# A new era in transcription factor research

**Edgar Wingender**

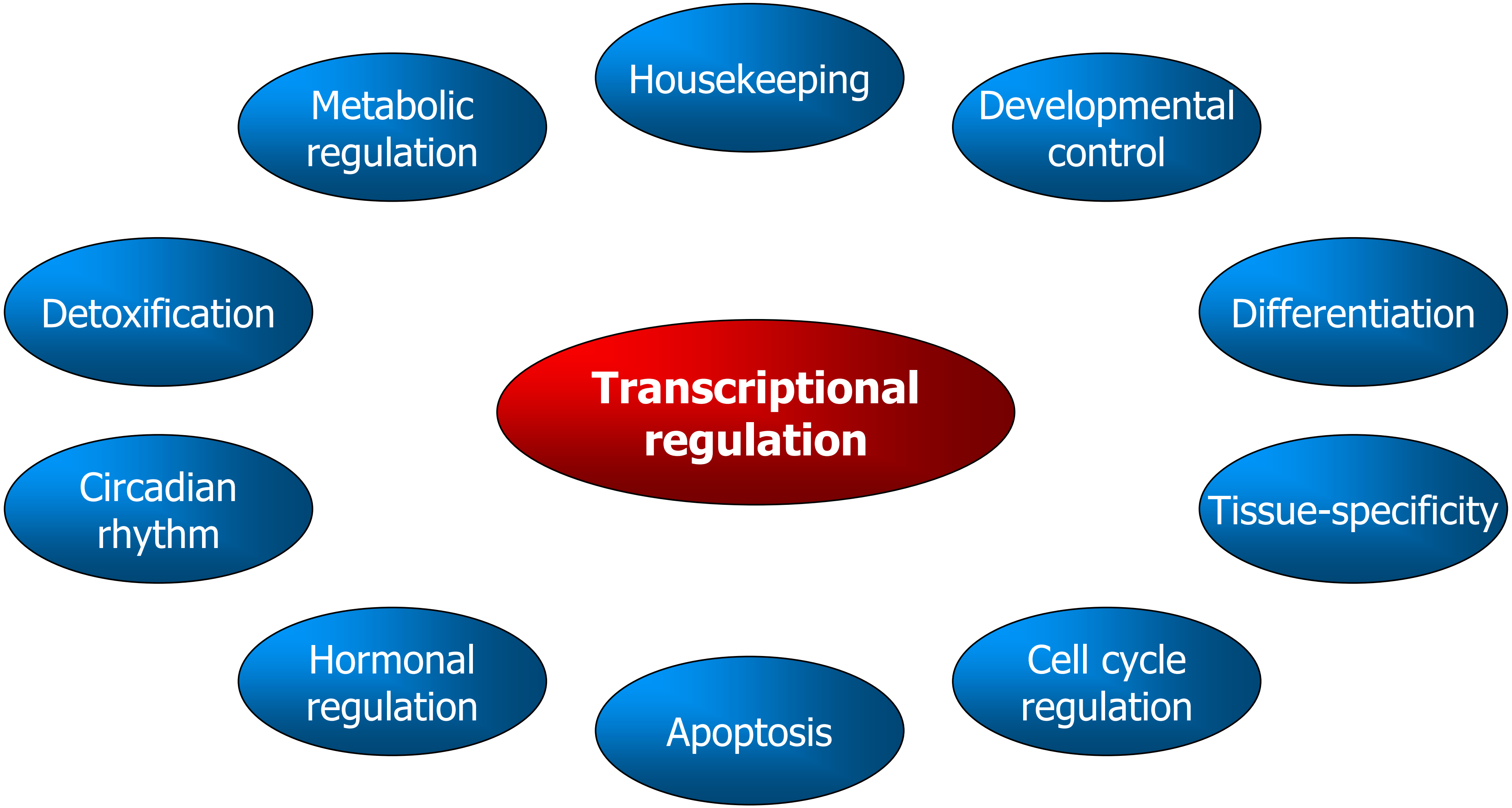
UNIVERSITÄTSMEDIZIN GÖTTINGEN UMG

Dept. Bioinformatics,
Medical School
Georg August University
Göttingen, Germany

BIOBASE
BIOLOGICAL DATABASES

Wolfenbüttel, Germany
Beverly, MA / USA
Bangalore, India
Yokohama, Japan

# Impact of transcriptional regulation



## The goals of transcription factor research:

(1) To understand how the gene-specificity of transcriptional regulation is achieved

(2) To develop genome-wide maps of transcription factor binding sites (TFBSs)

(3) To enable prediction of new TFBSs

(4) To comprehend the complex structure of regulatory genome regions (promoters, enhancers, etc.)

(5) To predict the DNA-binding specificity of new transcription factors (TFs)

(6) To construct system-wide transcription networks

(7) To understand transcriptional dysregulation under disease conditions

(8) To render transcriptional regulation amenable for targeted alterations

## The goals of transcription factor research:

(1) To understand how the gene-specificity of transcriptional regulation is achieved

(2) To develop genome-wide maps of transcription factor binding sites (TFBSs)

**Biology**

(3) To enable prediction of new TFBSs

(4) To comprehend the complex structure of regulatory genome regions (promoters, enhancers, etc.)

(5) To predict the DNA-binding specificity of new transcription factors (TFs)

**Bioinformatics**

(6) To construct system-wide transcription networks

**Systems biology**

(7) To understand transcriptional dysregulation under disease conditions

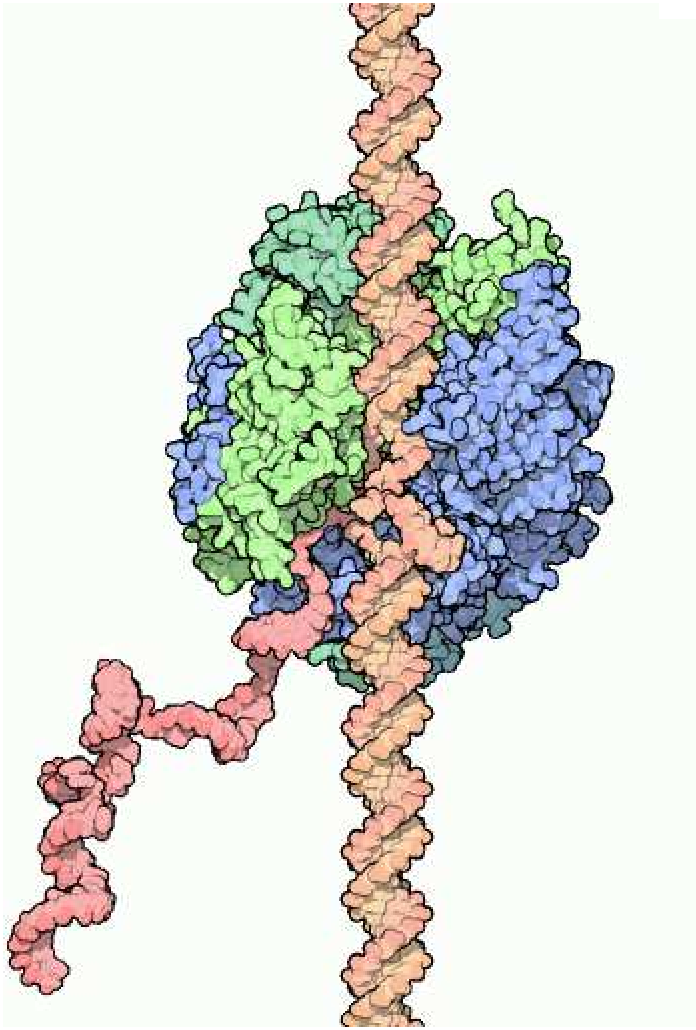
(8) To render transcriptional regulation amenable for targeted alterations

**Synthetic biology**

## The DNA-dependent RNA-polymerases

RNA polymerase II
from *S. cerevisiae*

PDB entry
1I6H



## The DNA-dependent RNA-polymerases

In *E. coli*:

| | | |
|---|---|---|
| RNA polymerase (RNAP): | all genes | 5 subunits |

In Eukaryotes:

| | | |
|---|---|---|
| RNA polymerase I (A): | 45S-rRNA | 7-14 subunits |
| RNA polymerase II (B): | mRNA | 12 subunits |
| RNA polymerase III (C): | tRNA, 5S-rRNA | 10 subunits |
| RNA polymerase IV: | siRNA in plants | |

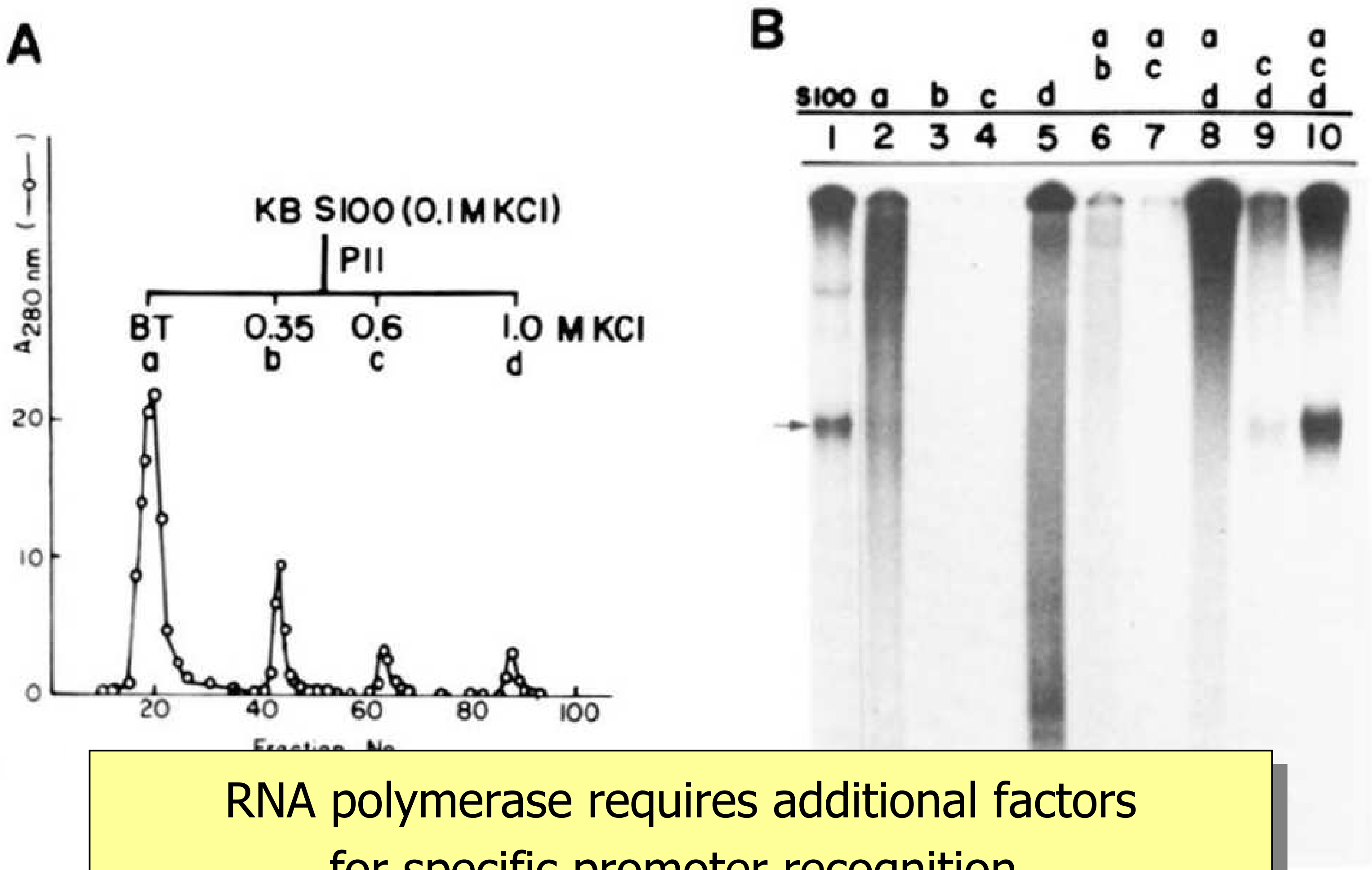
## The DNA-dependent RNA-polymerases

Goal #1:

Understand how the specificity of eukaryotic transcriptional regulation is achieved?

## Specificity of transcriptional regulation

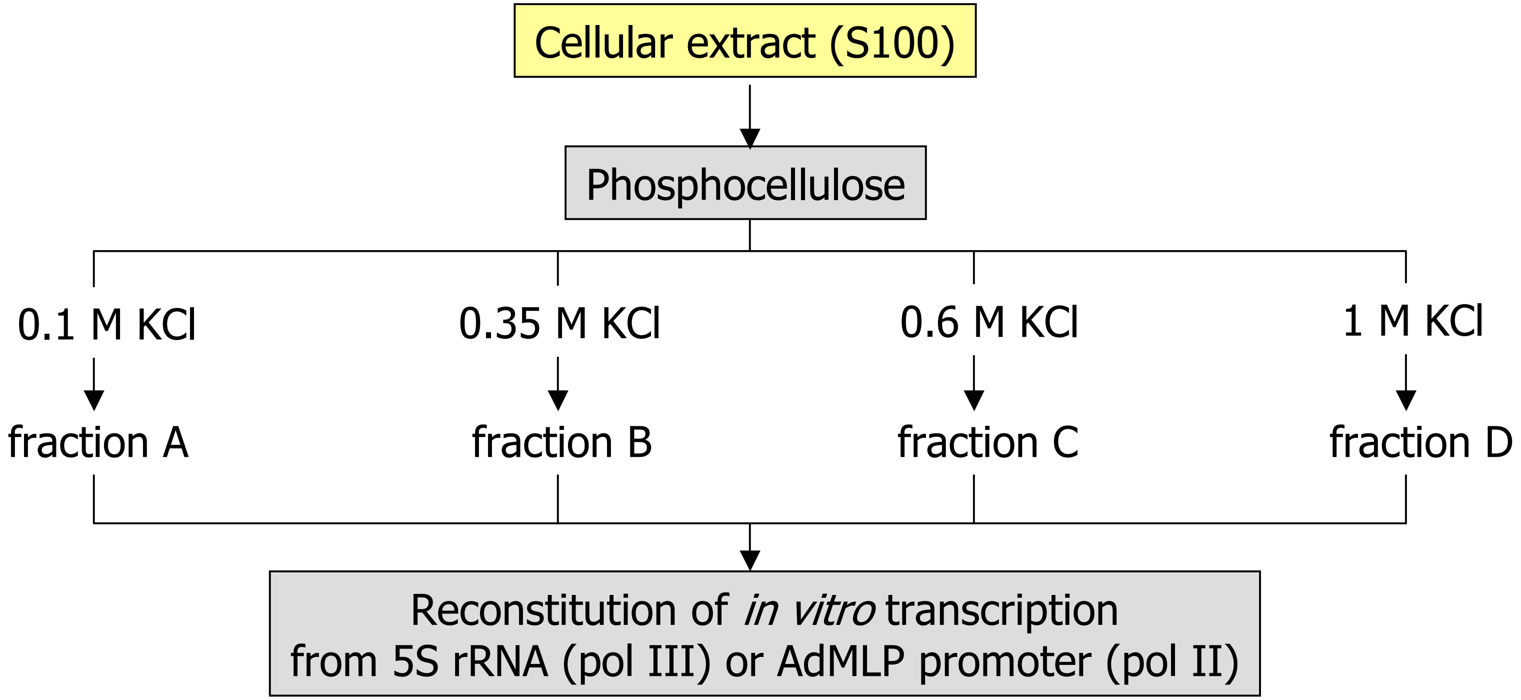
### General transcription factors




RNA polymerase requires additional factors for specific promoter recognition.


Multiple factors required for accurate initiation of transcription by purified RNA polymerase II. Matsui T, Segall J, Weil PA, Roeder RG. J Biol Chem. 1980 Dec 25;255(24):11992-6.
PMID: 7440580

## General transcription factors



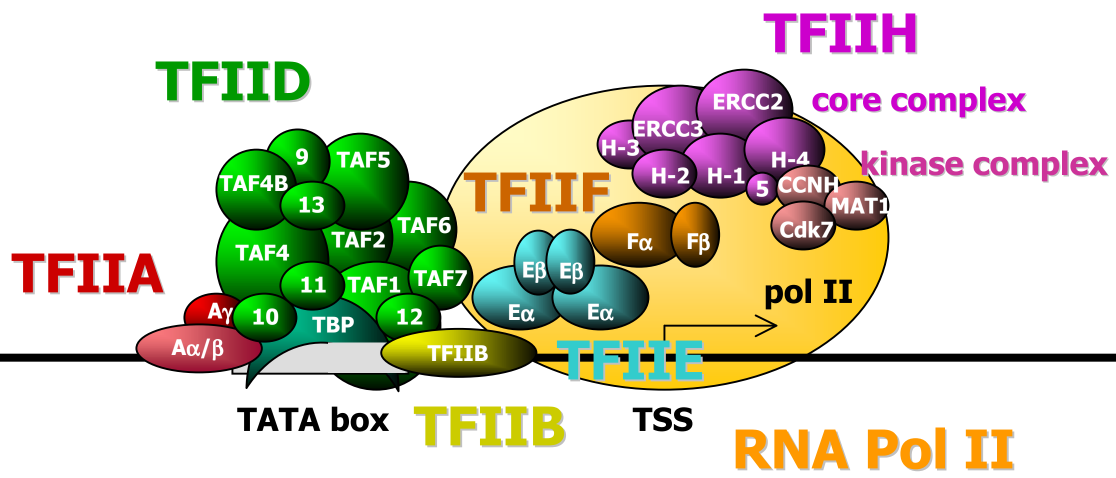

| | | | | |
|---|---|---|---|---|
| **pol III:** | TFIIIA | TFIIIB | TFIIIC | -- |
| **pol II:** | TFIIA | -- | TFIIB | TFIID |

Multiple factors required for accurate initiation of transcription by purified RNA polymerase II. Matsui T, Segall J, Weil PA, Roeder RG. J Biol Chem. 1980 Dec 25;255(24):11992-6.
PMID: 7440580

# Specificity of transcriptional regulation

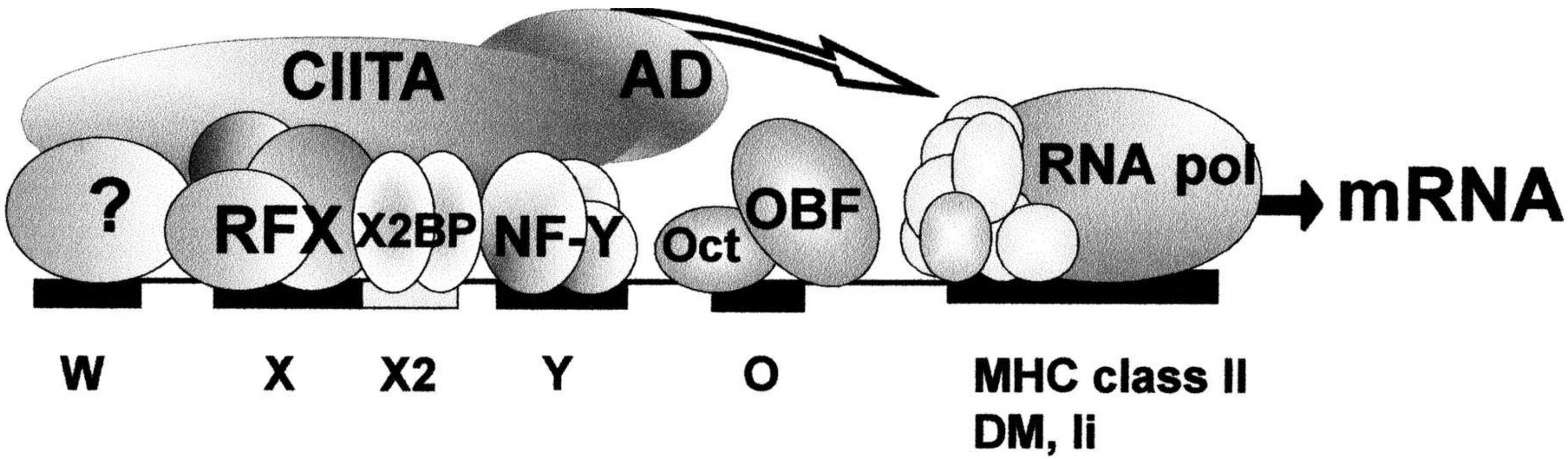
## General transcription factors

The pre-initiation transcription complex



## Specificity of transcriptional regulation

### Enhanceosome




Master

In addition to the assembly of general transcription factors, „upstream“ factors are required. Many of them are sequence-specific DNA-binding proteins.
Their task is to provide a favorable chromatin structure and/or to facilitate the assembly of the general transcription factors.

## Definition

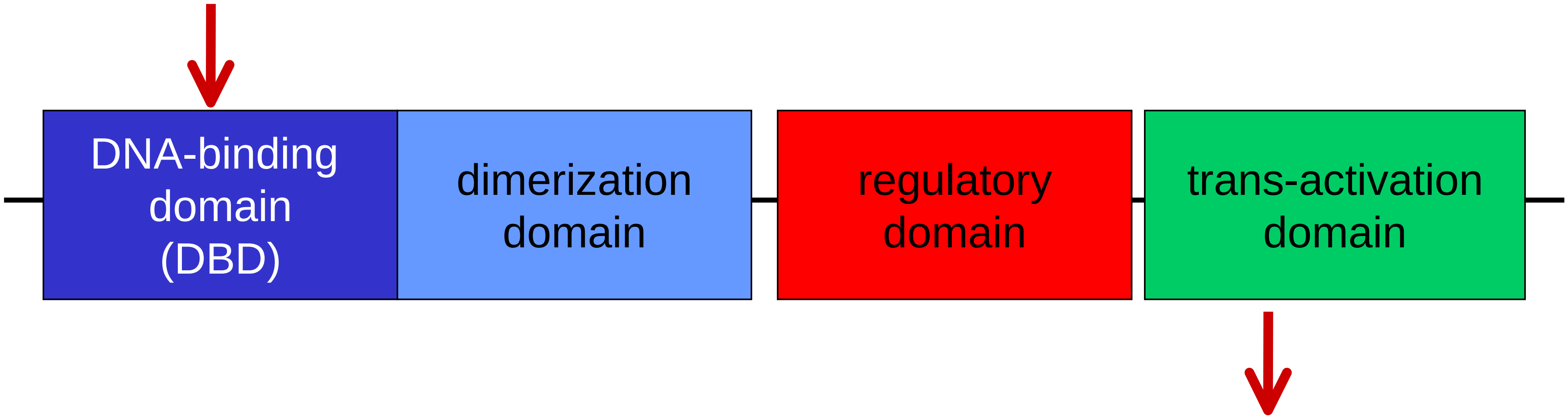
# What is a transcription factor?

A transcription factor is a protein that regulates
transcription
by specific interaction with DNA
or, after nuclear translocation, by stoichiometric
interaction with a protein that can be assembled into a
sequence-specific DNA-protein complex.

