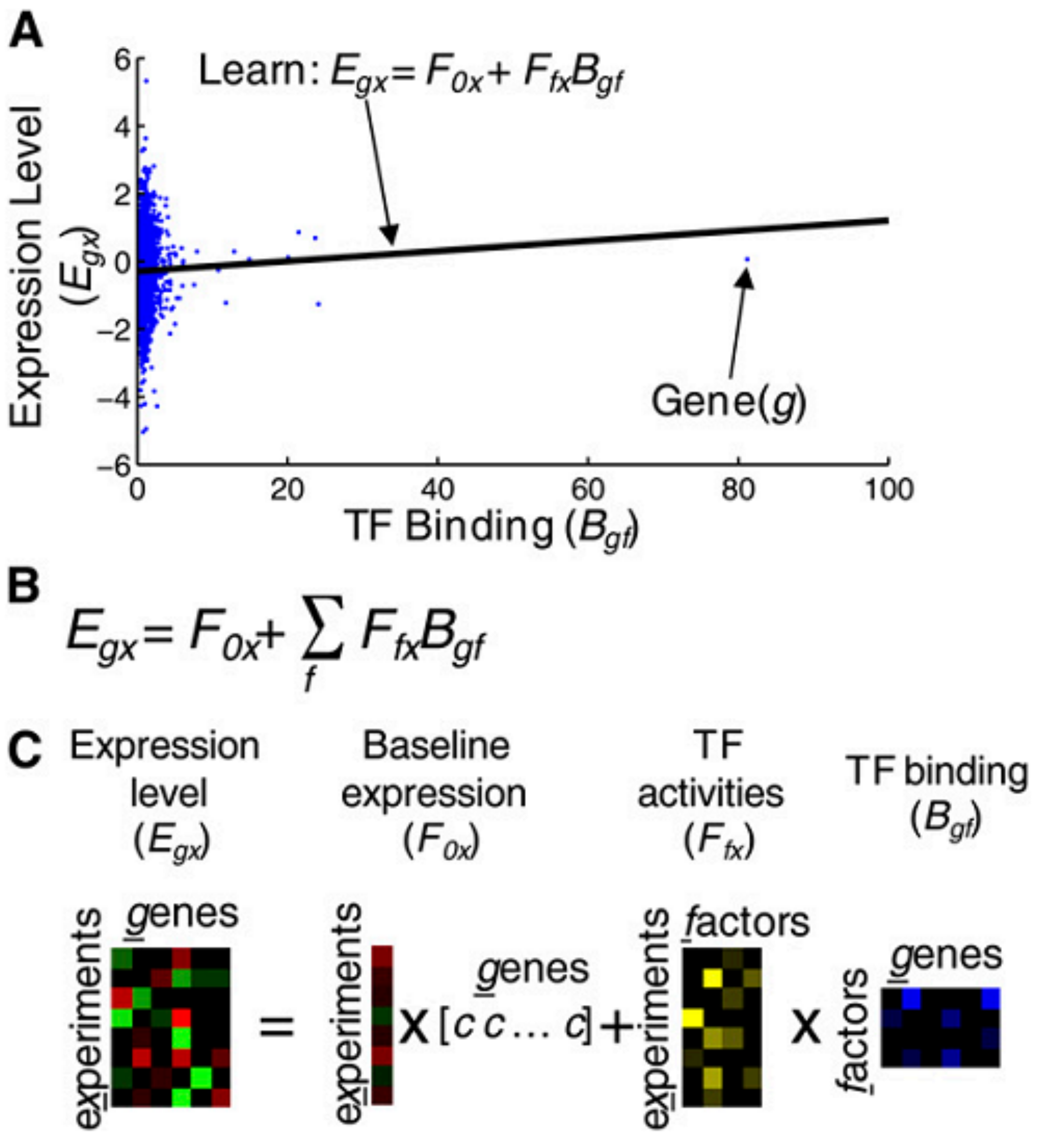
However, across 1000 microarray expression experiments for yeast, the **correlation** between a TF's expression and that of its ChIP-based targets was typically **very low** (only between 0 and 0.25)!

At least some of this (small) correlation can be accounted for by the fact that a subset of TFs autoregulate themselves.

→ In yeast, TF expression accounts for only a minority of the regulation of TF activity.

Hughes, de Boer (2013) Genetics 195, 9-36

Using regression to predict gene expression



(A) Example where the relationship between expression level (E_{gx}) and TF binding to promoters (B_{gf}) is found for a single experiment (x) and a single TF (f). Here, the model learns 2 parameters: the background expression level for all genes in the experiment (F_{0x}) and the activity of the transcription factor in the given experiment (F_{fx}).