Wingender, E. (1997). Classification scheme of eukaryotic transcription factors. Mol. Biol. 31, 483-497 (1997).

## Modular structure of a transcription factor

Primary function:
To recognize *cis*-regulatory elements




Ultimate function:
To activate transcription

Goal #2:

To generate a comprehensive map of genome-bound proteins.

## The first compilation

Volume 16 Number 5 1988

Nucleic Acids Research

**Compilation of transcription regulating proteins**

Edgar Wingender

Gesellschaft für Biotechnologische Forschung mbH, Mascheroder Weg 1, D-3300 Braunschweig, FRG

Received November 28, 1987; Revised and Accepted January 28, 1988

Wingender E. Compilation of transcription regulating proteins. Nucleic Acids Res. 1988 Mar 25;16(5):1879-902.
PMID: 3282223

## The first compilation

| gene, gene product | species/tissue[a] | protein interacting region[b] | method[c] | sequence motif | factor bound | ref. |
|---|---|---|---|---|---|---|
| α-actin | chicken/rat myocytes | -83 to -78 | 3, 4b | | myoc.-spec.f. | 1 |
| | /rat non-myocytes | " | 3, 4b | | distinct factor | 1 |
| β-actin | rat/HeLa | | 3 (compet. with c-fos) | small, dyad symmetry | SRF ? | 2 |
| actin 5C | Drosophila | -38 to +34 | 1a | TATAAAA | B factor | 3 |
| actin (cytoskeletal) | Xenopus laevis /HeLa | -94 to -75 (SRE) | 1a, 4a, 4b | AAGATgcCCAtATtTGGcgATCTT | SRF (?) | 4 |
| Ad2MLP (adenovirus 2 major late prom.) | adenovirus/HeLa | -68 to -49 | 1a, 3, 4a, 4b | TGTAGGCCACGTGACCGG | UEF, USF, MLTF | 5-10 |
| | | -63 to -52 | 1b | GGCCACGTGACC | USF | 9 |
| | | -50 to -10 | 1a | TATAAAA | ? | 8 |
| | | -40 to +35 | 1a | TATAAAA | TFIID | 9 |
| | | -34 to -22 | 1b | TATAAAA | TFIID | 9 |
| Adh (alcohol dehydrogenase) -distal prom. | Drosophila | - 85 to - 47 | 1a | GCTGCtGCTGCatcCGTCGaCGTCG | Adf-1 | 11 |
| -proxim.prom. | | -269 to -229 | 1a | TACTAA (4x) | | 11 |
| | | -151 to -105 | 1a | GCAGCGCTGCCGTCGccggctgaGCAGC | Adf-1 ? | 11 |
| | | - 98 to - 77 | 1a | GAGATCGCGTAACGGTAGATAA | | 11 |
| Adh1 (alcohol dehydrogenase) | maize | -190 to -186 | 4a (in vivo) | CCACG | | 12 |
| | | -145 to -138 (after ind.) | 4a (in vivo) | CCCCGG | | 12 |
| | | -120 to -117 (after ind.) | 4a (in vivo) | CGTGG | | 12 |
| | | -108 to -100 | 4a (in vivo) | CCCACAGGC | | 12 |
| ADH2 (alcohol dehydrogenase) | yeast | -257 to -216 | deletions | 22 bp dyad symmetry | ADR1 | 13 |
| albumin | rat/liver | -156 to -141 | 1a | GCAAGGGATTTAGTTA | | 14 |
| | | -126 to -107 | 1a | TTTTTGGCAAGGAT | NF-1 | 14 |
| | | -105 to - 89 | 1a | ATTTTGTAAT | | 14 |
| | | - 72 to - 35 | 1a | TGGTTAATGATCTACAGTTATTGGTTA | | 14 |
| A-MuLV (amphotrop. murine leukemia virus) | /F9, PCC4 | -87 to -59 | 1a, 3 | CCAAT | EPBF | 15 |
| aP2 (adipocyte P2) | mouse/adipocytes | -124 to -108 | 1a, 3 | AACATGACTCAGAGGAA | c-fos ? | 16 |

Wingender E. Compilation of transcription regulating proteins. Nucleic Acids Res. 1988 Mar 25;16(5):1879-902. PMID: 3282223

# The genome-wide map: TRANSFAC Site Table

## The present status

A BKL Locus Report (rat albumin)

Locus Report

**Rat Alb (Serum albumin)** Albumin, a drug and fatty acid transporter that acts in protein nitrosylation and complex formation, MAPKKK cascade, and TGF-beta1 production; human ALB is associated with familial form of hyperthyroxinemia and hypoalbuminemia

Navigator [expand]

ALB [orthogene]
- ALB [cattle, Bos taurus] [gene]
- ALB [Human] [gene]
- Alb [Mouse] [gene]
- **Alb [Rat] [gene]**
- ALB [taxonomic class Mammalia] [gene]

Albumin [orthogroup]

Gene

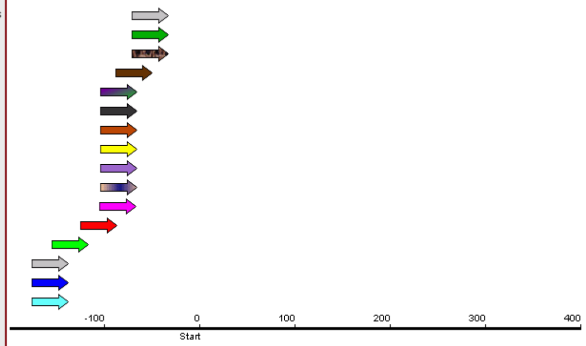
Gene Regulation [hide]

Visualization: Click icon to view in the BKL Pathfinder.

Chromosome: 14p22

Promoter: 19141231..19152231 PM000039510; albumin

Regulatory Elements




* Note: Only binding sites whose location is relative to the TSS are graphically displayed.

Binding Sites (18 interactions)

| | Identifier | Location | Binding Factor(s) | DNA Binding Reaction | Effect |
|---|---|---|---|---|---|
| | RAT$ALBU_24 | -177 to -154 | HNF-3alpha(h) | HNF-3alpha(h) --> ALB(r) | DNA binding |
| | RAT$ALBU_24 | -177 to -154 | HNF-3alpha(r) | HNF-3alpha(r) --> ALB(r) | DNA binding |
| | RAT$ALBU_24 | -177 to -154 | HNF-3beta(r) | HNF-3beta(r) --> ALB(r) | transactivation |
| | RAT$ALBU_18 | -156 to -141 | | | |
| | RAT$ALBU_25 | -152 to -132 * | HNF-3alpha(r) | HNF-3alpha(r) --> ALB(r) | DNA binding |
| | RAT$ALBU_25 | -152 to -132 * | HNF-3beta(r) | HNF-3beta(r) --> ALB(r) | transactivation |
| | RAT$ALBU_19 | -126 to -107 | NF-1(r) | NF-1(r) --> ALB(r) | DNA binding |
| | RAT$ALBU_20 | -106 to -88 | | | |
| | RAT$ALBU_21 | -105 to -89 | C/EBPalpha(m.s.) | C/EBPalpha(m.s.) --> ALB(r) | transactivation |
| | RAT$ALBU_21 | -105 to -89 | C/EBPalpha(r) | C/EBPalpha(r) --> ALB(r) | transactivation |
| | RAT$ALBU_21 | -105 to -89 | C/EBPbeta(LAP)(r) | C/EBPbeta(LAP)(r) --> ALB(r) | DNA binding |
| | RAT$ALBU_21 | -105 to -89 | C/EBPbeta(r) | C/EBPbeta(r) --> ALB(r) | transactivation |
| | RAT$ALBU_21 | -105 to -89 | C/EBPdelta(m) | C/EBPdelta(m) --> ALB(r) | DNA binding |
| | RAT$ALBU_21 | -105 to -89 | C/EBPepsilon(r) | C/EBPepsilon(r) --> ALB(r) | DNA binding |
| | RAT$ALBU_22 | -89 to -64 | ACF(r) | ACF(r) --> ALB(r) | DNA binding |
| | RAT$ALBU_23 | -72 to -35 | HNF-1alpha(r) | HNF-1alpha(r) --> ALB(r) | DNA binding |
| | RAT$ALBU_23 | -72 to -35 | HNF-1beta(r) | HNF-1beta(r) --> ALB(r) | DNA binding |
| | RAT$ALBU_23 | -72 to -35 | HNF-3alpha(r) | HNF-3alpha(r) --> ALB(r) | DNA binding |

[less]

# The genome-wide map: TRANSFAC Site Table

## The prese

Site Report

| R26428 | Site: RAT$ALBU_24 - ALB (albumin) |
|---|---|

**Binding Site Information** [hide]

| | |
|---|---|
| **Identifier** | RAT$ALBU_24 |
| **Gene** | Rat Alb (albumin) |
| **Region** | promoter |
| **Sequence** | agcttcagaTGGCAAACATACgca |
| **Export** | FASTA |
| **Sequence Type** | DNA |
| **Reference Point for Sequence Start** | |
| **Element** | Site X |
| **Element Range** | from -177 to -154 |
| **External Identifiers** | EMBL/GenBank/DDBJ: S82890; (327:350) |
| **Element Mapped to Promoter(s)** | |
| **Binding Factors** | HNF-3beta(r) Quality:1<br>HNF-3alpha(r) Quality:2<br>HNF-3alpha(h) Quality:5 |
| **Factor Source** | rec(rat-COS-7); Rat; recombinant expression: rat factor has been expressed in COS-7 cells<br>HepG2; Human; |
| **Method** | functional analysis, direct gel shift, gel shift competition |
| **Comments** | Positive regulatory element [1]; |

A Site entry

**References** [hide] **(1 entry)**

**[1]** *PMID 10024498*. Hsiang, C. H., Marten, N. W., Straus, D. S. Upstream region of rat serum albumin gene promoter contributes to promoter activity: presence of functional binding site for hepatocyte nuclear factor-3. Biochem J 338 241-9. (1999)

# The genome-wide map: TRANSFAC

## The present status

Number of Entries

| | Compilation 1988 | TRANSFAC 2009.1[a] |
|---|---|---|
| **Transcription factors** | 145 | 12,183[b] |
| **Binding sites** | 464 | 24,745[c] |
| **Factor-site links** | 361 | 33,513 |
| **Genes** | 122 | 36,317 |
| **ChIP-chip fragments** | -- | 155,306 |
| **Matrices** | -- | 885 |
| **References** | 209 | 20,072 |

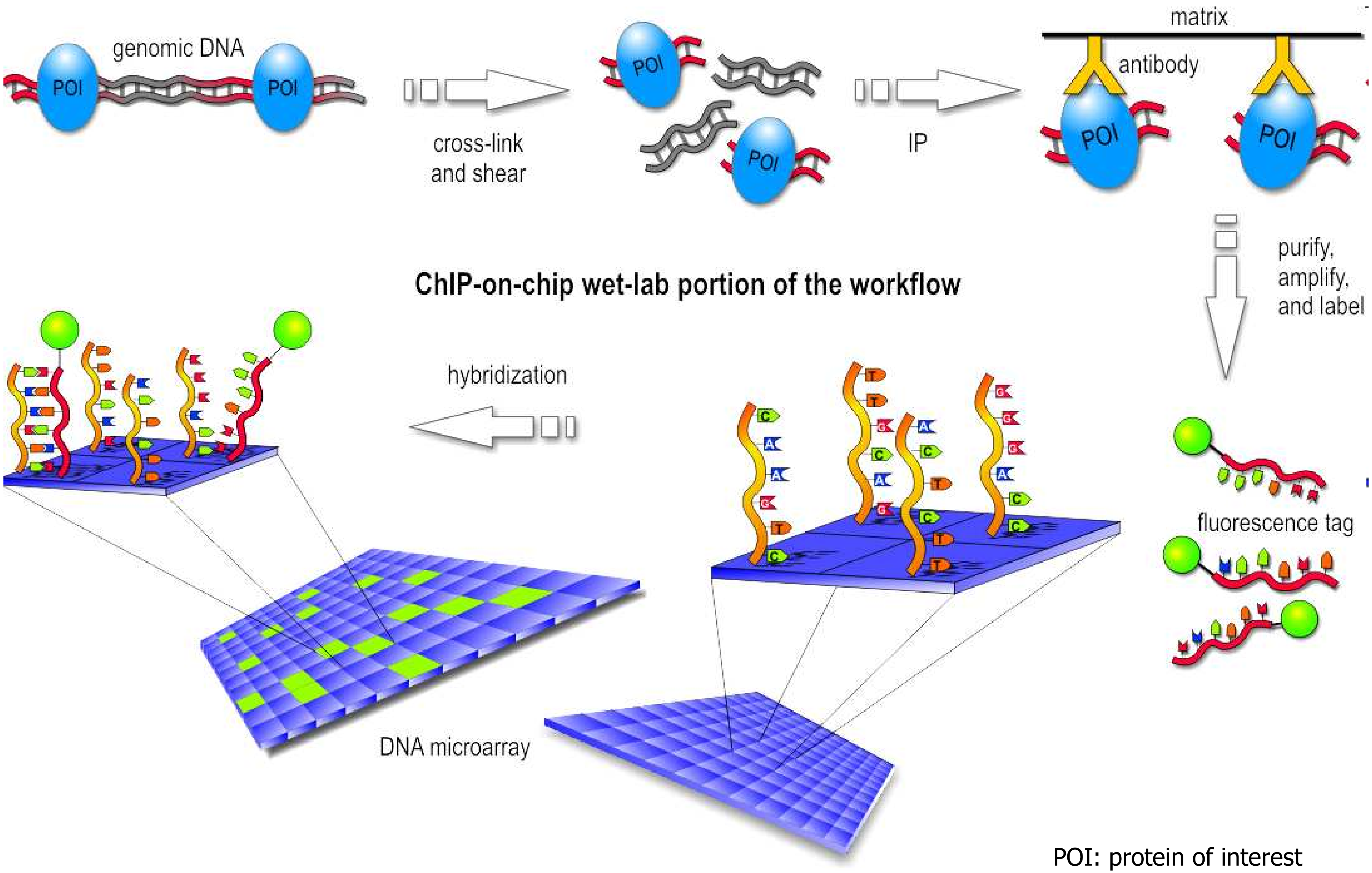
[a] as of March 31, 2009
[b] including 237 miRNAs
[c] including 577 miRNA target sites

New high-throughput technologies quickly populate the genome-wide map.

Additional 130,000 ChIP-Seq fragments coming up with the Summer release.

Wingender E., Nucleic Acids Res. 16, 1879-1902 (1988).
Matys et al., Nucleic Acids Res. 34, D108-D110 (2006).

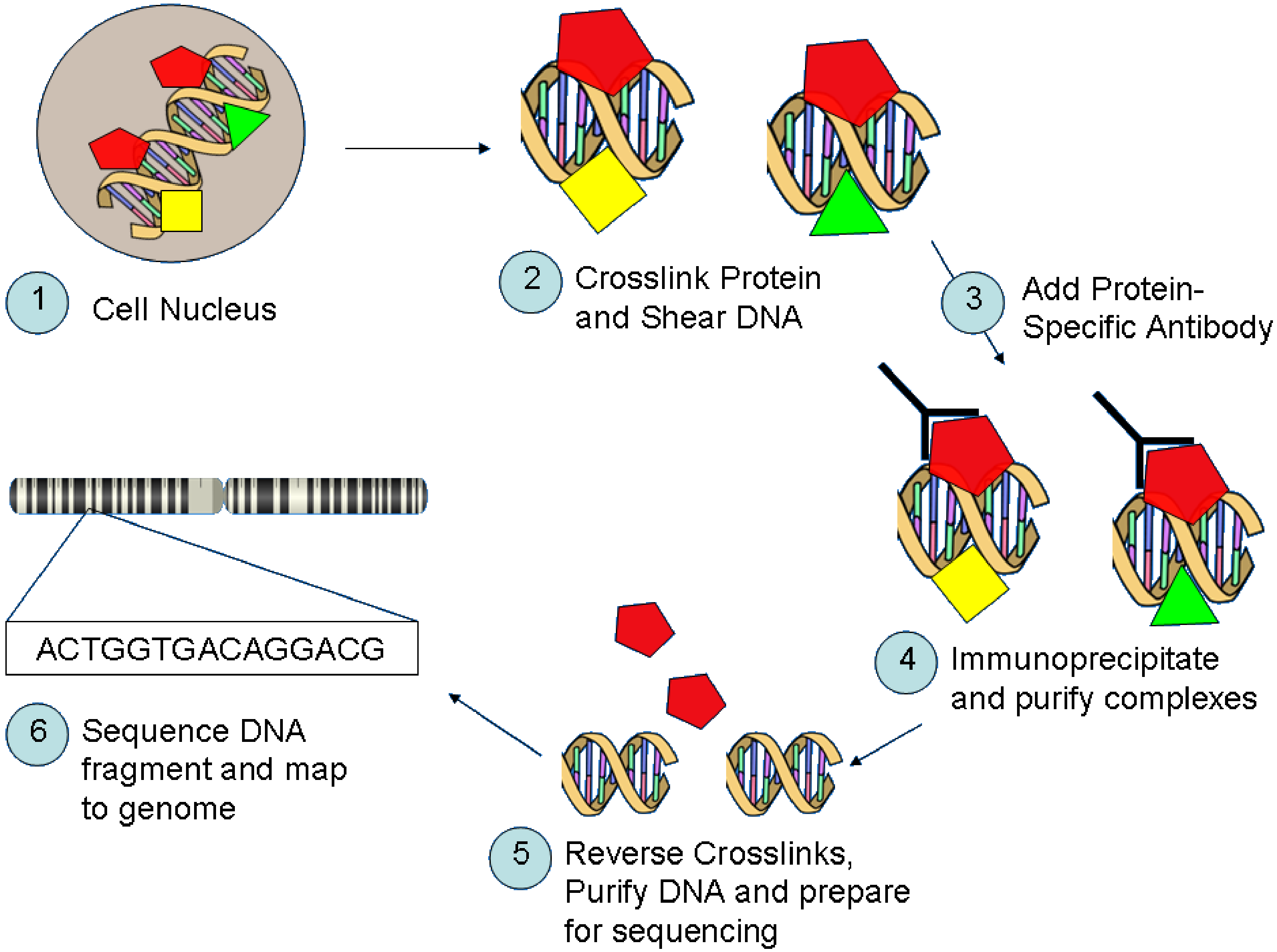
## High throughput approaches: ChIP-chip




http://en.wikipedia.org/wiki/ChIP-chip

POI: protein of interest
IP: immunoprecipitation

## High throughput approaches: ChIP-seq




http://en.wikipedia.org/wiki/ChIP-Seq

## The genome-wide map: TRANSFAC

### The present status

Additional 130,000 ChIP-Seq fragments coming up with the Summer release …

**Current data sets (TRANSFAC 2009.1)** — ChIP-chip

| | | |
|---|---|---|
| T00140 | c-Myc(h) | 751 |
| T00163 | CREB(h) | 215 |
| T02284 | CTCF(h) | 13711 |
| T00368 | HNF-1alpha-A(h) | 288 |
| T03828 | HNF-4alpha(h) | 2085 |
| T03286 | HNF-6(h) | 309 |
| T08969 | Nanog(m) | 2932 |
| T00671 | p53(h) | 586 |
| T09363 | POU5F1(m) | 1049 |
| T00594 | RelA-p65(h) | 208 |
| T00759 | Sp1(h) | 125062 |
| T00781 | TAF(II)250-isoform2(h) | 8523 |
| | sum | 155719 |

**Available with release of Sep09** — ChIP-seq

| | | |
|---|---|---|
| T10194 | (GABP-alpha(h))2:(GABP-beta(h))2 | 16489 |
| T02284 | CTCF(h) | 27119 |
| T02512 | HNF-3A(h) | 14761 |
| T06124 | REST(h) | 9607 |
| T00764 | SRF(h) | 27822 |
| T04759 | STAT1(h) | 34334 |
| | sum | 130132 |
| | **total sum** | **285851** |

**interacting fragments for 17 TFs**

UMG

## The present status

… and more coming up until the end of 2009, totaling ~790,000 genome fragments interacting with 40 different TFs

**Further upcoming datasets for human TFs:**

| | |
|---|---|
| c-Fos | 29391 |
| c-Myc | 27921 |
| JunD | 6352 |
| NF-E2 | 13325 |
| SREBP-1a | 1147 |
| Tcf4 | 52293 |
| c-Jun | 60574 |
| GATA-1 | 11097 |
| Max | 27084 |
| Pol2 | 56968 |
| SREBP2 | 1145 |
| ZNF263 | 32101 |
| Sum | 319398 |

**Further upcoming datasets for mouse TFs:**

| | |
|---|---|
| c-Myc | 3422 |
| CTCF | 39601 |
| E2F1 | 20696 |
| ESRRB | 21644 |
| Klf4 | 10873 |
| NANOG | 10343 |
| n-Myc | 7182 |
| Oct4 | 376 |
| P300 | 524 |
| SmaD1 | 1126 |
| Sox2 | 4526 |
| Stat3 | 2546 |
| Suz12 | 4215 |
| TCFCP2l1 | 26908 |
| Zfx | 10336 |
| FoxA2 | 11462 |
| P300 | 2453 |
| Sum | 181617 |

Most of the fragments come from the international ENCODE project.

## The ENCODE Project

- **ENC**yclopedia **O**f **D**NA **E**lements
- **Goal:** To compile a comprehensive parts list of functional elements in the human genome
- **Pilot Project:**
  - 2003-2007
  - Experiments focused on a limited set of genomic regions comprising about 1% of the human genome
- In September 2007, the ENCODE project was scaled up from pilot to productive phase, aiming at the coverage of the entire human genome

http://www.genome.gov/10005107

## The ENCODE Project

- Protein-coding and non-protein coding genes
  - Full-length coding sequence and variants
  - Transcriptional regulatory elements
  - All pseudogenes
- Global sequence features:
  - Methylation/CpG islands, sequence variation, evolutionary history of sequence blocks, and repetitive elements
  - Non-coding chromosomal elements:
    - Origins of replication, nuclease hypersensitive sites, matrix attachment sites, and histone modifications

http://www.genome.gov/10005107

## The ENCODE Project: experimental protocols applied

| Feature Class | Experimental Technique(s) | Abbreviations | No. Exp. Data Points |
|---|---|---|---|
| Transcription | Tiling array, Integrated annotation | TxFrag, RxFrag, GENCODE | 63,348,656 |
| 5′ Ends of transcripts | Tag sequencing | GIS-PET, CAGE | 864,964 |
| Histone modifications | Tiling array | Histone nomenclature, RFBR | 4,401,291 |
| Chromatin structure | QT-PCR, Tiling array | DHS, FAIRE | 15,318,324 |
| Sequence-specific factors | Tiling array, tag sequencing, Promoter assays | STAGE, ChIP-Chip, ChIP-PET, RFBR | 324,846,018 |
| Replication | Tiling array | TR50 | 14,735,740 |
| Computational analysis | Computational methods | CCI, RFBR Cluster | NA |
| Comparative sequence analysis | Genomic sequencing, multi-sequence alignments, computational analyses | CS | NA |
| Polymorphisms | Resequencing, copy no. variation | CNV | NA |

http://www.genome.gov/10005107

## The ENCODE Project: Status

- Pilot project results (Nature **447**, 799-816, 2007):
  - Transcription more complex than expected
  - Transcription Start Sites (TSS) more numerous than protein-coding genes
  - Regulatory information is distributed:
    - clustered across the genome, distribution near TSSs is symmetrical
  - Many distal DNaseI Hypersensitive Sites (DHSs)
  - Replication correlated with histone structure in a more detailed manner than previously known
- 9 Dec 2008 - First ENCODE whole-genome data freeze completed

http://www.genome.gov/10005107

However, even with the new HTP technologies, we will not be able to investigate the complete „promotome" for:

- all TFs (~1800 TF genes in human) in
- all cell types (~300 human cell types),
- all organs (~14000 morphologically distinguishable structures) at
- all developmental stages ($\geq$24 in human) and
- under all environmental conditions (--> ∞).

But what we can do is to learn about the rules behind the observations and to develop predictive models.

Goal #3:

To enable prediction of new transcription factor binding sites.

UMG

## Prediction-relevant contents in TRANSFAC

Number of Entries

| | Compilation 1988 | TRANSFAC 2009.1[a] |
|---|---|---|
| **Transcription factors** | 145 | 12,183[b] |
| **Binding sites** | 464 | 24,745[c] |
| **Factor-site links** | 361 | 33,513 |
| **Genes** | 122 | 36,317 |
| **ChIP-chip fragments** | -- | 155,306 |
| **Matrices** | -- | 885 |
| **References** | 209 | 20,072 |

[a] as of March 31, 2009
[b] including 237 miRNAs
[c] including 577 miRNA target sites

Sites and matrices can be used to predict TFBS.

Wingender E., Nucleic Acids Res. 16, 1879-1902 (1988).
Matys et al., Nucleic Acids Res. 34, D108-D110 (2006).

## Matrix construction

### TRANSFAC

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 3 | 3 | 1 | 11 | 0 | 0 | 0 | 0 | 2 | 2 | 0 | 0 |
| C | 6 | 2 | 1 | 0 | 12 | 0 | 0 | 0 | 7 | 3 | 2 | 4 |
| G | 1 | 7 | 3 | 1 | 0 | 12 | 0 | 12 | 3 | 6 | 8 | 4 |
| T | 2 | 0 | 7 | 0 | 0 | 0 | 12 | 0 | 0 | 1 | 2 | 4 |

```
          GCCCTACGTGCTGTCTCA
         CAGGCAACGTGCAGCCGGA G
       CAGTGCATACGTGGCTCCA
     CTTTGTGTGTACGTGCAGGAA
          GAAATACGTGCGCTTTGTGTG
      CGCGAGCGTACGTGCCTCAGG
      CCCCCTCGGACGTGACTCGGACCAC
       AGGGCCGGACGTGGGGCCCC
          GGAGTACGTGACGGAGCCCC
      ACGCTGAGTGCGTGCGGGAC
     CCCAGCCTACACGTGGGGTTC
  GGAGCCCAGCGGACGTGCGGGAA
```

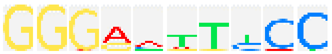
Matys et al., Nucleic Acids Res. 34, D108-D110 (2006).

# Prediction of TFBSs

| | |
|---|---|
| Accession Number | M00054 |
| Identifier | V$NFKAPPAB_01 |
| Created<br>Updated | 13.04.1995 by h<br>30.07.1996 by e |
| Copyright | Copyright (C), B |
| Name | NF-kappaB |
| Factor Description | NF-kappaB |
| Binding factors | T00587; NF-kappaB; Species: rat, Rattus norvegicus.<br>T00588; NF-kappaB; Species: mouse, Mus musculus.<br>T00590; NF-kappaB; Species: human, Homo sapiens.<br>T00593; p50; Species: human, Homo sapiens.<br>T00594; RelA-p65; Species: human, Homo sapiens. |
| Binding Matrix | (see matrix below) |

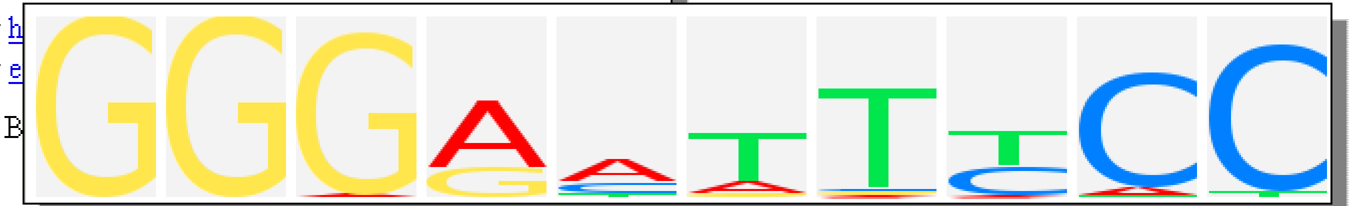
**Binding Matrix**

| A | C | G | T | Consensus |
|---|---|---|---|---|
| 0 | 0 | 40 | 0 | G |
| 0 | 0 | 40 | 0 | G |
| 1 | 0 | 39 | 0 | G |
| 27 | 0 | 13 | 0 | A |
| 21 | 13 | 1 | 5 | M |
| 8 | 1 | 3 | 28 | T |
| 1 | 2 | 2 | 35 | T |
| 2 | 18 | 0 | 20 | Y |
| 3 | 36 | 0 | 1 | C |
| 0 | 38 | 0 | 2 | C |



Click for fullsize view!

| | |
|---|---|
| Basis | 40 binding sites from 30 genes (26 cellular genes and 4 viral genomes) |



**TRANSFAC**
**Matrix Table**

Matys et al., Nucleic Acids Res. 34, D108-D110 (2006).

## TFBS detection with TRANSFAC matrices and the Match™ algorithm:

$$q = \left( \sum_{i=1}^{L} I(i) f_{i,b_i} - \sum_{i=1}^{L} I(i) f_i^{\min} \right) \Big/ \sum_{i=1}^{L} I(i) f_i^{\max}$$

with: $b_i$, nucleotide *b* found in the *i*-th position of test sequence,
$f_{bi}$, frequency of nucleotide *b* in the *i*-th position of the aligned training sequences,
$f_i^{min}$, minimum frequency in position *i*,
$f_i^{max}$, maximum frequency in position *i*,
and

$$I(i) = \sum_{B \in \{A,T,G,C\}} f_{i,B} \ln(4 f_{i,B}), \quad i = 1,2,\ldots,L$$

Kel et al., Nucleic Acids Res. 31, 3576-3579 (2003).

## Validation of potential TFBSs by comparative genome analysis:

General idea:

Genomic sites that are functionally important are under evolutionary pressure and, thus, are more conserved among related genomes than the genomic background („phylogenetic footprinting").

This could add an independent criteria to the pattern-based prediction of TFBSs.

## Validation of potential TFBSs by comparative genome analysis:

However:

What has to be conserved, the sequence or the pattern?

## Validation of potential TFBSs by comparative genome analysis:

Sequence-only conservation: *core_sim/matrix_sim*

```
GGGGAATTTCC (NF-κB consensus):        1.000 / 0.997
**  *********
GGAGAATTTCC (10/11 match, 91%):       0.713 / 0.834
**  *  ******
GGAGTATTTCC (9/11 match, 82%):        0.426 / 0.672
```

Matrix V$NFKB_Q6, M0194

Pattern-only conservation:

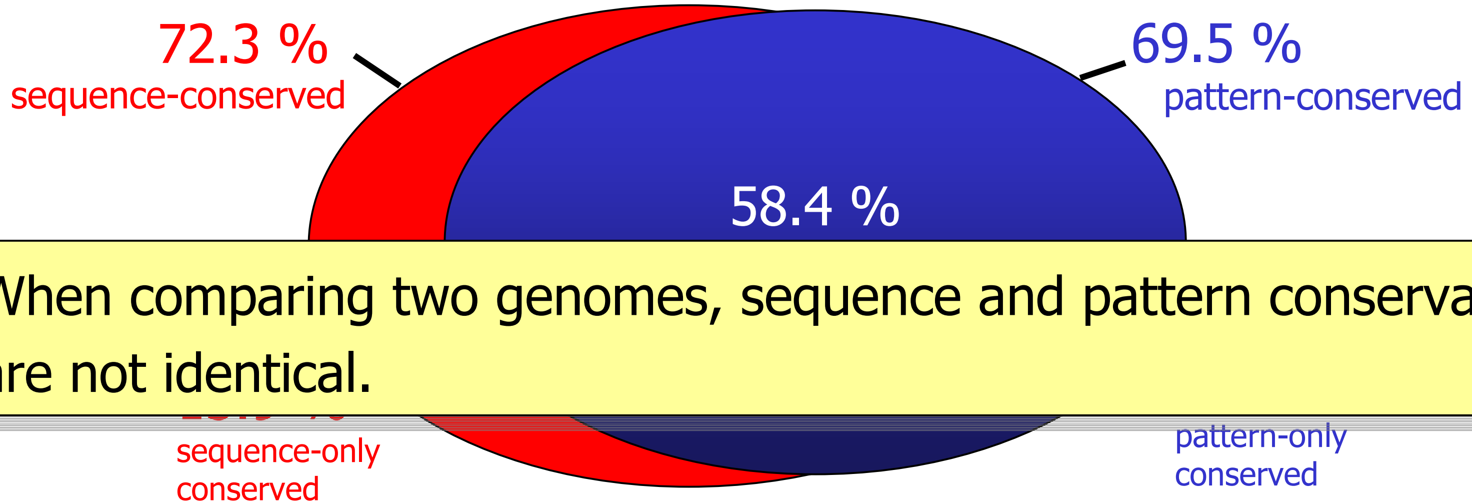
```
AATGCCTGAGGCGCT (AP-2αA):             0.991 / 0.983
   ***   ***       (6/15 match, 40%)
TTCGCCCCAGGGCGC (AP-2αA):             0.957 / 0.951
```

Matrix V$AP2ALPHA_02, M01045

## Validation of potential TFBSs by comparative genome analysis:

Human / rodent comparison for all corresponding TRANSFAC sites:

| | |
|---|---|
| General sequence conservation of TFBS: | 72.3 % |
| Background conservation in upstream sequences: | 35.2 % |
| Pattern-conserved TFBS: | 69.5 % |




When comparing two genomes, sequence and pattern conservation are not identical.

Sauer et al., Bioinformatics 22:403 (2006)

## Validation of potential TFBSs by comparative genome analysis:

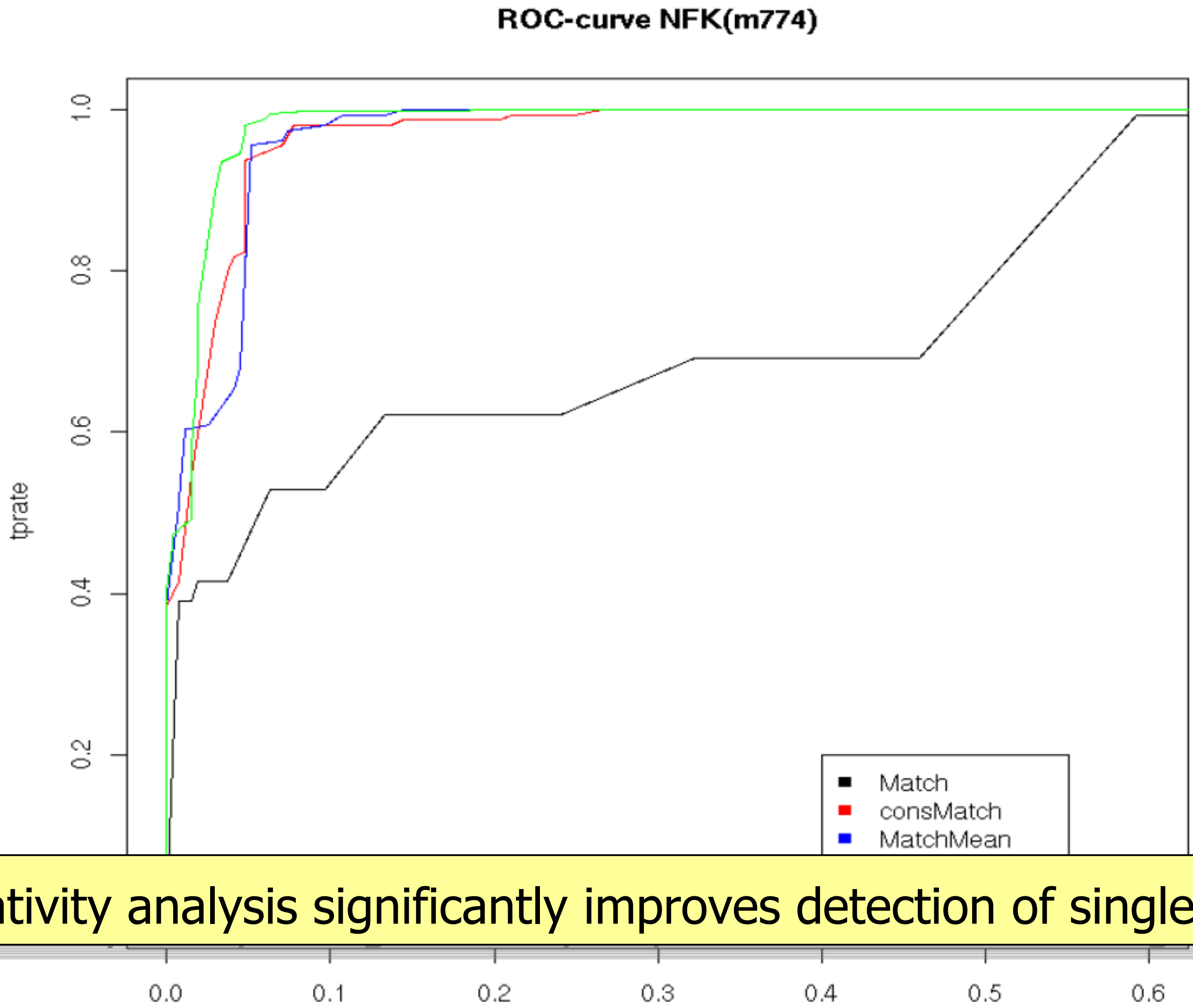
human c-*jun* gene

```
                                      R00960 (AP-1)
cow   : GCTGCG------------CTCGAGAGAGCTCCGTGAGTGACCGCGACTTT @ 953/1171
dog   : GCCGCGCCCCGAGAGCGCCCCGAGCGCGccccgtgggtgacCGCGACTTT @ 945/1153
human : GCTGCG-----------CACGAAGAGCGCTCAGTGAGTGACCGCGACTTT @ 1001/1217
mouse : GCTGCT-----------CCCCGAGAGCGCTCCGTGAGTGACCGCGACTTT @ 964/1173
rat   : GCTGCT-----------TCCCGAGAGCGCTCCGTGAGTGACCGCGACTTT @ 957/1165
      = 1051      1061      1071      1081      1091
```

```
                                      M00925 (AP-1)
cow   : T--CAAAGCCGGGCGGCGCGCG---AGCGCACAAGTAAGAGCGCGGGC @ 998/1171
dog   : TCCCCGAGGCGGGCAGCGCGCGCCAGCTGACCCAGGAGGAGCGCGG-- @ 993/1153
human : T--CAAAGCCGGGTAGCGCGCG---AGTCGACAAGTAAGAGTGCGGGA @ 1046/1217
mouse : T--CAAAGCTCGGCATCGCGGG---AGCCTACCAACGTGAGTGCTAGC @ 1009/1173
rat   : T--CAAAGCTCCGGATCGCGGG---AGCCAACCAACGTGAGTGCAAGC @ 1002/1165
      = 1101      1111      1121      1131      1141
```

Multiple genome comparison may better reveal conserved sites, and may pinpoint to „surrogate“ sites.

## Validation of potential TFBSs by comparative genome analysis:

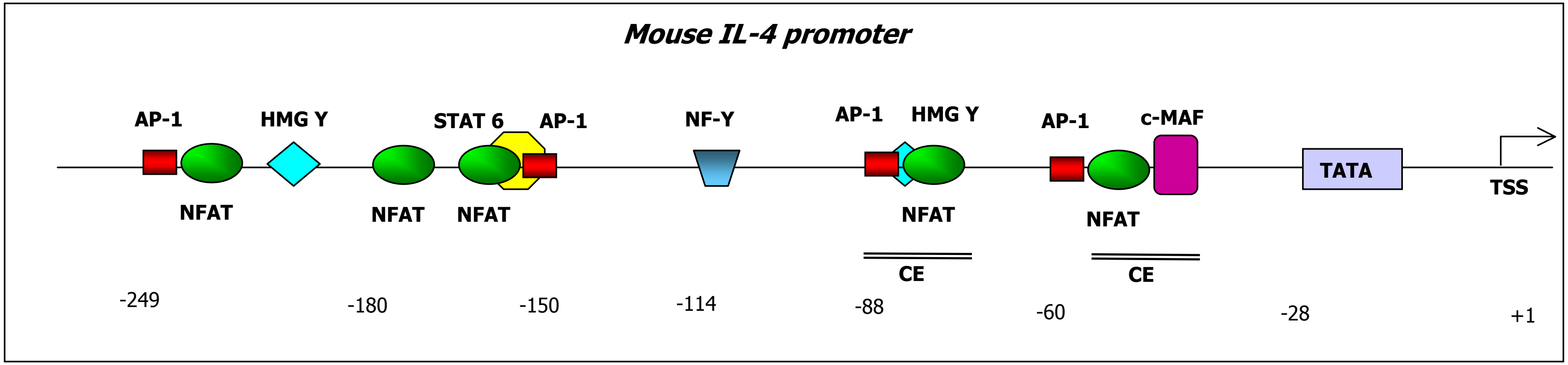


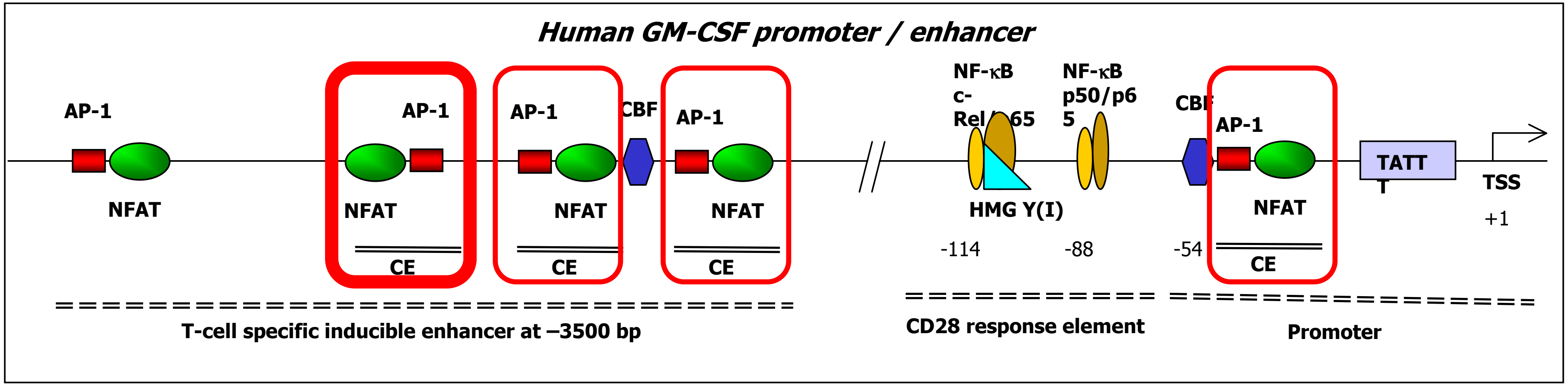

Conservativity analysis significantly improves detection of single TFBSs.

Goal #4:

To reveal the complex structure of regulatory genome regions (promoters, enhancers, etc.).

# Promoters and enhancers: examples



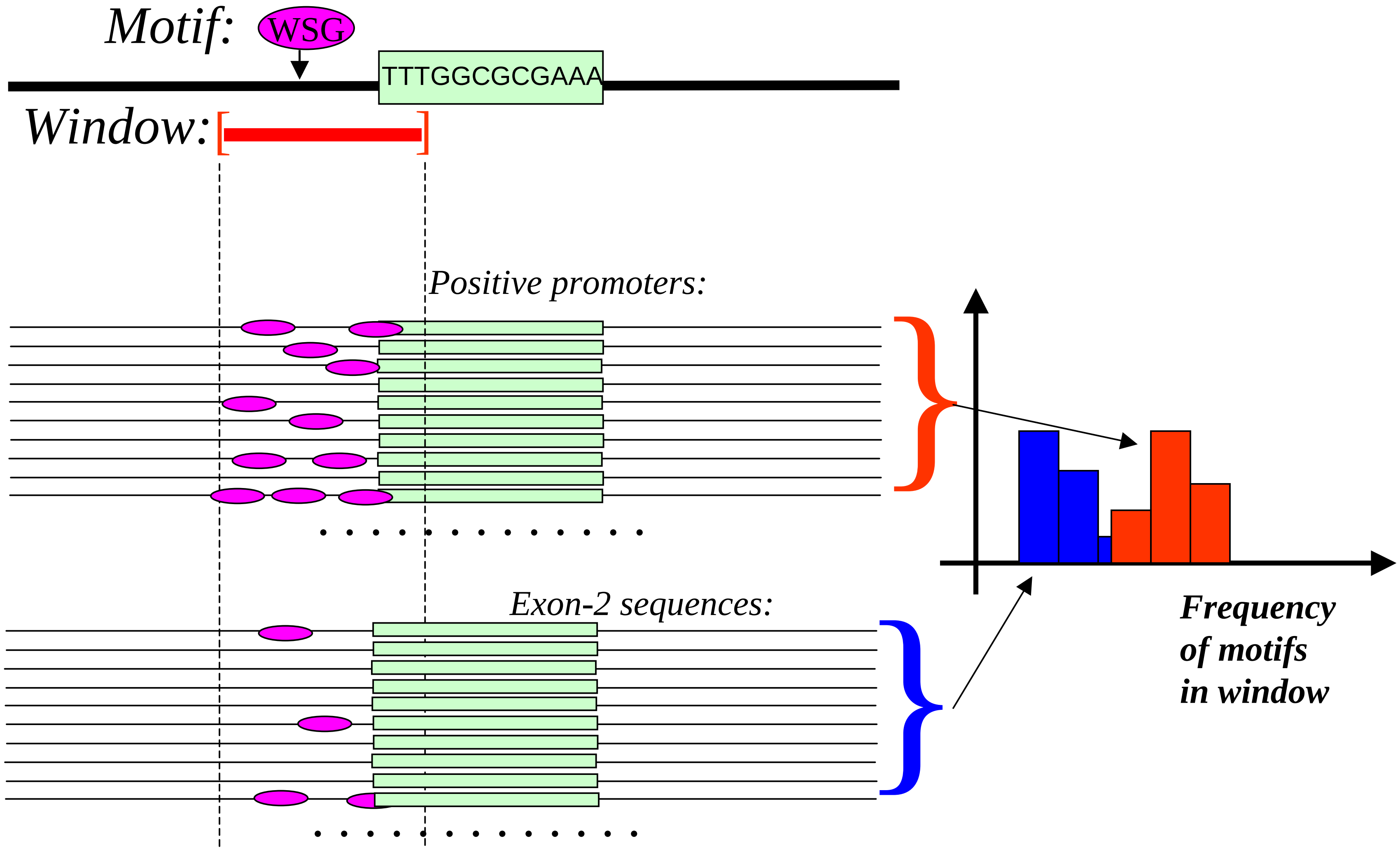



## Local context analysis

Task:

Finding nucleotide patterns around specific TFBSs which may facilitate or impede TF binding.