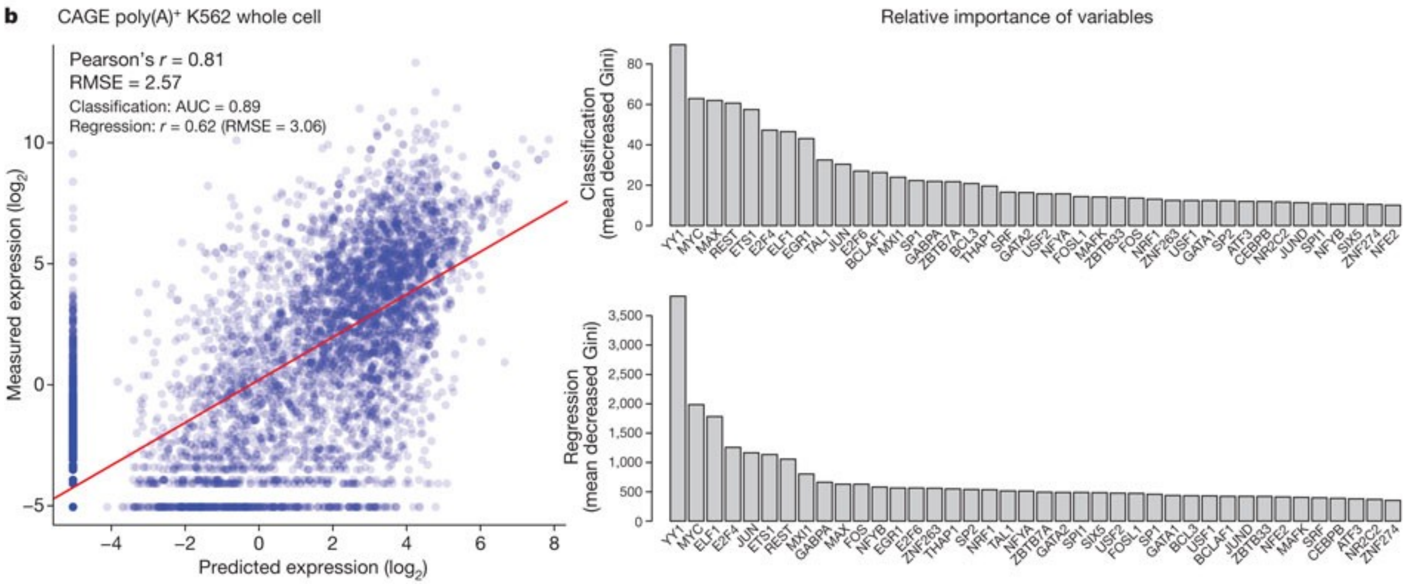
(B) The generalized equation for multiple factors and multiple experiments.

(C) Matrix representation of the generalized equation.

Baseline expression is the same for all genes and so is represented as a single vector multiplied by a row vector of constants where c = 1/(no. genes).

Hughes, de Boer (2013) Genetics 195, 9-36

ENCODE



AUC: area under curve;

Gini: Gini coefficient;

RMSE: root mean square error.

The ENCODE project studied how well the occupancy of TFBS is correlated with RNA production in human K562 cells.

(left) Scatter plot comparing a linear regression curve (red line) with observed values for RNA production (blue circles).

(right) Bar graphs showing the most important TFs both in the initial classification phase (top) or the quantitative regression phase (bottom). Larger values indicate increasing importance of the variable in the model.

ENCODE Project Consortium, Nature 489, 57 (2012)

Transcription factors in human: ENCODE

Some TFs can either activate or repress target genes.

The TF YY1 shows the largest mixed group of target genes.

TF	Ubiquitously activated	Ubiquitously repressed
YY1	COQ5[cd]	AC091153.1
	CPNE1	ATP5O
	CPSF2 [cd]	BIRC6[d]
	CR613718	CAPZA2
	IP6K2[a]	CXorf26
	NARS[ac]	DKFZp434H247
	PAK4[d]	EFHA1
	PSMB4[ac]	MRPS10[c]
	UBR5	MRPS18B[acd]
		NUP160
		OXCT1
		PSMD8[ac]
		SNX27
		SNX3[ad]
		SRP68[ad]
		TNKS

1UBD.pdb
human YY1

Whitfield et al. Genome Biology 2012, 13:R50

Summary transcription

- Gene transcription (mRNA levels) is controlled by transcription factors (activating / repressing) and by microRNAs (degrading) (see later lecture)

- Binding regions of TFs are ca. 5 – 10 bp stretches of DNA

- Global TFs regulate hundreds of target genes

- Global TFs often act together with more specific TFs

- TF expression only weakly correlated with expression of target genes (yeast)

- Some TFs can activate or repress target genes. Use similar binding motifs for this.