## Local context analysis




Kel et al., *Genome Biol.* 2008, **9:**R36.

## Local context analysis

$$f(\lambda, w, S) = \frac{N(\lambda, w, S)}{t_2 - t_1}$$

Context Score:

$$d(S) = \beta + \sum_{i=0}^{k} \alpha_i \times f(\lambda_i, w_i, S)$$

$N(\lambda,w,s)$: Number of motifs $\lambda$ in window $w=[t_1, t_2]$ of sequence $S$

| | N | Motif ($\lambda$) | Window ($w$)[1)] | $\hat{f}^{Y} / \hat{f}^{N}$ [2)] | Utility | $\alpha_i$ |
|---|---|---|---|---|---|---|
| Positive characteristics | 1 | MGCG | [27,34] | 0.0048 / 0.0041 = 1.179 | 0.80 | 0.394 |
| | 2 | TTT | [39,41] | 0.0112 / 0.0032 = 3.536 | 0.75 | 0.9618 |
| | 3 | CGSK | [17,38] | 0.0851 / 0.0341 = 2.499 | 0.90 | 0.5353 |
| | 4 | HKCG | [13,16] | 0.0675 / 0.0095 = 7.071 | 0.79 | 0.5904 |
| | 5 | VDWW | [17,46] | 0.1233 / 0.0536 = 2.299 | 0.72 | 0.223 |
| | 6 | DWTT | [21,26] | 0.0337 / 0.0000 | 0.80 | 0.5036 |
| | 7 | GSDM | [3,69] | 0.0980 / 0.0559 = 1.754 | 0.82 | 0.595 |
| Negative characteristics | 8 | VWS | [7,66] | 0.1258 / 0.1932 = 0.651 | 0.91 | -0.095 |
| | 9 | HSWY | [26,65] | 0.0413 / 0.0813 = 0.508 | 0.79 | -0.2297 |
| | 10 | VTV | [19,34] | 0.0427 / 0.1354 = 0.315 | 0.71 | -0.261 |
| | 11 | BAY | [7,65] | 0.0274 / 0.0614 = 0.447 | 0.78 | -0.566 |
| | | | | | | β=-5.6767 |

E2F binding sites start at pos. 31; Y and N: positive and negative sequence set; "utility": -1 < U < +1, calculated from the average difference, distribution overlap, normality of the distribution, *bootstrap* test, etc.

Kel et al., *Genome Biol.* 2008, **9:**R36.

## Global context analysis

Task:

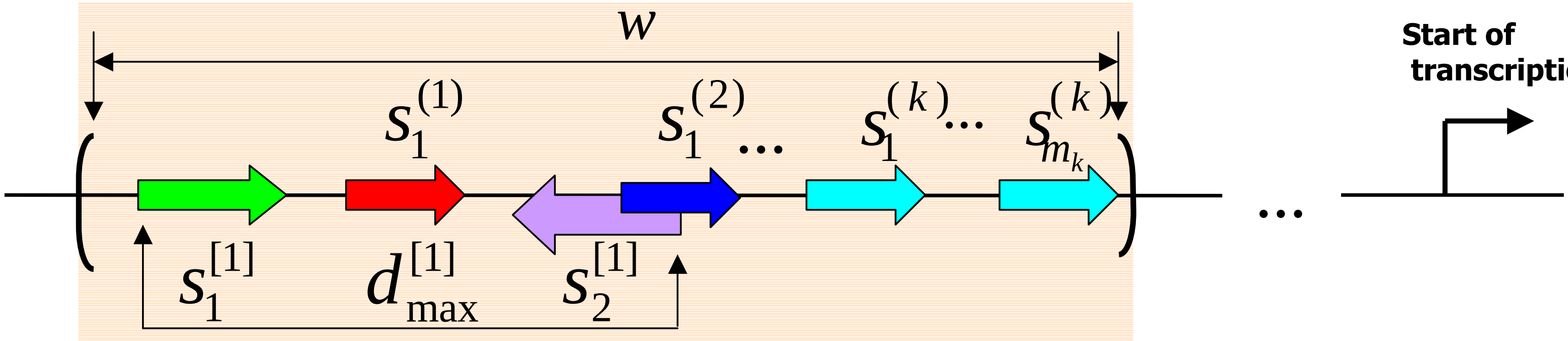
Finding (a) combination(s) of TFBS („Composite Modules") that is/are characteristic for a given set of promoters.

Kel et al., *Genome Biol.* 2008, **9:**R36.

## Composite module analyzer (CMA)

Kel et al., Bioinformatics 22, 1190-1197 (2006).

### Mathematical model (v.2)

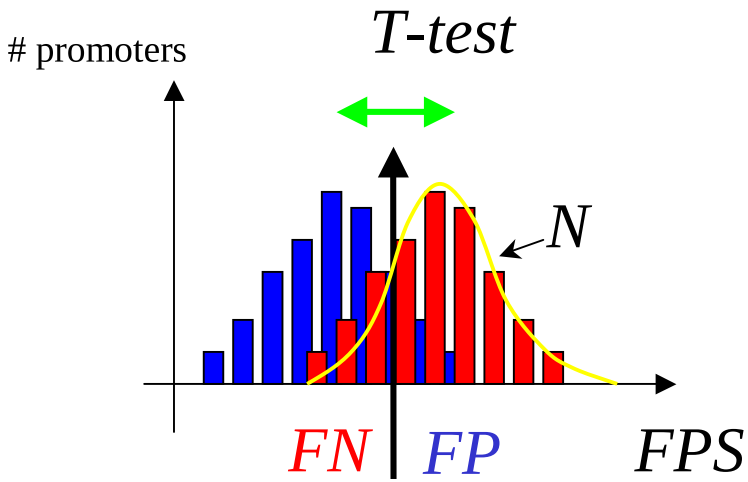


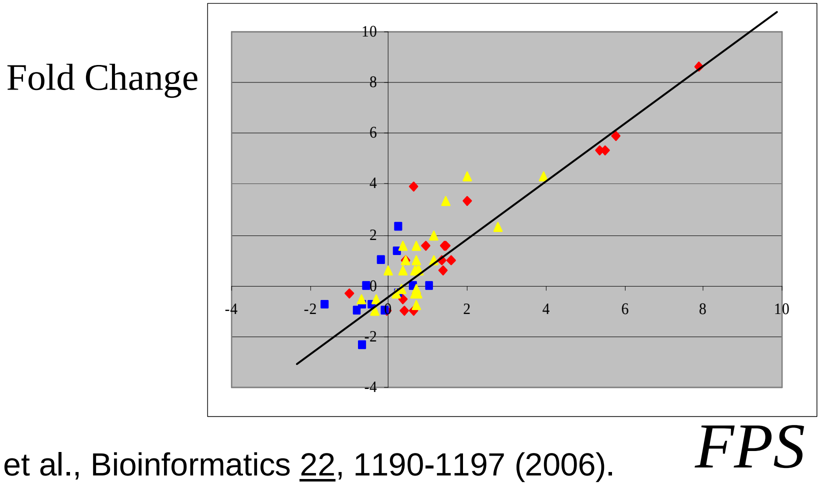

A CM contains single elements as well as composite elements (site pairs)

| | | | | |
|---|---|---|---|---|
| $d_{\max}^{[1]}$ | $d_{\max}^{[1]}$ | ... | $d_{\max}^{[R]}$ | *Parameters of the model to be estimated by a Genetic Algorithm* |
| $q_{cut-off}^{(1)}$ | $q_{cut-off}^{(2)}$ | ... | $q_{cut-off}^{(k)}$ | |
| $\phi^{(1)}$ | $\phi^{(2)}$ | ... | $\phi^{(k)}$ | |

## Fitness function of the Genetic Regression Algorithm (GRA)

$$F = \alpha \cdot R + \beta \cdot (1 - FN) + (1 - \beta) \cdot (1 - FP) + \gamma \cdot T + \delta \cdot N - \mu \cdot k$$




Kel et al., Bioinformatics 22, 1190-1197 (2006).

*R* – linear regression

*FN* – false negatives

*FP* – false positives

*T* – T-test (difference between mean values)

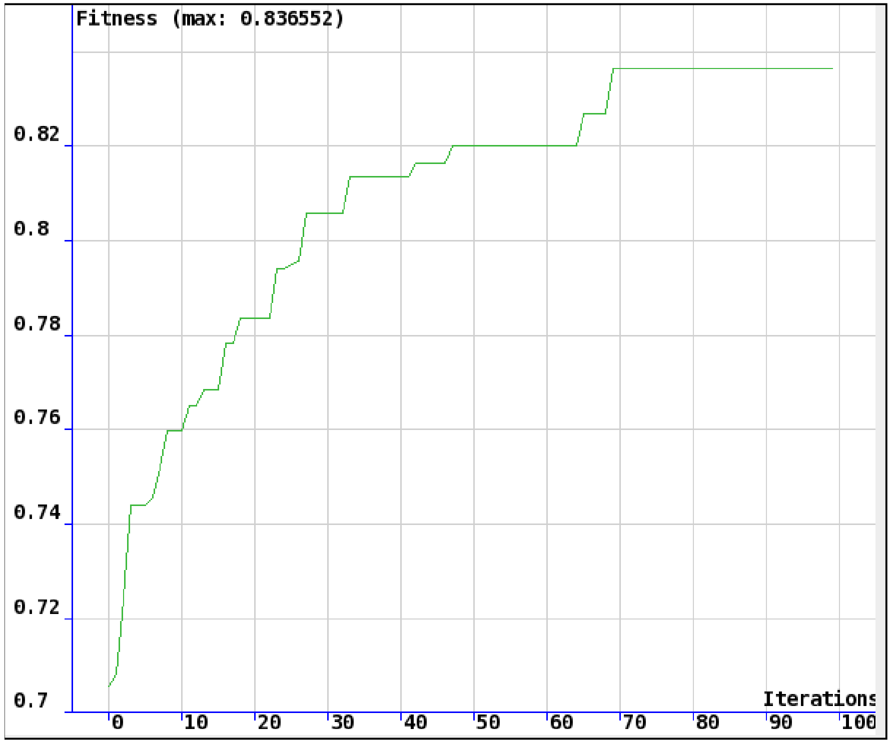
*N* – normal likeness

*k* – number of free parameters

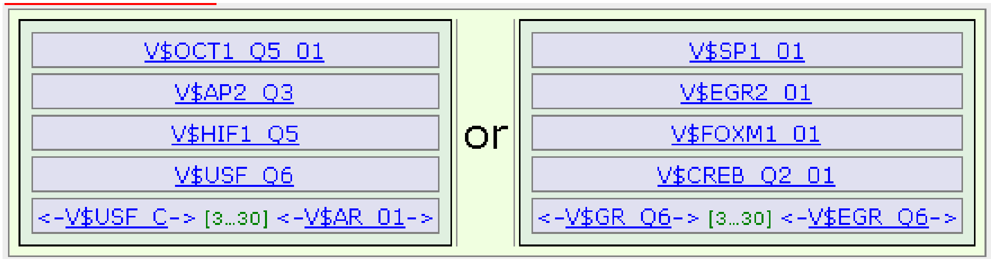
# Composite module analyzer (CMA)

## Promoter model generation

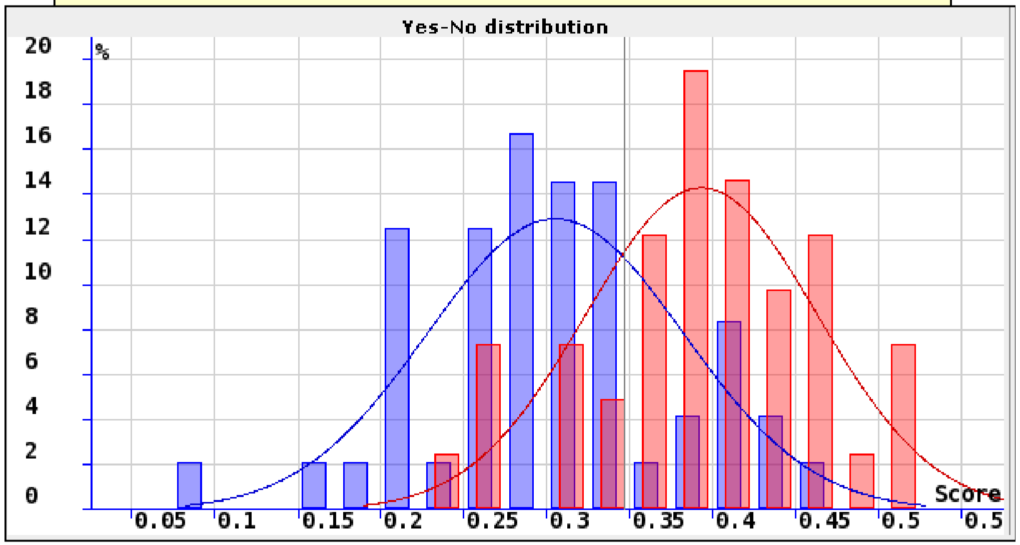
**Increase of Fitness function with number of iterations**




**Composition of the promoter model**



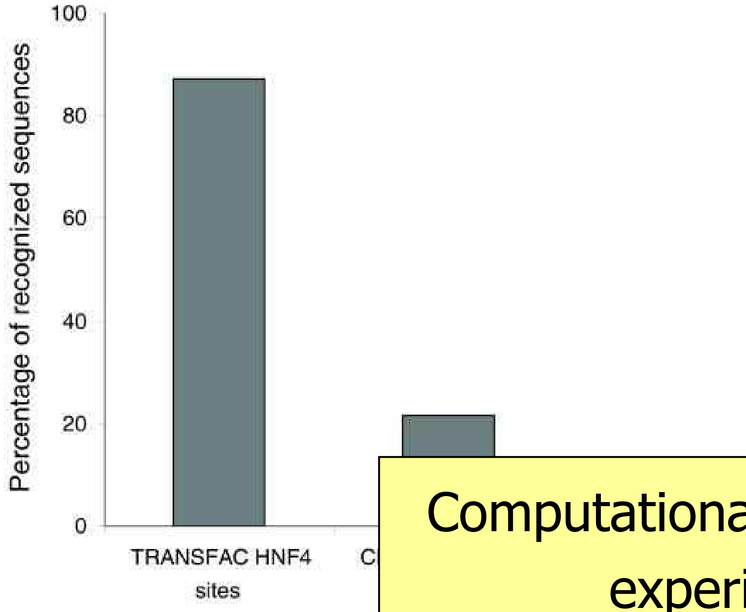

**Sequences of YES and NO sets are well separated by the selected promoter model**

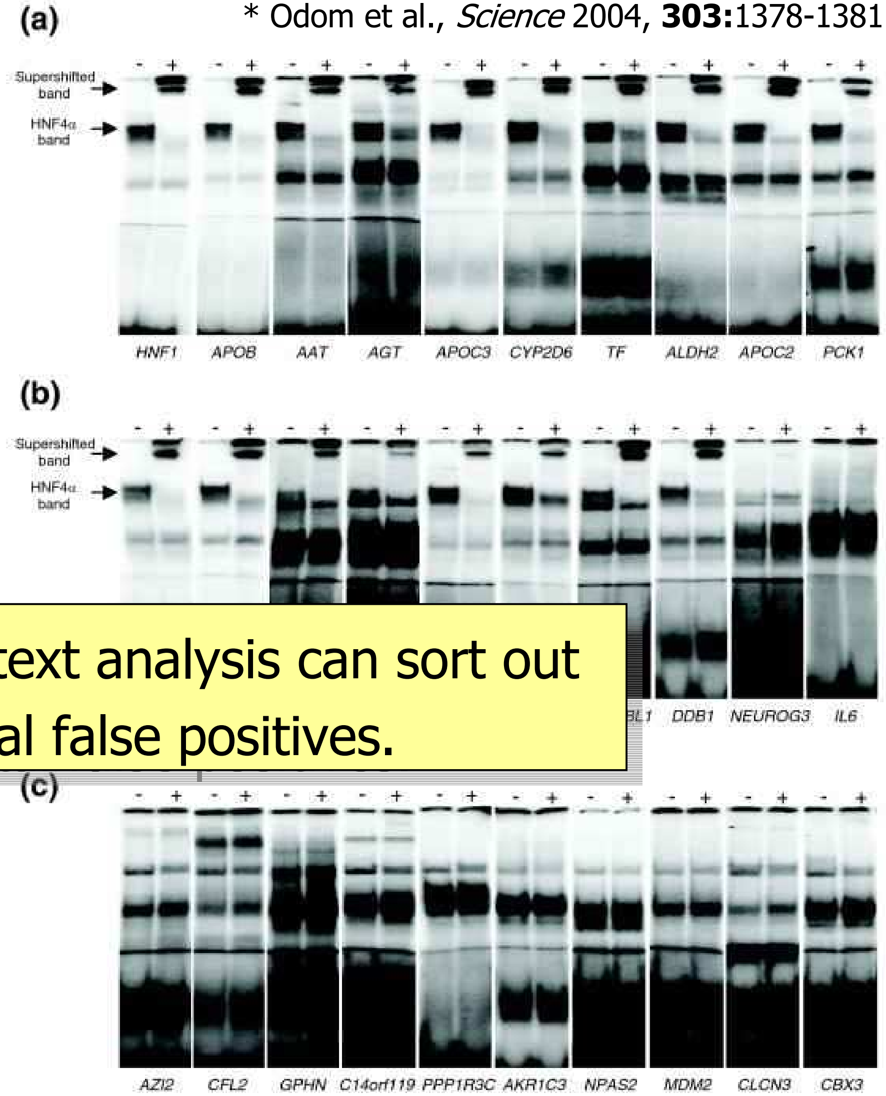


## Composite module analyzer (CMA)

HNF-4α sites of ChIP-chip fragments* revisited:

* Odom et al., *Science* 2004, **303:**1378-1381.



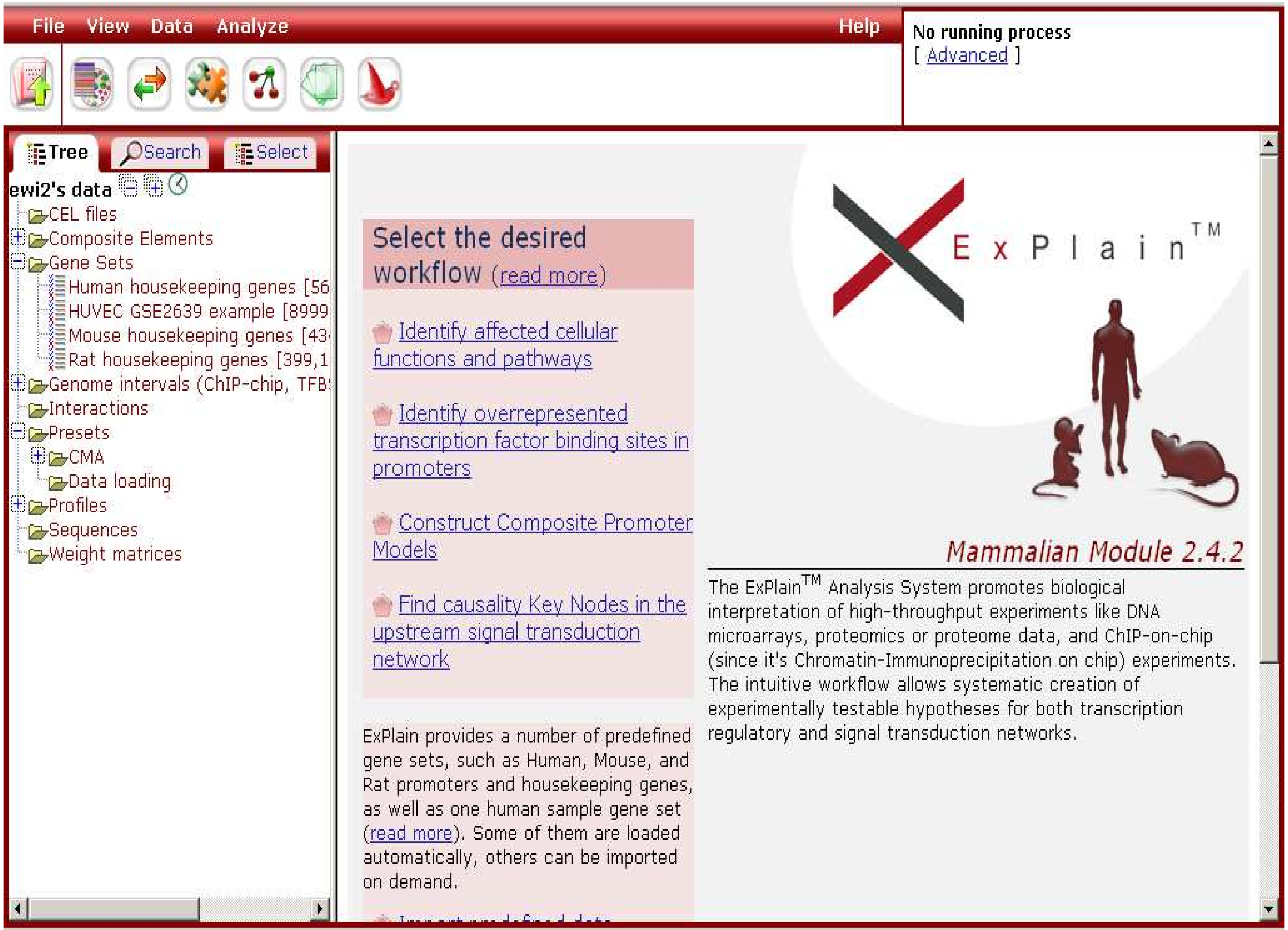




Computational context analysis can sort out experimental false positives.

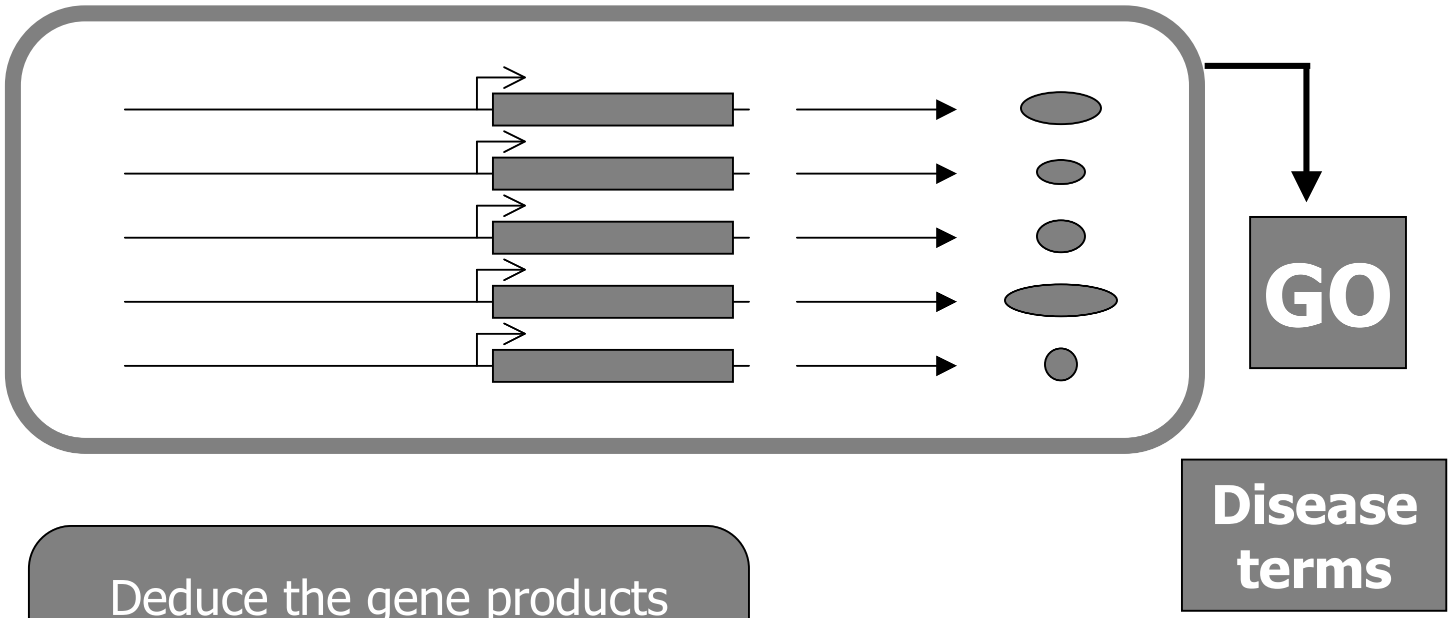
Kel et al., *Genome Biol.* 2008, **9:**R36.

# Promoter structure recognition

## The ExPlain™ System



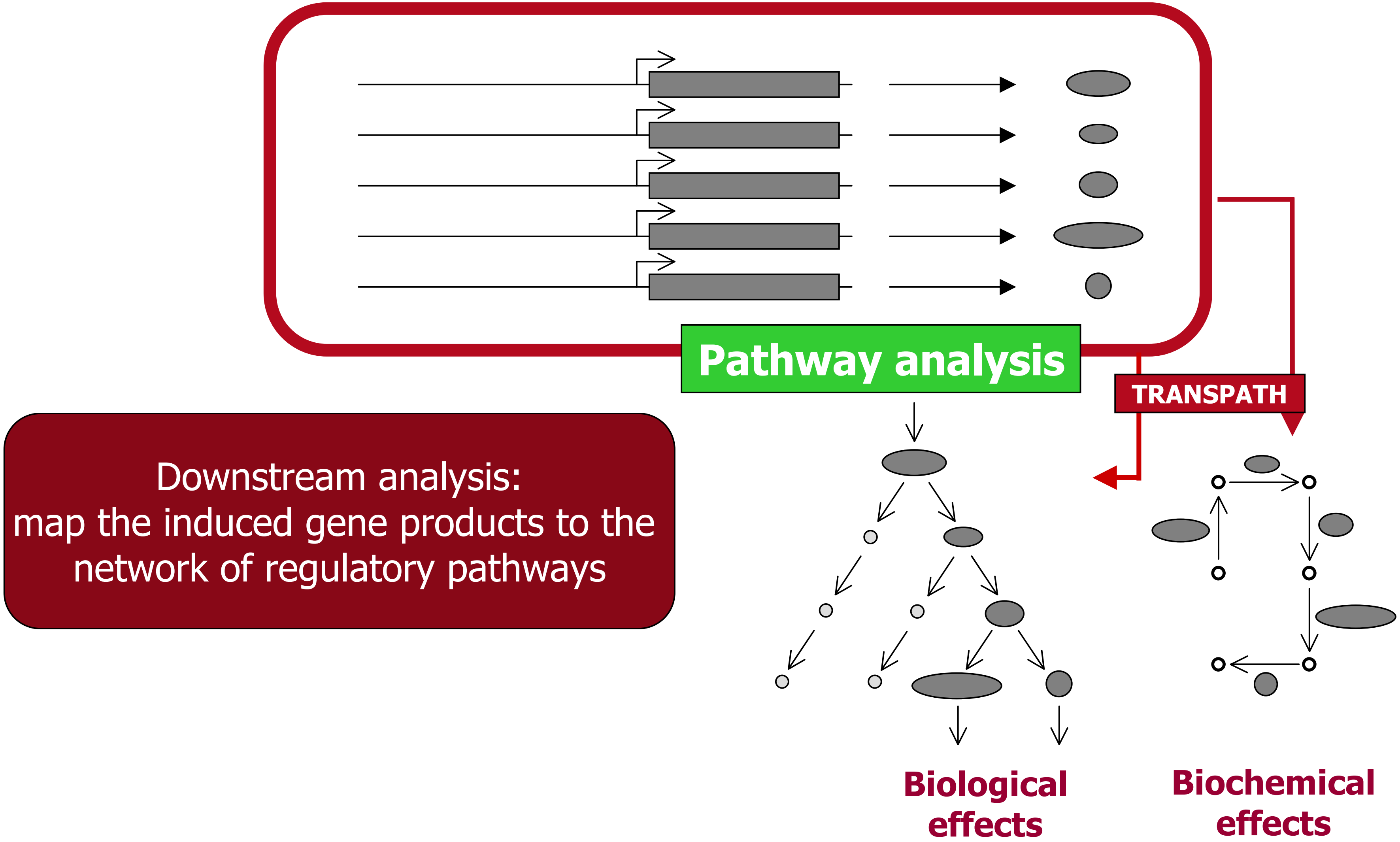
## The ExPlain™ System



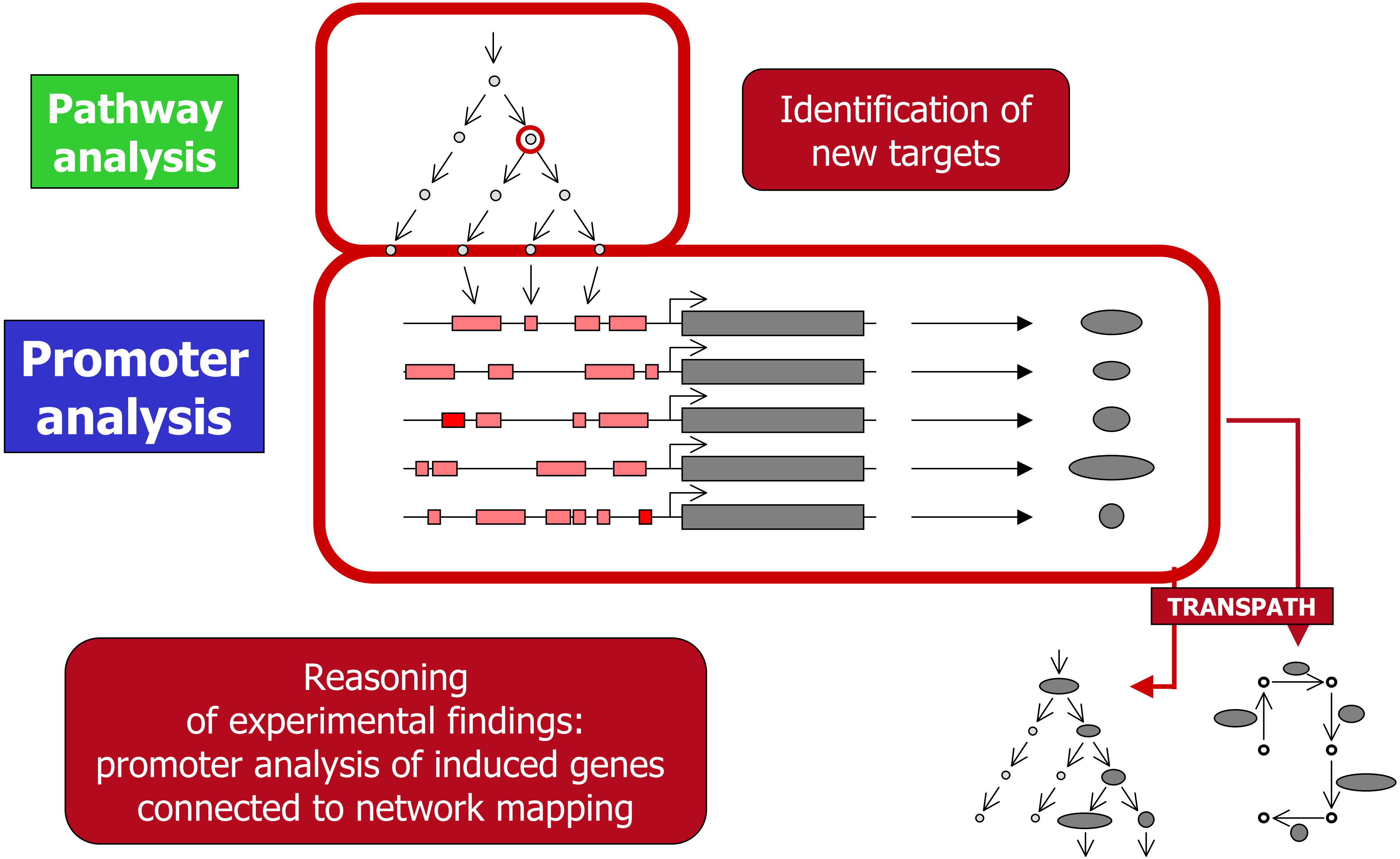

Deduce the gene products and map them to functional categories

**Functional analysis**

## The ExPlain™ System



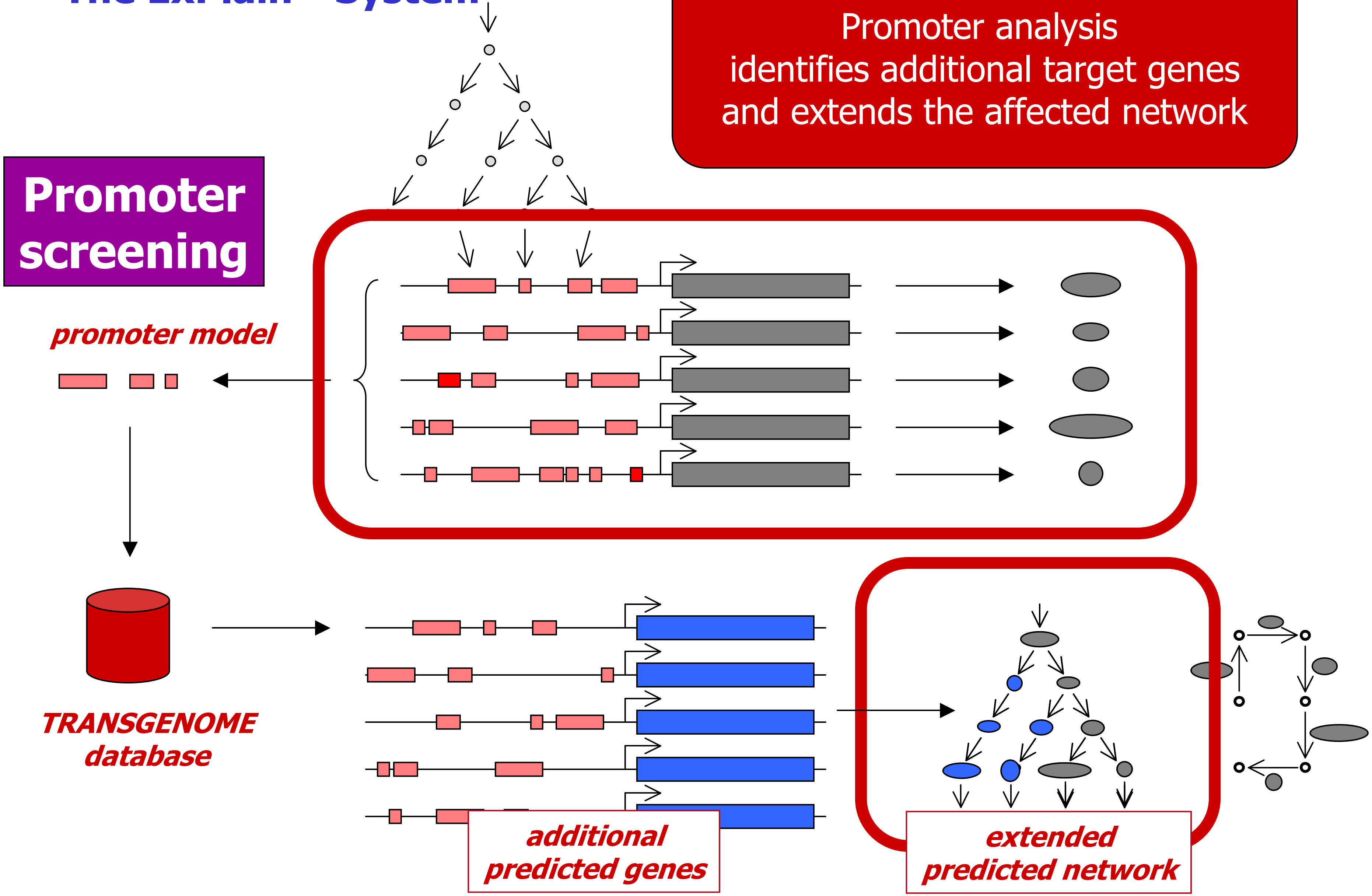
## The ExPlain™ System



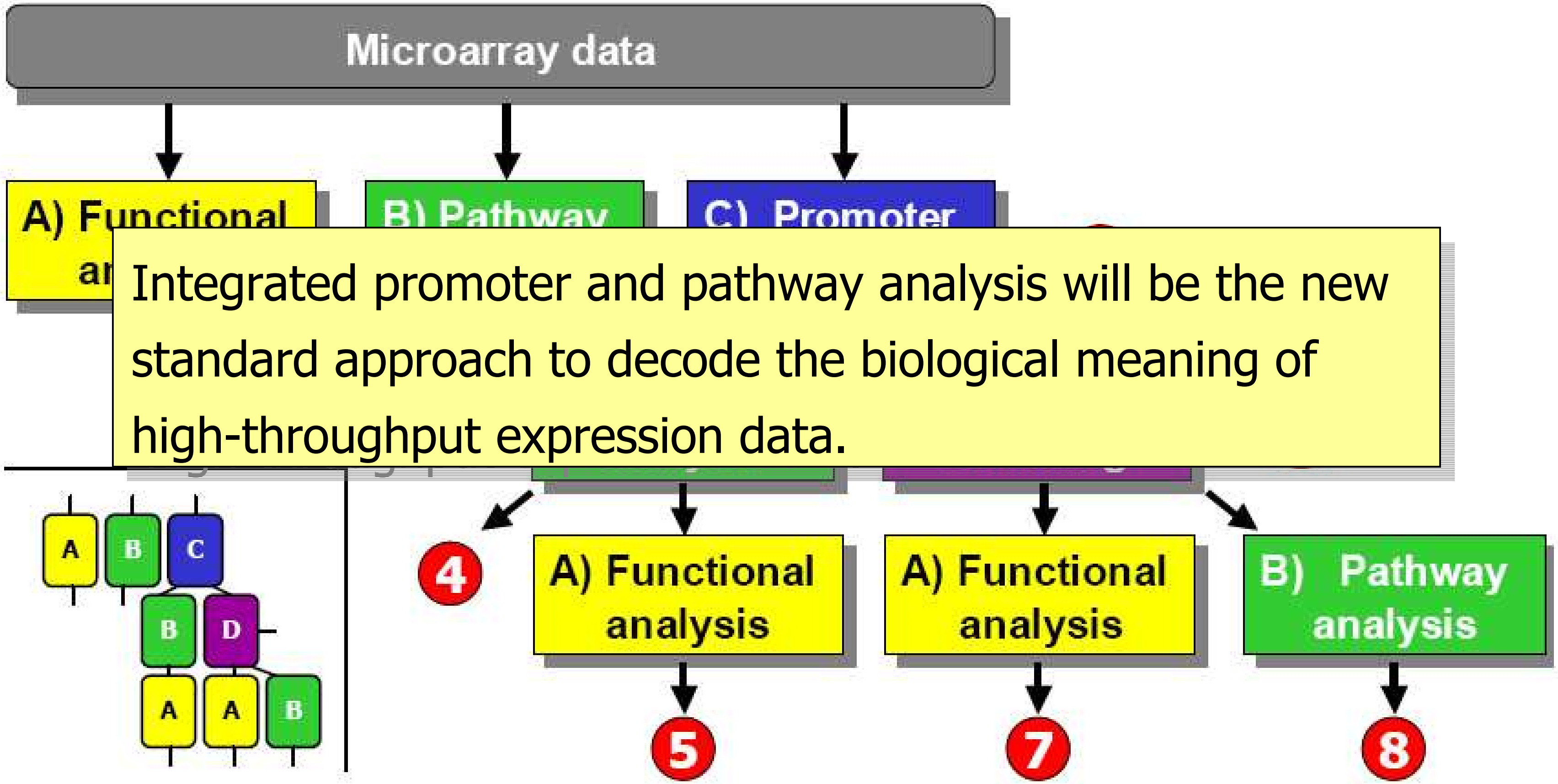
# Promoter structure recognition

## The ExPlain™ System

Promoter analysis
identifies additional target genes
and extends the affected network



## The ExPlain™ System: workflow example




Integrated promoter and pathway analysis will be the new standard approach to decode the biological meaning of high-throughput expression data.

UMG

Goal #5:

Deciphering the Protein-DNA interaction code in order to predict the DNA-binding specificity of new transcription factors.

# Protein-DNA interaction code

## The first compilation: The FACTOR table

The zinc-finger structure: The eukaryotic alternative to the bacterial helix-turn-helix motif?

| factor | source | gene | M. W. (kDa)/ finger str. | synonyms, equivalents | ref. |
|---|---|---|---|---|---|
| 60K protein | soybean | lectin | 60 | | 104 |
| α4 protein | human(HeLa) | α0 (HSV) | 2x | ICP4,Vmw175 | 84, 176 |
| | | α4 " | 170 | | 84 |
| Adf-1 | Drosophila | adh | | | 11 |
| ADR1 | yeast | ADH2 | 151 / 2 CH | | 13, 177 |
| AP-1 | human | collagenase | 47 | | 21, 110, 178 |
| | | IL-2 (?) | | | 21 |
| | | metallothionein IIA | | | |
| | | polyoma virus (?) | | | |
| | | rat stromelysin (?) | | | |
| | | SV40 enhancer | | | |
| AP-2 | HeLa | BPV | 50; 52 | KBF1 (mouse) ? | 18, 110 |
| | | GH | | | 18 |
| | | metallothionein IIA | | | 18 |
| | | c-myc | | | 18 |
| | | MHC class I H-2K$^b$ | | | 18 |
| | | SV40 enhancer | | | 18, 110, 162 |
| AP-3 | HeLa | enhancer (SV40) | | | 110, 162 |
| B factor | Drosophila | actin 5C | | TFIID (HeLa) ? | 3 |
| | " | histone H3 | | | |
| | " | histone H4 | | | |
| CBF | sea urchin | histone H2B-1 | | | 64 |
| CBP | mouse | ε-globin | | CTF ? | 51 |
| | human | hsp70 ? | | | 79 |
| | rat | LTR (MSV) | | | 125 |
| | rat | tk (HSV) | | | 125 |
| CDF | sea urchin | histone H2B-1 | | | 64 |
| CCAAT-binding factor | human | hsp70 | | CTF | 78 |
| | sea urchin | histone H2B-1 | | | 64 |
| | mouse | $\alpha_2$(I) collagen | | | 20 |
| CCAAT box bind.protein | mouse, rat, human(HeLa) | $\epsilon_A$ globin | 64 ? | CCAAT-bind.f.; CTF;CBP ? | 51 |

Wingender E. Compilation of transcription regulating proteins. Nucleic Acids Res. 1988 Mar 25;16(5):1879-902. PMID: 3282223

# Protein-DNA interaction code

## The first compilation: Zinc finger alignments

| gene, gene product[*] (ref.) | position | finger sequences |
|---|---|---|
| TFIIIA | 1 | MGEKAL |
| (210, 211) | 7 | PVVYKRYICSFADCGAAYNKNWKLQA HLCK H |
| | 38 | TGEK PFPCKEEGCEKGFTSLHHLTR HSLT H |
| | 68 | TGEK NFTCDSDGCDLRFTTKANMKK HFNRFH |
| | 99 | NIKICVYVCHFENCGKAFKKHNQLKV HQFS H |
| | 130 | TQQL PYECPHEGCDKRFSLPSRLKR HEKV H |
| | 160 | AG YPCKKDDSCSFVGKTWTLYLKHVAECH |
| | 189 | QD LAVC DVCNRKFRHKDYLRD HQKT H |
| | 215 | EKERTVYLCPRDGCDRSYTTAFNLRS HIQSFH |
| | 247 | EEQR PFVCEHAGCGKCFAMKKSLER HSVV H |
| Xfin | 103 | SAKK SHIC S HTGKLFSCTAAVVRHQRM H |
| (212) | 131 | QLQK SHHC PH CKKSFVQRSDFIK HQRT H |
| | 159 | TGER PYQC VE CQKKFTERSALVN HQRT H |
| | 187 | TGER PYTC LD CQKTFNQRSALTK HRRT H |
| | 215 | TGER PYRC SV CSKSFIQNSDLVK HLRT H |
| | 243 | TGEK PYEC PL CVKRFAESSALMK HKRT H |
| | 271 | STHR PFRC SE CSRSFTHNSDLTA HMRK H |
| | 299 | TEFR ... |
| | 320 | NVASSPYSC SK CRKTFKRWKSFLN HQQT H |
| | 349 | SREK PYLC SH CNKGFIQNSDLVK HFRT H |
| | 377 | TGER PYQC AE CHKGFIQKSDLVK HLRT H |
| | 405 | TGEK PFKC SH CDKKFTERSALAK HQRT H |
| | 433 | TGEK PYKC SD CGKEFTQRSNLIL HQRI H |
| | [illegible] | [illegible] |

Wingender E. Compilation of transcription regulating proteins. Nucleic Acids Res. 1988 Mar 25;16(5):1879-902.
PMID: 3282223

## The helix-turn-helix motif as paradigm of prokaryotic DNA-binding domains

Sauer RT, Yocum RR, Doolittle RF, Lewis M, Pabo CO (1982) Nature 298, 447-451:

Homology among DNA-binding proteins suggests use of a conserved super-secondary structure.

The amino acid sequences of the repressor and cro proteins of phages lambda, 434 and P22 are homologous, especially in a region in which repressor and lambda cro have a similar **alpha-helix-turn-alpha-helix secondary structure**. Model-building studies indicate that this structure is important in DNA binding, and we suggest it may be a common feature of many DNa-binding proteins.

PMID: 6400820

## The helix-turn-helix motif as paradigm of prokaryotic DNA-binding domains

PDB entry
1CGP

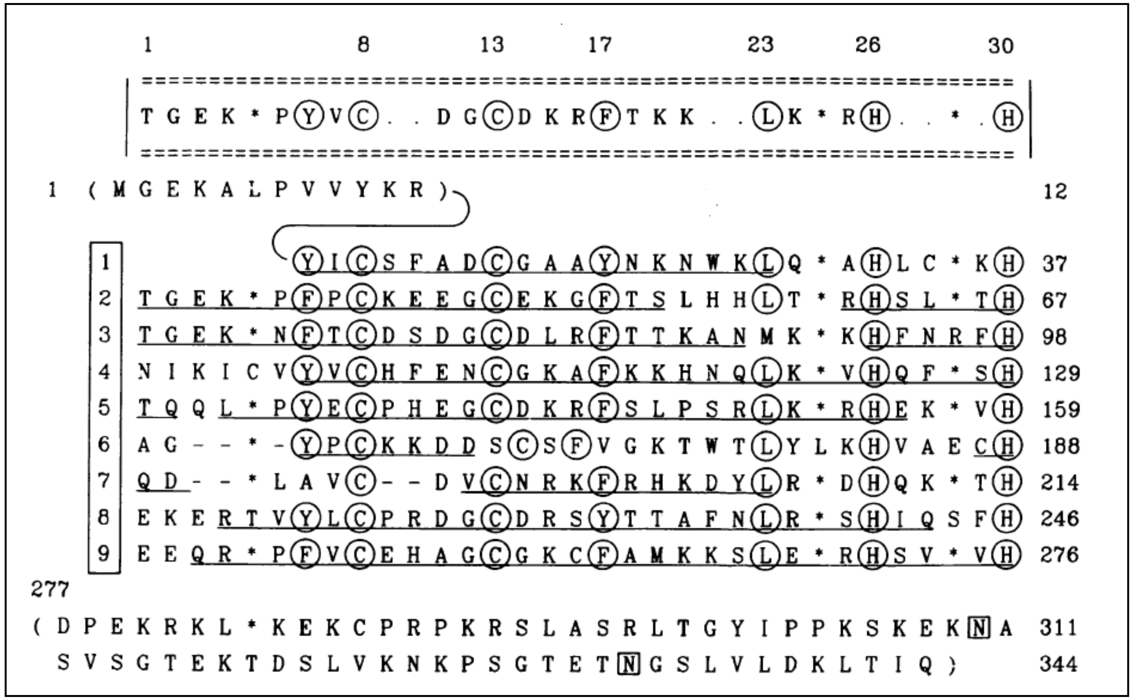
Schultz, S.C., Shields, G.C., Steitz, T.A. (1991) Crystal structure of a CAP-DNA complex: the DNA is bent by 90 degrees. Science **253:** 1001-1007

*dating back to:*

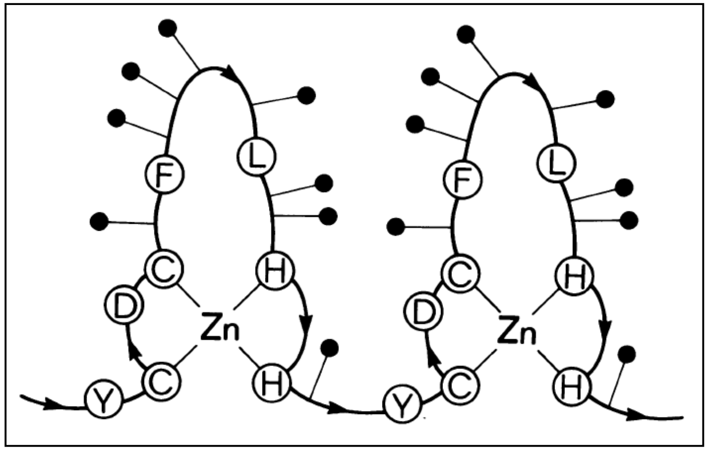
McKay, D.B., Steitz, T.A. (1981) Structure of catabolite gene activator protein at 2.9 A resolution suggests binding to left-handed B-DNA. Nature **290**:744-749.

## Protein-DNA interaction code

### The zinc finger motif as general paradigm for eukaryotic DNA-binding domains?



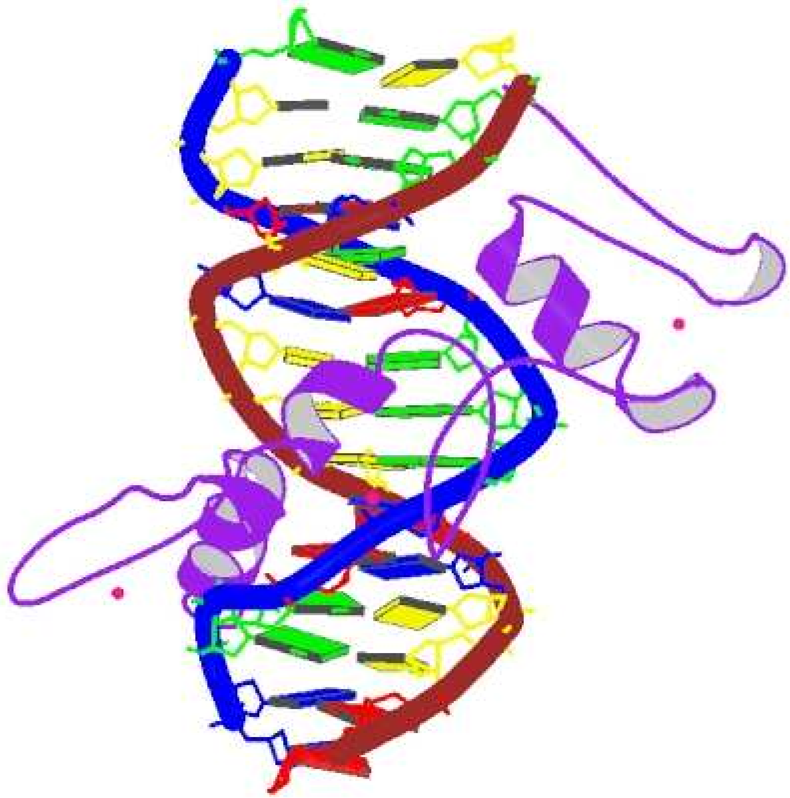

9 Consecutive $Cys_2$-$His_2$ motifs in TFIIIA identified, with a zinc finger structure proposed.




Figs. 3 and 4 in:
Miller J, McLachlan AD, Klug A. (1985) Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. EMBO J. **4**:1609-1614.
PMID: 4040853

## The zinc finger motif as general paradigm for eukaryotic DNA-binding domains?

PDB entry
1TF3

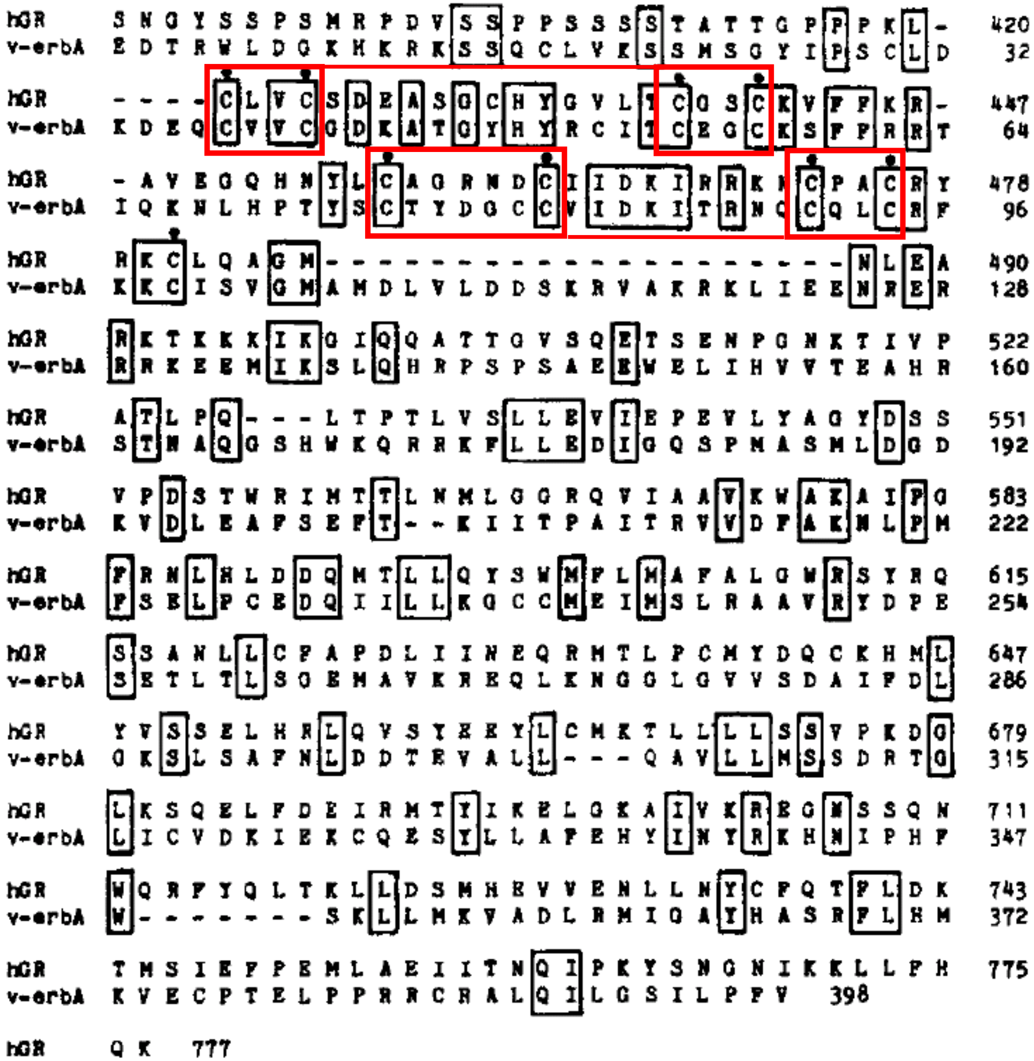


Foster, M.P., Wuttke, D.S., Radhakrishnan, I., Case, D.A., Gottesfeld, J.M., Wright, P.E. (1997) Domain packing and dynamics in the DNA complex of the N-terminal zinc fingers of TFIIIA. Nat.Struct.Biol. **4:** 605-608

## The zinc finger motif as general paradigm for eukaryotic DNA-binding domains?

### GR zinc finger

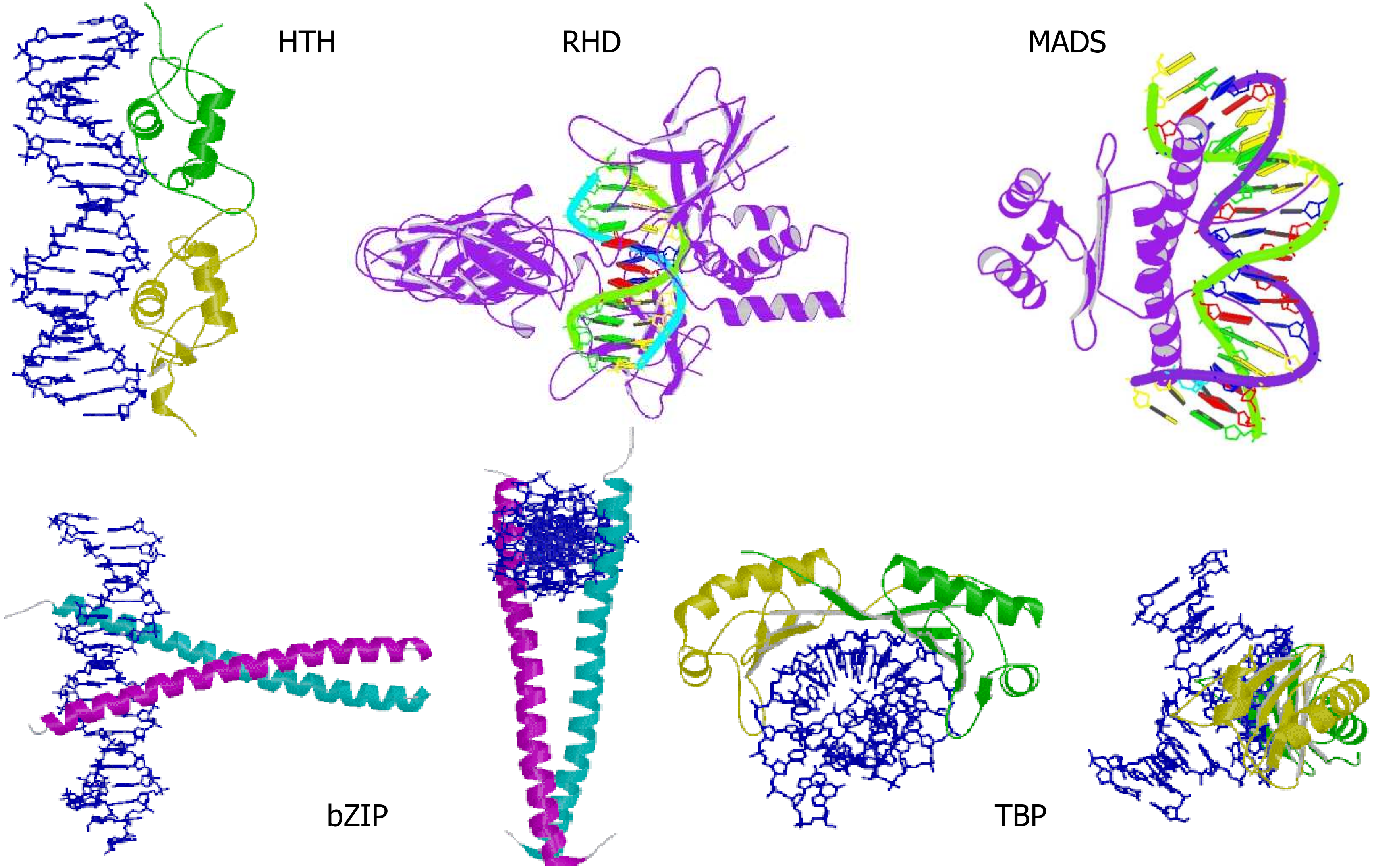
Similarity with TFIIIA zinc finger structure suggested for the first time.



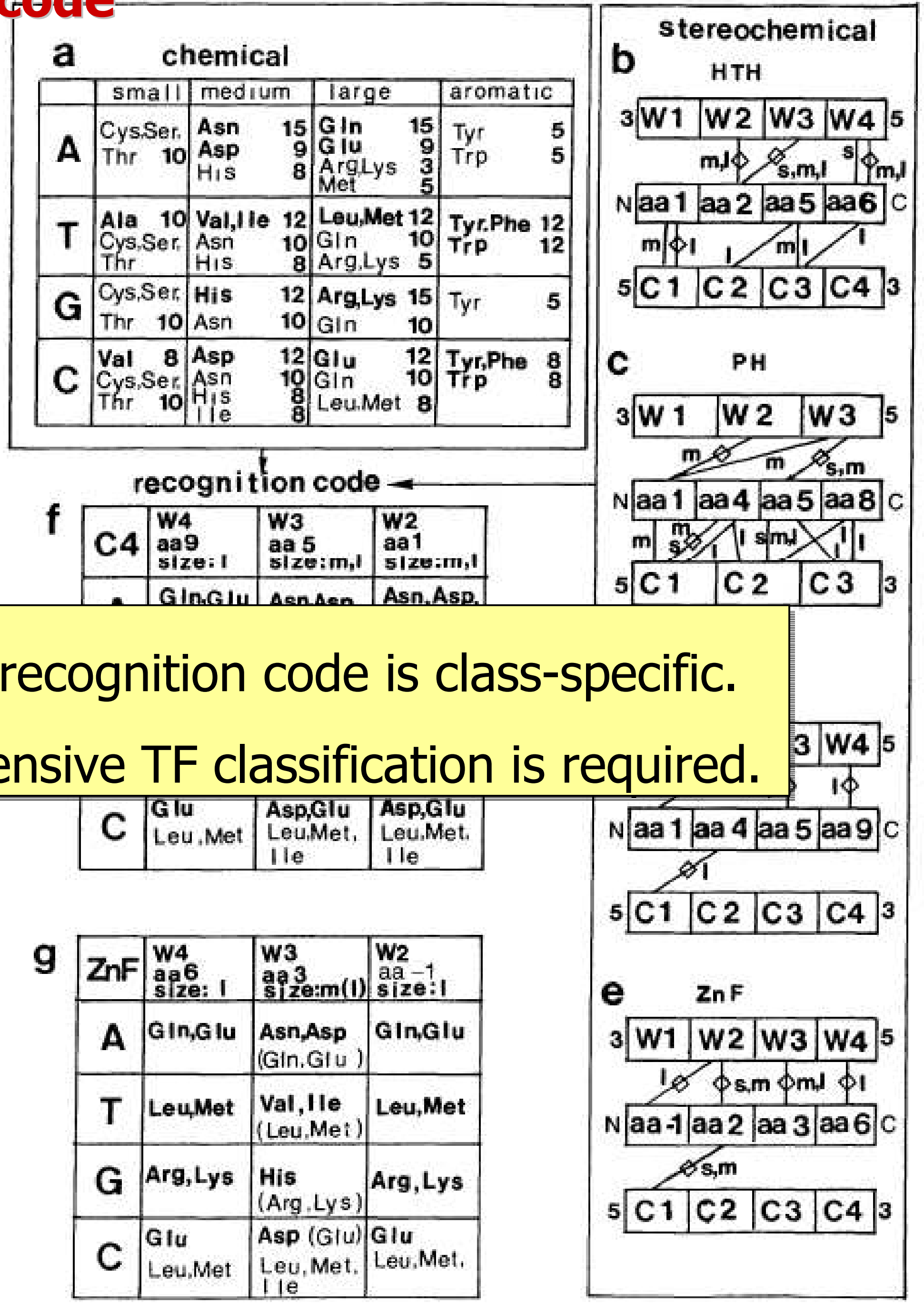
Weinberger C, Hollenberg SM, Ong ES, H
Identification of human glucocorticoid receptor complementary DNA clones by epitope selection. Science. 1985 May 10;228(4700):740-2.
PMID: 2581314

# Protein-DNA interaction code

## Many more structural classes of eukaryotic DNA-binding domains found later on



## Protein-DNA interaction code




### Protein-DNA recognition code?

The protein-DNA recognition code is class-specific.

Thus, a comprehensive TF classification is required.

Suzuki & Yagi, PNAS 91:12357-12361, 1994

UMG

## Transcription factor classification

**Table 1. Classification scheme for eukaryotic transcription factors**

| Level | Pattern of classification no. | group designation | description | example |
|---|---|---|---|---|
| 1 | N | superclass | general topology of DBD | zinc-coordinating domains |
| 2 | N.N | class | structural blueprint of the DBD | zinc-finger nuclear receptors |
| 3 | N.N.N | family | functional criteria such as protein-DNA-complex formation (DNA-binding specificity, multimerization behaviour) or biological effect | $T_3R$/RAR (in contrast to steroid hormone receptors) |
| 4 | N.N.N.N | subfamily | mainly according to sequence similarity of the DBDs | RAR (retinoic acid receptor) |
| 5 | N.N.N.N.N | genus | according to factor gene | RAR-$\alpha$, RAR-$\beta$ |
| 6 | N.N.N.N.N.N | factor "species" | initiation/ splice/ processing variants | RAR-$\alpha$1, RAR-$\alpha$2 |

## Protein-DNA interaction code

### Transcription f

| | |
|---|---|
| 1 | Basic Domains |
| 1.1 | Leucine zipper factors (bZIP) |
| 1.2 | Helix-loop-helix factors (bHLH) |
| 1.3 | Helix-loop-helix / leucine zipper factors (bHLH-ZIP) |
| 1.6 | bHSH |
| 2 | Zinc-coordinating DNA-binding domains |
| 2.1 | Cys4 zinc finger of nuclear receptor type |
| 2.2 | diverse Cys4 zinc fingers |
| 2.3 | Cys2His2 zinc finger domain |
| 2.4 | Cys6 cysteine-zinc cluster |
| 2.5 | Zinc fingers of alternating composition |

DNA-binding domains of eukaryotic TFs can be classified in (at least) 4 superclasses and 30 classes.

| | |
|---|---|
| 4 | beta-Scaffold Factors with Minor Groove Contacts |
| 4.1 | RHR (Rel homology region) |
| 4.2 | STAT |

More recent data, however, suggest that there may be up to 13 superclasses (mostly spreading out from previous superclass 4).

| | |
|---|---|
| 0 | Other Transcription Factors |
| 0.1 | Copper fist proteins |
| 0.2 | HMGI(Y) |
| 0.3 | Pocket domain |
| 0.4 | E1A-like factors |
| 0.5 | AP2/EREBP-related factors |

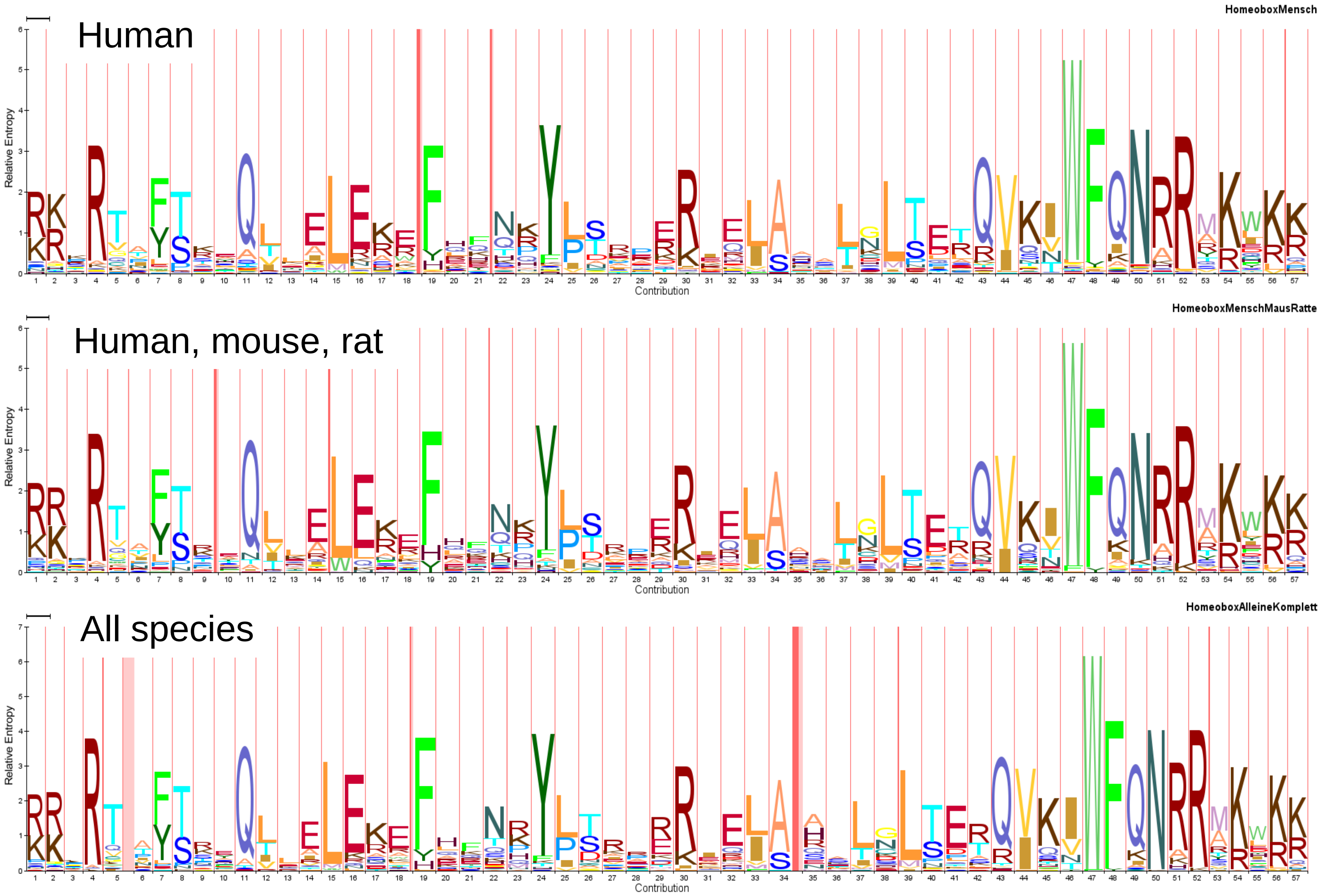
UMG

## Transcription factor classification

| | |
|---|---|
| 1 | Basic Domains |
| 2 | Zinc-coordinating DNA-binding domains |
| 2.1 | Cys4 zinc finger of nuclear receptor type |
| 2.1.1 | Steroid hormone receptors (NR3) |
| 2.1.1.1 | Corticoid receptors (NR3C) |
| 2.1.1.1.1 | **GR (NR3C1):** GR GR GR GR GR GR GR |
| 2.1.1.1.1.1 | **GR-alpha :** GR-alpha |
| 2.1.1.1.1.2 | **GR-beta:** GR-beta |
| 2.1.1.1.2 | **MR (NR3C2):** MR MR |
| 2.1.1.2 | Progesterone receptor (NR3C) |
| 2.1.1.3 | Androgen receptor (NR3C) |
| 2.1.1.4 | Estrogen receptor (NR3A) |
| 2.1.1.5 | Estrogen related receptor (NR3B) |
| 2.1.2 | Thyroid hormone receptor-like factors (NR0, NR1, NR2, NR4, NR5) |
| 2.2 | diverse Cys4 zinc fingers |
| 2.3 | Cys2His2 zinc finger domain |
| 2.4 | Cys6 cysteine-zinc cluster |
| 2.5 | Zinc fingers of alternating composition |
| 3 | Helix-turn-helix |
| 4 | beta-Scaffold Factors with Minor Groove Contacts |
| 0 | Other Transcription Factors |

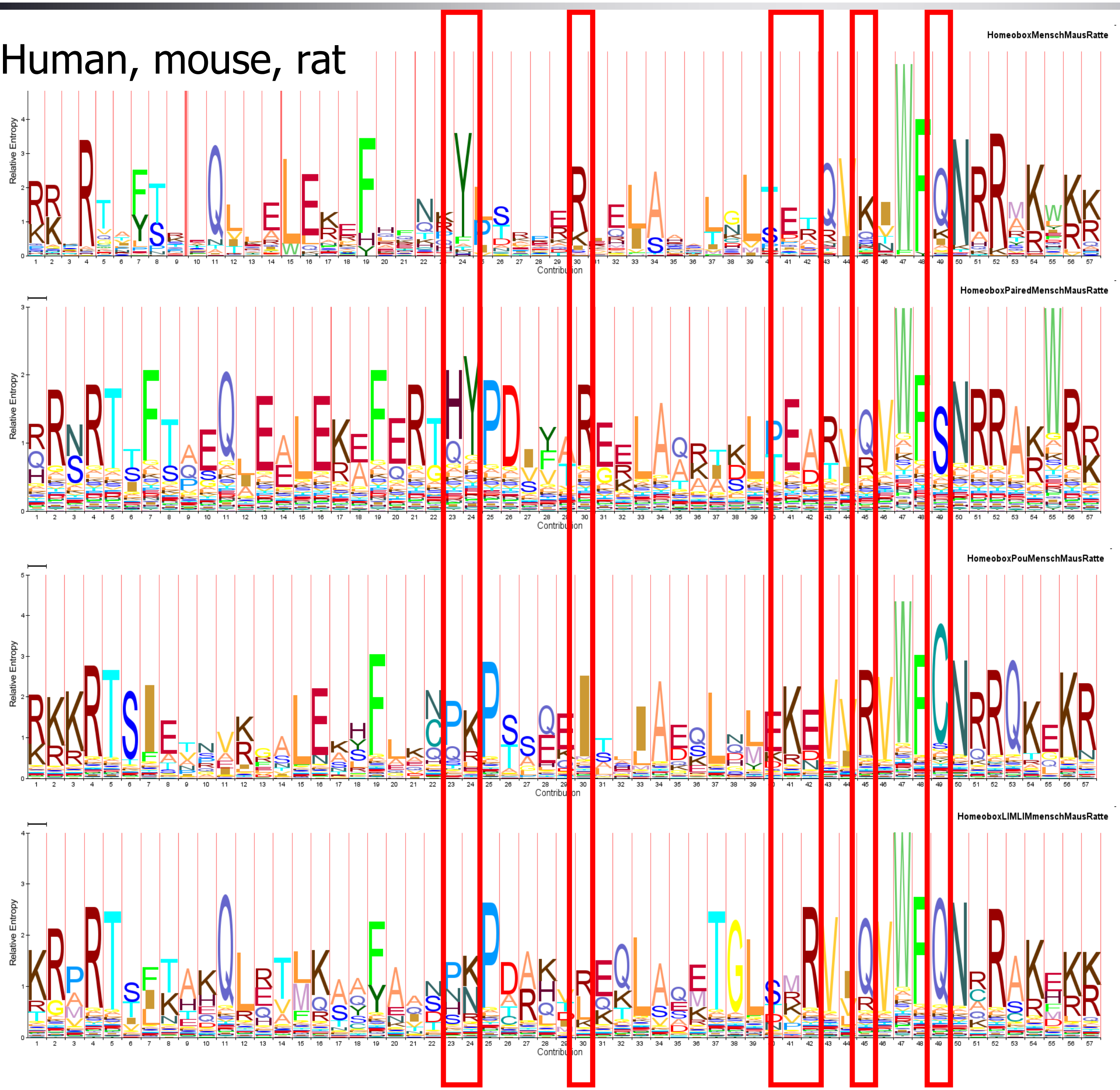
## Protein-DNA interaction code

### Homeo domains



# Protein-DNA interaction code

Human, mouse, rat




Homeo domain only

Homeo domain in hom-paired proteins

Homeo domain in hom-POU proteins

Homeo domain in hom-lim proteins

# Protein-DNA interaction code

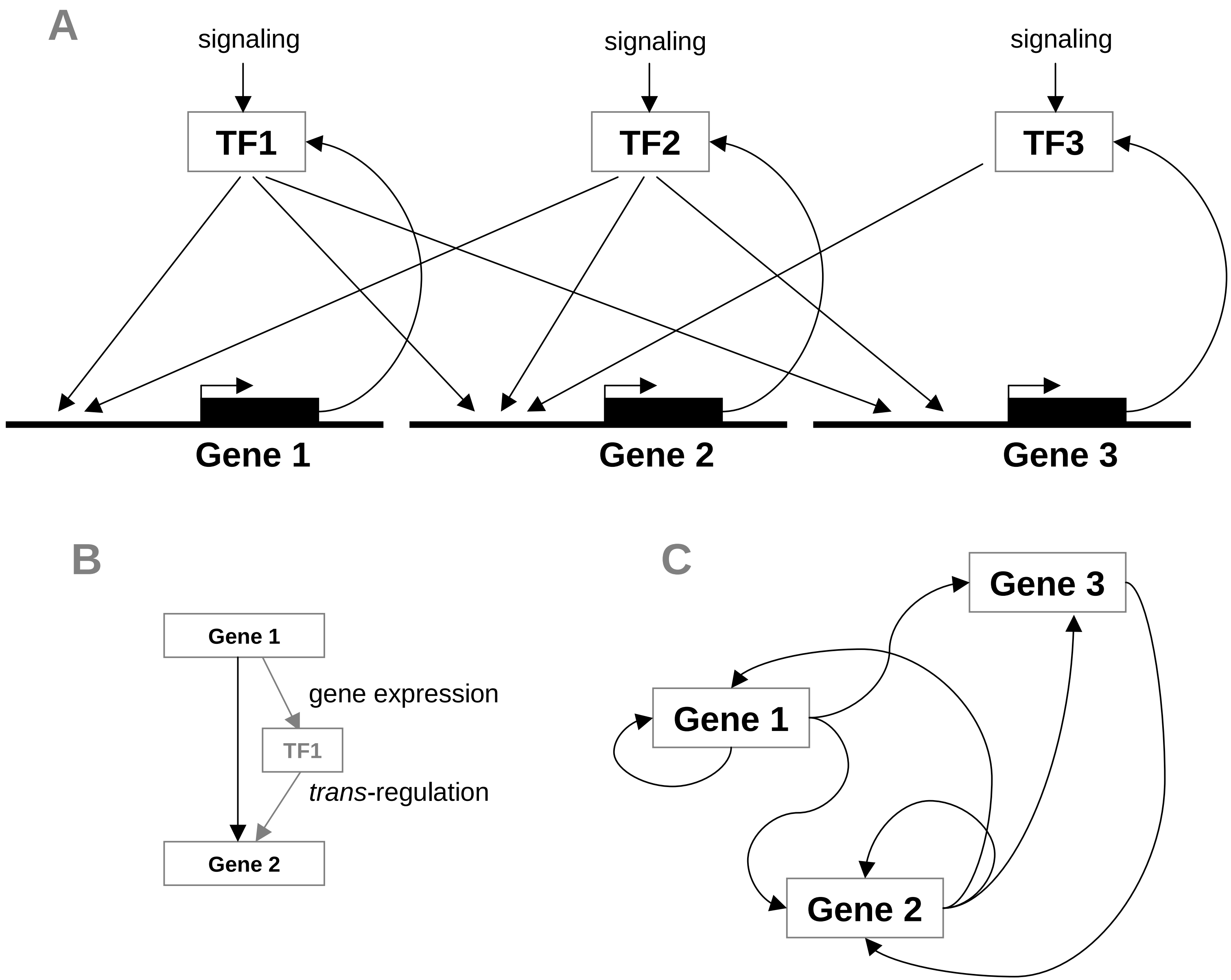
In a next step, it will be investigated whether whole DBDs or just a few residues are required to reveal mutual dependencies between DBDs and their target DNA sequences.

Goal #6:

Re-engineering and analyzing the complete transcription network of a cell / an organism.

UMG

## Definition



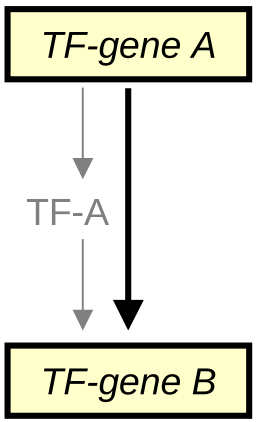
# Transcription network

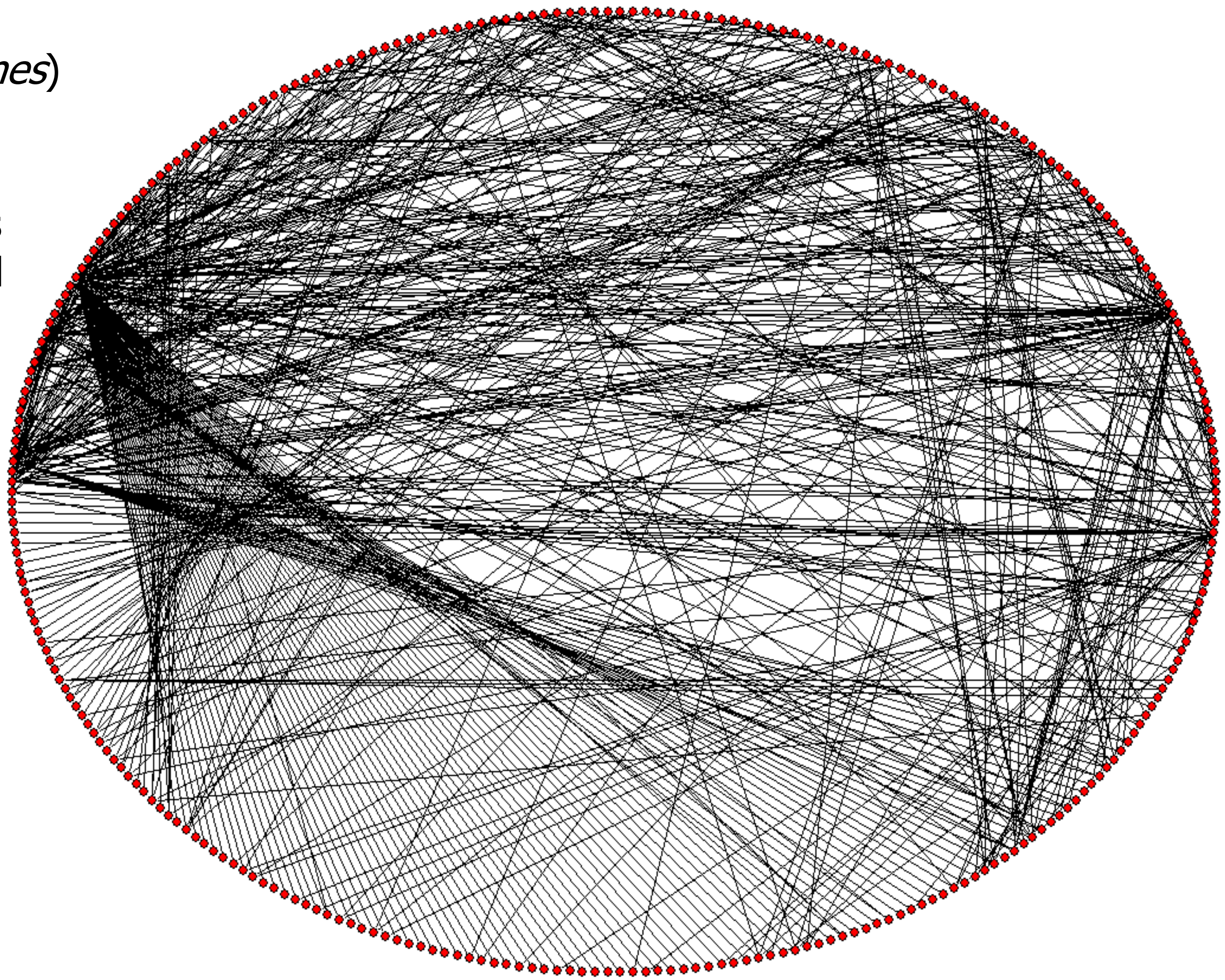
## Mammalian network of transcription factor genes (TFG_RN)

298 vertices (*TF-genes*)

638 edges

Each edge combines gene expression and *trans*-regulation



## Disconnectivity network analysis

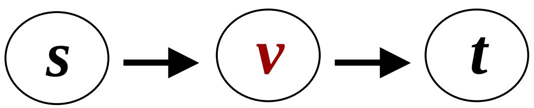
### Disconnectivity: Effect of removing a vertex

$$Dis(v) = \frac{N_0 - N_{-v}}{N_0} = 1 - \frac{N_{-v}}{N_0}$$

$N_0$ - number of all-pairs shortest paths

$N_{-v}$ - number of all-pairs shortest paths after removing vertex *v* from the graph.

$$Dis(v) = \frac{\sum_{s \neq v \in V} \delta_{sv} + \sum_{s \neq v \neq t \in V} \delta_{st}(v) + \sum_{t \neq v \in V} \delta_{vt}}{\sum_{s \neq t \in V} \delta_{st}}$$



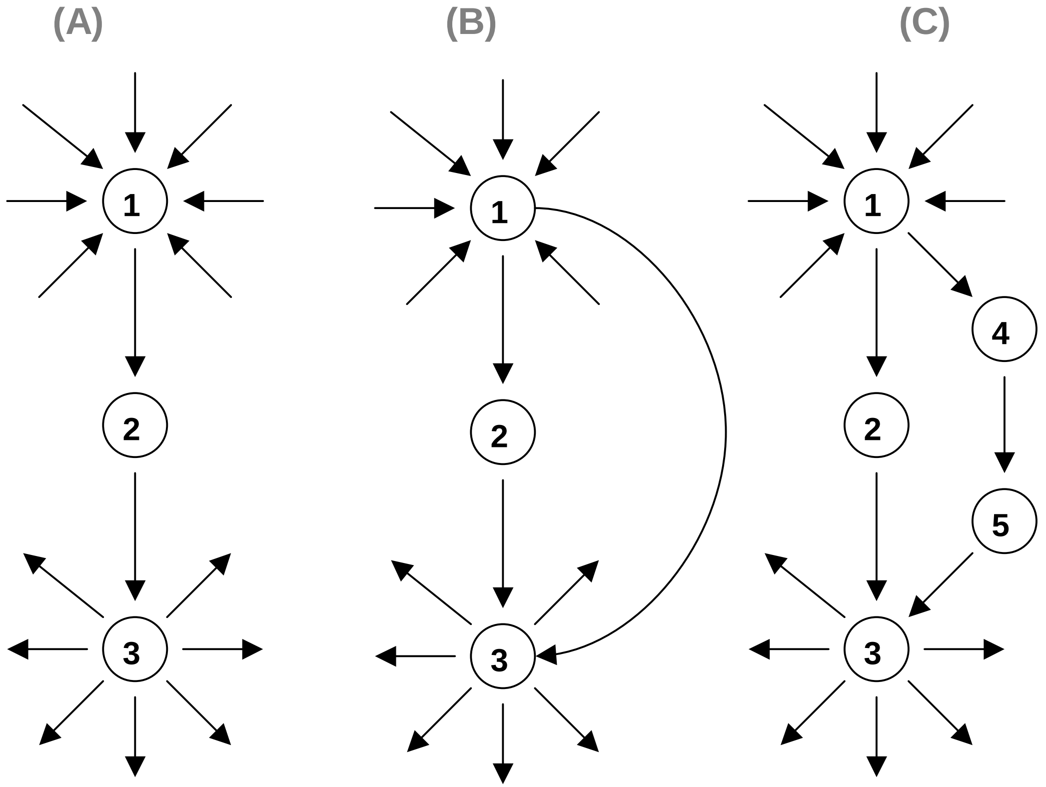

Fraction of all-pairs connections for which vertex ***v*** is crucial: there is no other parallel path(s) between vertices of such a pair and these vertices will not be connected anymore if vertex ***v*** is deleted.

The betweenness centrality and the disconnectivity index $D_i$ display different aspects of individual vertex topology in a whole network.

Potapov et al., BMC Bioinformatics, 9, 227 (2008)

## Disconnectivity network analysis

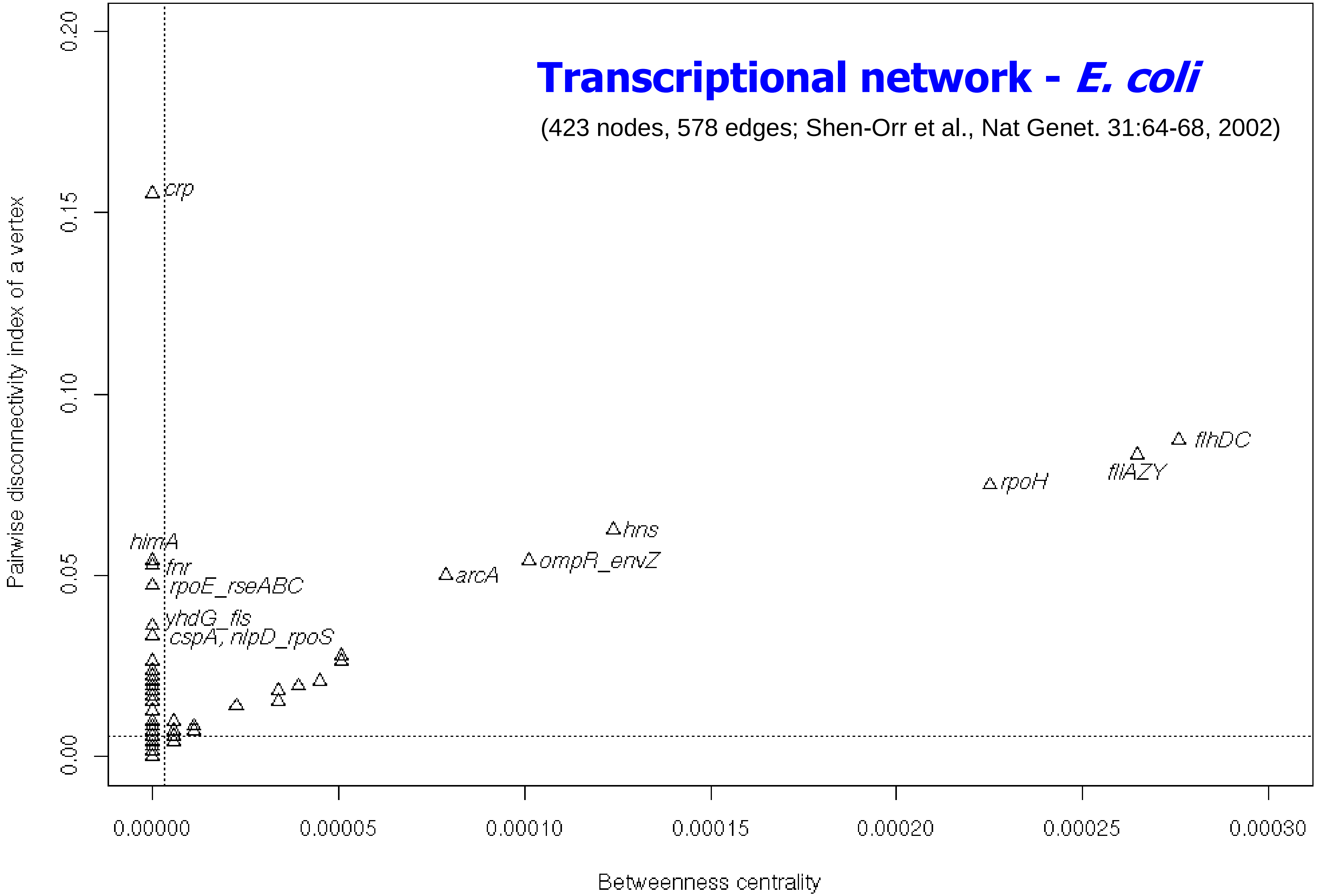
Disconnectivity sensitive towards bypasses



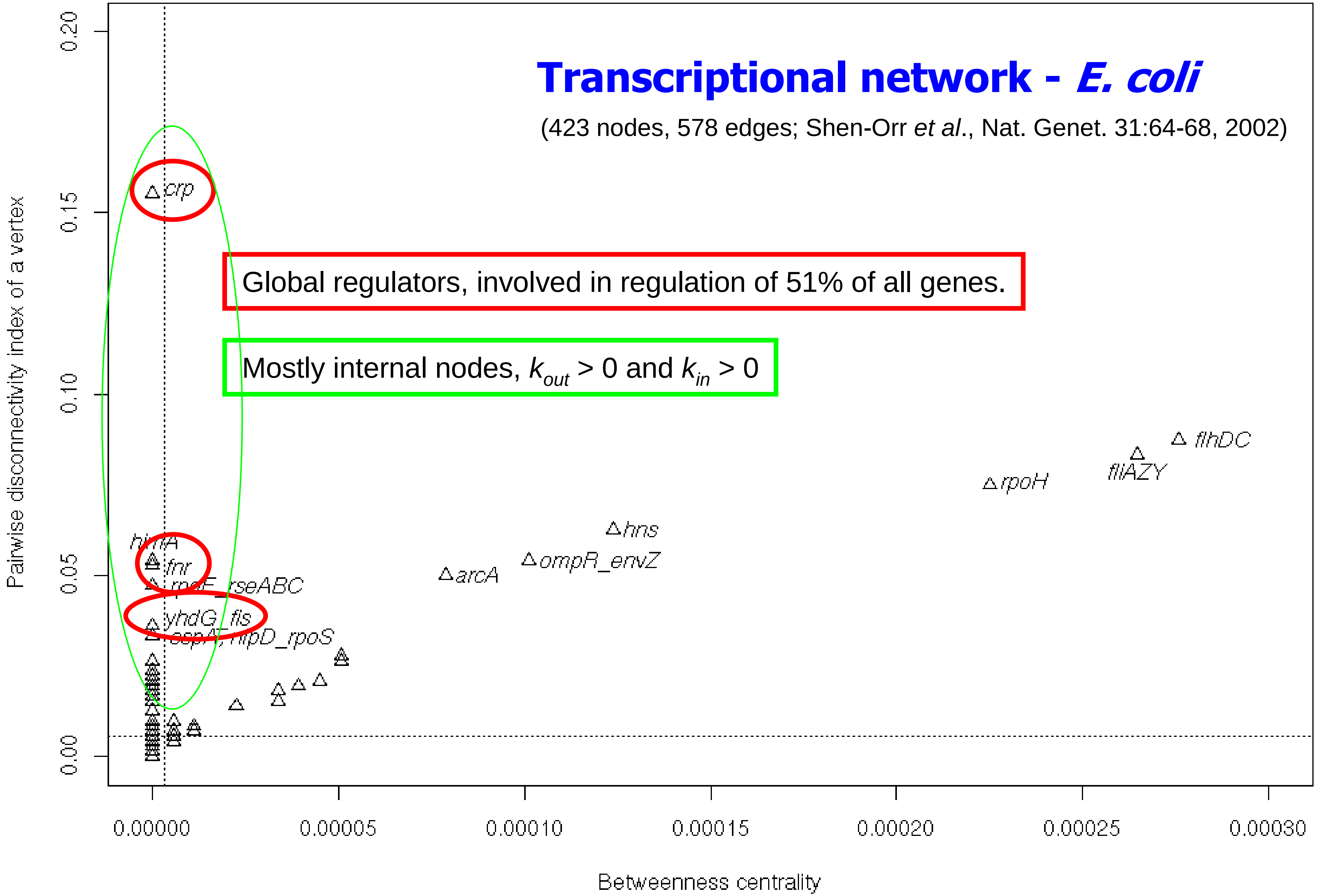

| | (A) | (B) | (C) |
|---|---|---|---|
| *B*(*v*2): | high | low | high |
| *Dis*(*v*2): | high | low | low |

Potapov et al., BMC Bioinformatics, 9, 227 (2008)

# Transcription network



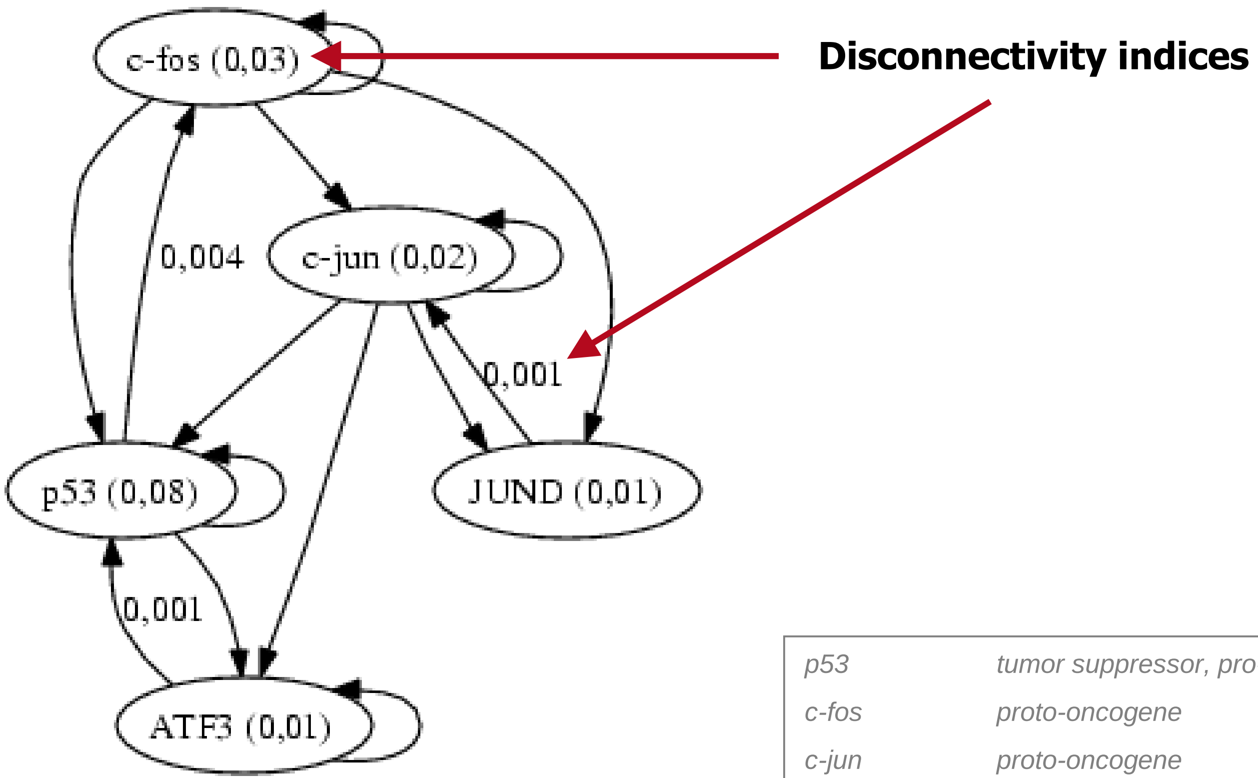

Potapov et al., BMC Bioinformatics, 9, 227 (2008)




Potapov et al., BMC Bioinformatics, 9, 227 (2008)

# Transcription network

## ‘p53’ strongly connected sub-graph in TFG_RN

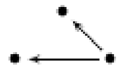
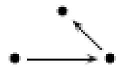
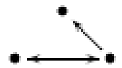
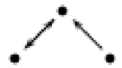
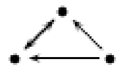
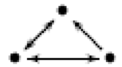
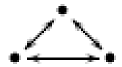



| | |
|---|---|
| *p53* | *tumor suppressor, pro-apoptotic* |
| *c-fos* | *proto-oncogene* |
| *c-jun* | *proto-oncogene* |
| *JUND* | *proto-oncogene* |
| *ATF3* | *tumor suppressor, pro-apoptotic* |

Potapov et al., BMC Bioinformatics, 9, 227 (2008)

# Transcriptional network

**Pattern / motif characterization by Disconnectivity**

| | | *E. coli* | | | *S. cerevisiae* | | | *Mammals* | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Pattern* | *ID* | *Freq* | *Z-Score* | *$\overline{Dis}$* | *Freq* | *Z-Score* | *$\overline{Dis}$* | *Freq* | *Z-Score* | *$\overline{Dis}$* |
|  | 6 | 4777 | 11.23 | 0.0039 | 11892 | 14.54 | 0.0018 | 1916 | -0.39 | 0.0023 |
|  | 12 | 160 | -11.21 | 0.0189 | 295 | -14.25 | 0.0135 | 1068 | -1.67 | 0.011 |
|  | 14 | - | - | - | 18 | -1.30 | 0.0063 | 73 | -10.47 | 0.0079 |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | …044 |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | …050 |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | …042 |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | …099 |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | …224 |
| [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | …137 |
|  | 164 | - | - | - | - | - | - | 197 | -6.52 | 0.0051 |
|  | 166 | - | - | - | 1 | 4.65 | 0.0052 | 20 | 7.12 | 0.0058 |
|  | 174 | - | - | - | - | - | - | 6 | 7.31 | 0.0109 |
|  | 238 | - | - | - | - | - | - | 1 | - | 0.0073 |

In transcription networks, the only 3-vertex motif appearing in all networks analyzed is the feed-forward loop (confirmed).

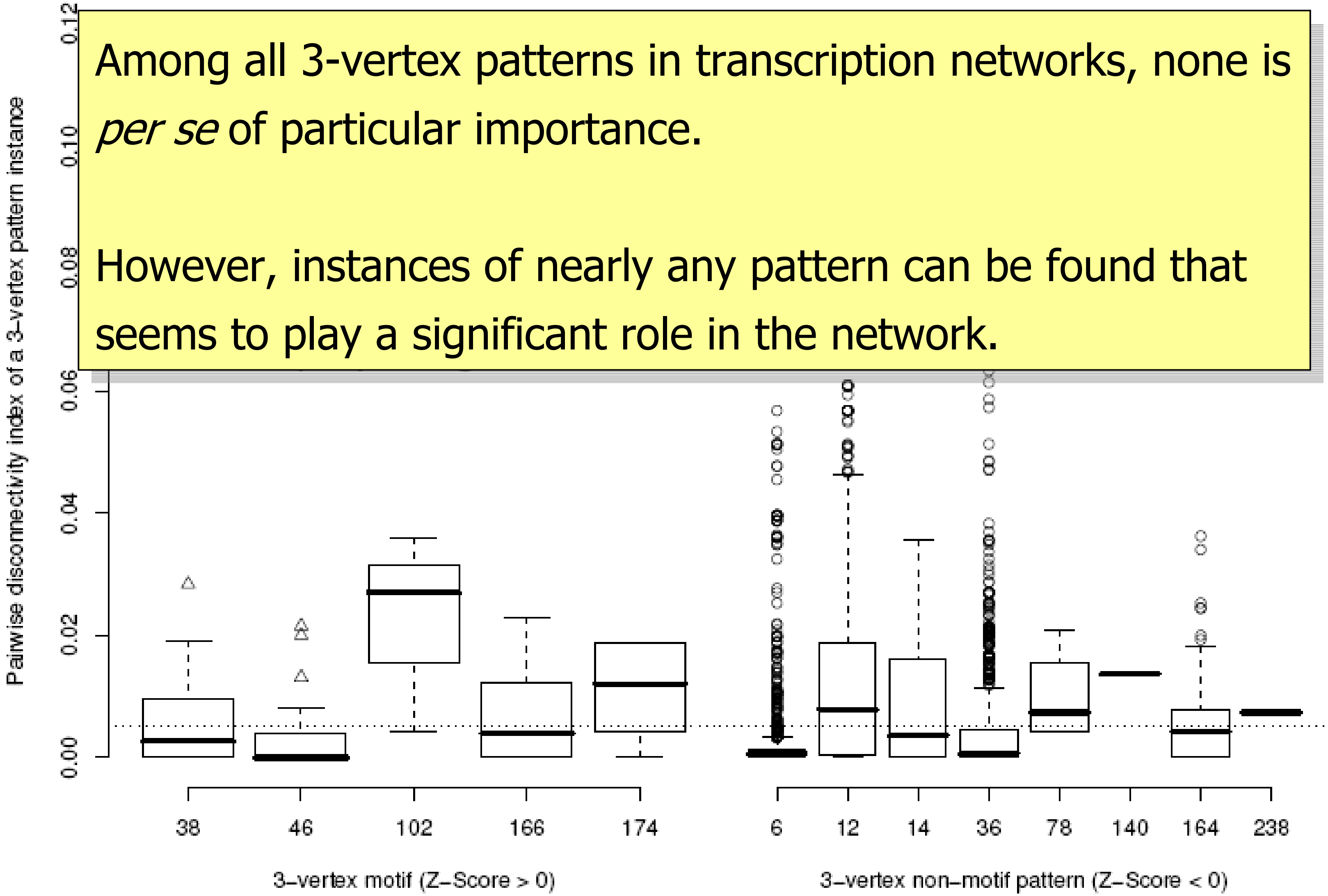
However, motifs *per se* do not play a more important role for the coherence of the network than non-motif patterns.

Goemann et al., BMC Syst. Biol. 3, 53 (2009)

# Transcriptional network

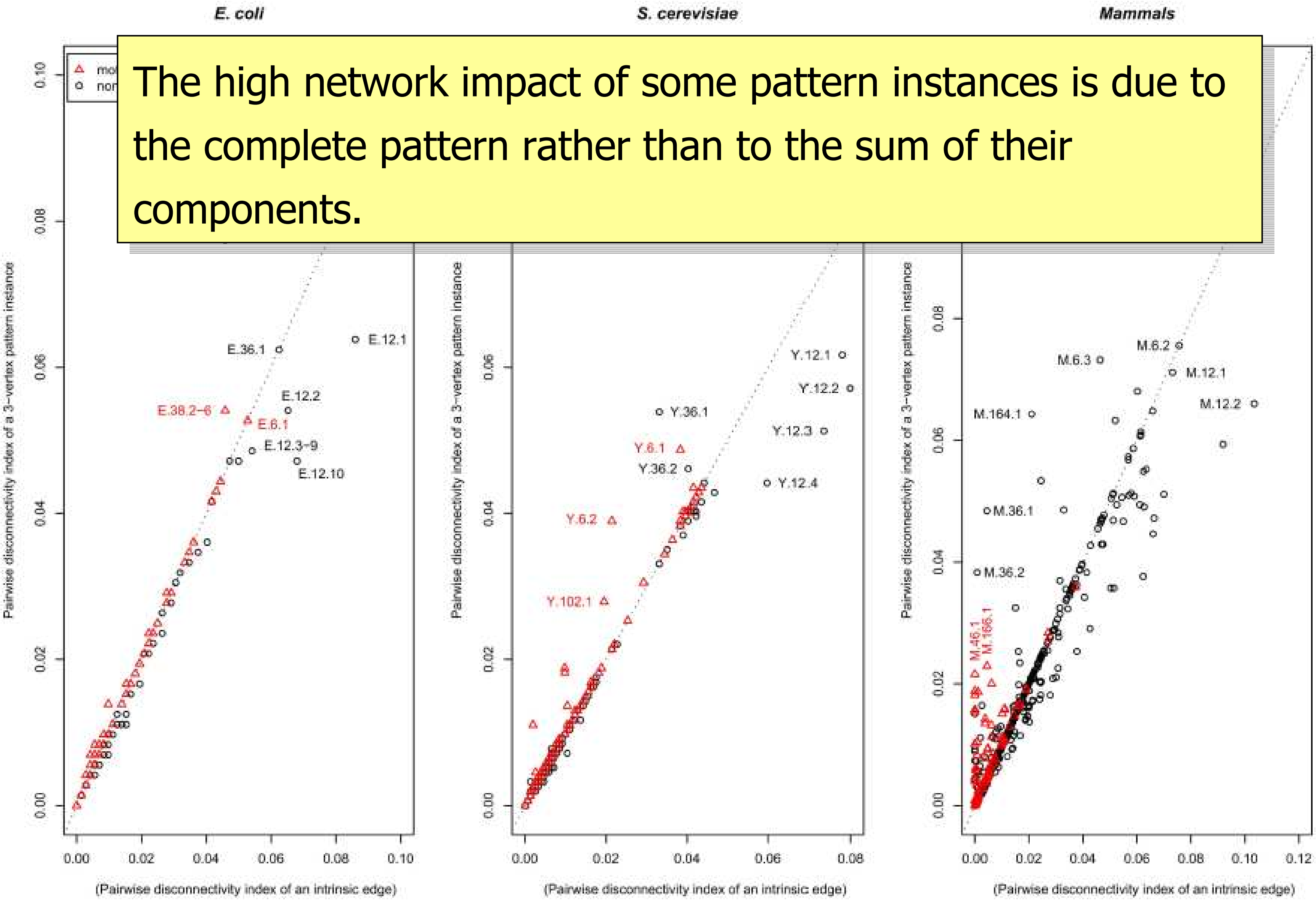
Among all 3-vertex patterns in transcription networks, none is *per se* of particular importance.

However, instances of nearly any pattern can be found that seems to play a significant role in the network.




Goemann et al., BMC Syst. Biol. 3, 53 (2009)

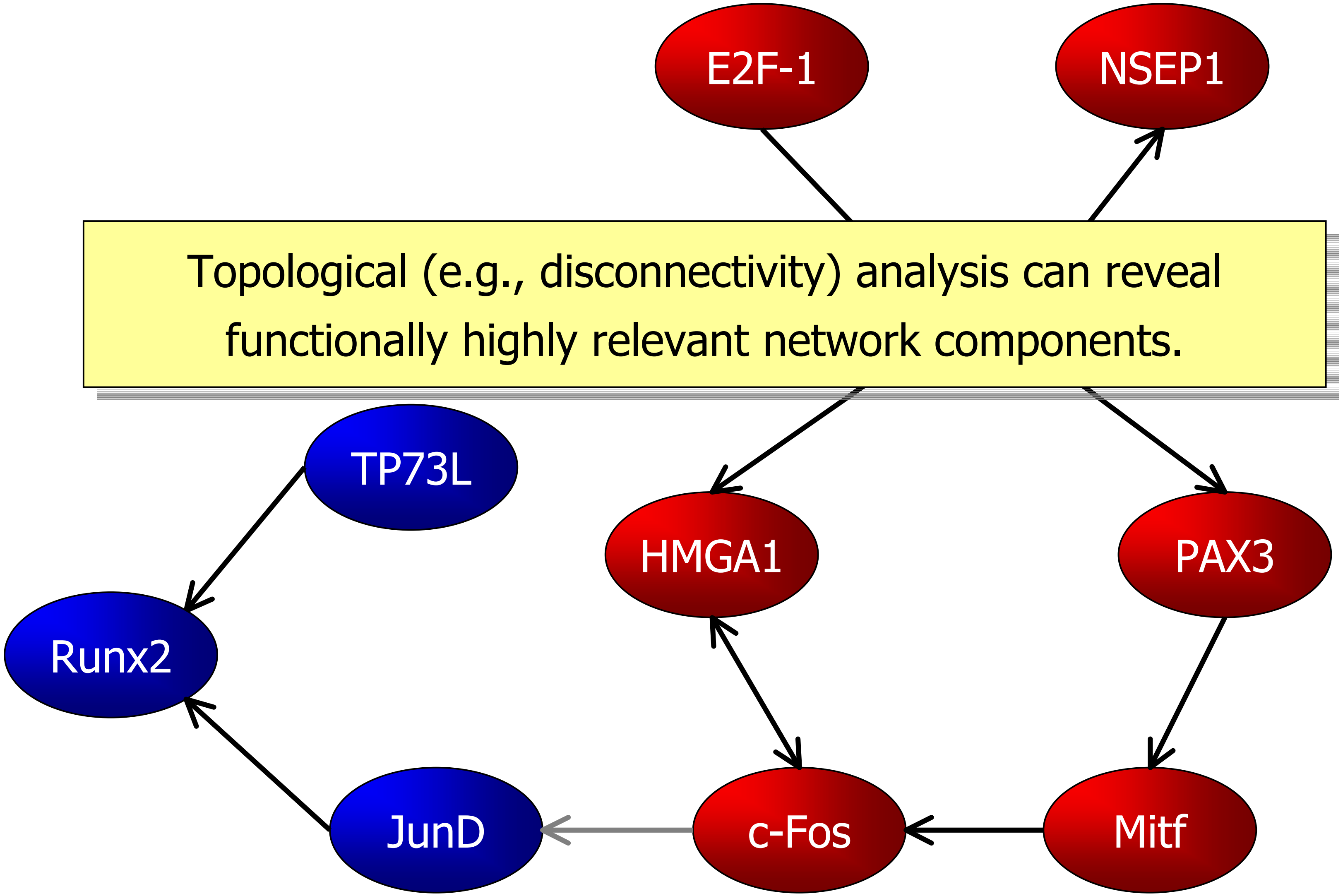
# Transcriptional network




The high network impact of some pattern instances is due to the complete pattern rather than to the sum of their components.

Goemann et al., BMC Syst. Biol. 3, 53 (2009)

## The largest mammalian subnetwork according to motif analysis is functionally linked with cell-cycle control




Topological (e.g., disconnectivity) analysis can reveal functionally highly relevant network components.

## Conclusions (1)

- Through the past 20 years, our understanding of the mode how transcription factors (TFs) operate has made remarkable progress.
- New HTP technologies have increased the chances to come up with a genome-wide map of TF binding sites (TFBS), which, however, will always be a kind of snapshot of a selection of cellular states.
- Context-dependent approaches to TFBS prediction that also make use of comparative genomics may provide a major breakthrough.

## Conclusions (2)

- While TRANSFAC provides the knowledge base for this, state-of-the-art algorithms implemented in ExPlain™ reveal the syntax of promoters.
- The knowledge compiled in the TRANSFAC database may enable us to even predict the DNA-binding specificity of newly discovered TFs.
- Analyzing the architecture of the transcriptional network of cells will reveal key regulators that may be ideal candidates for diagnostic, therapeutic or biotechnological purposes.