A new era in transcription factor research

Edgar Wingender

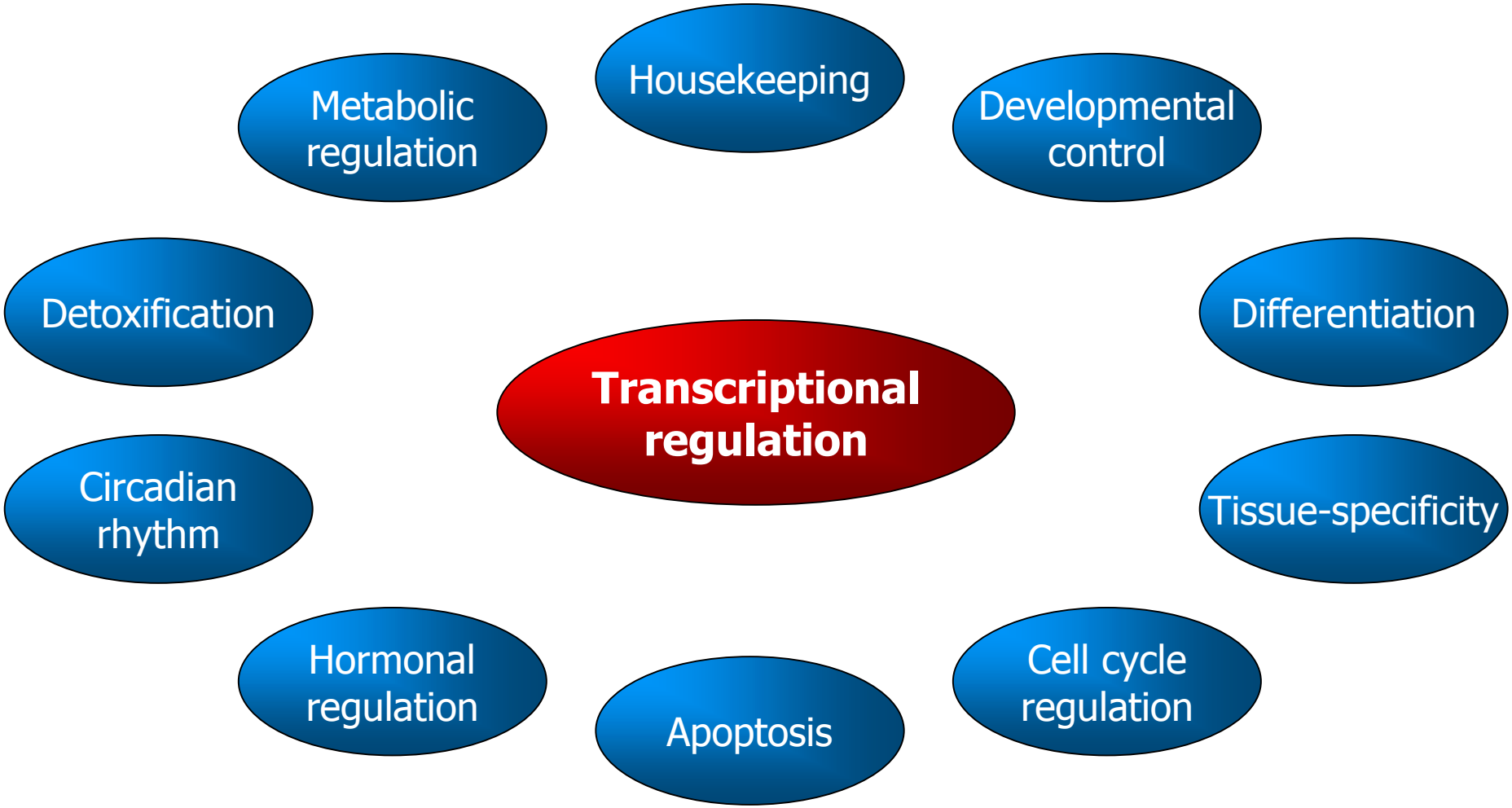
UNIVERSITÄTSMEDIZIN GÖTTINGEN UMG

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

Bioinformatics

(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

Systems biology

(7) To understand transcriptional dysregulation under disease conditions

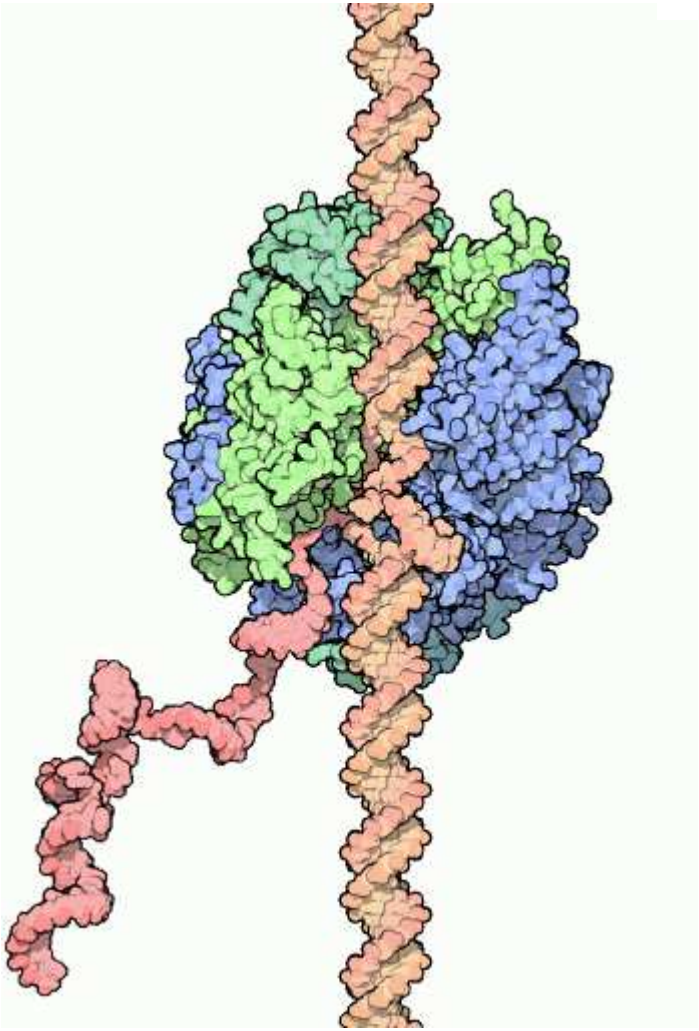
(8) To render transcriptional regulation amenable for targeted alterations

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:

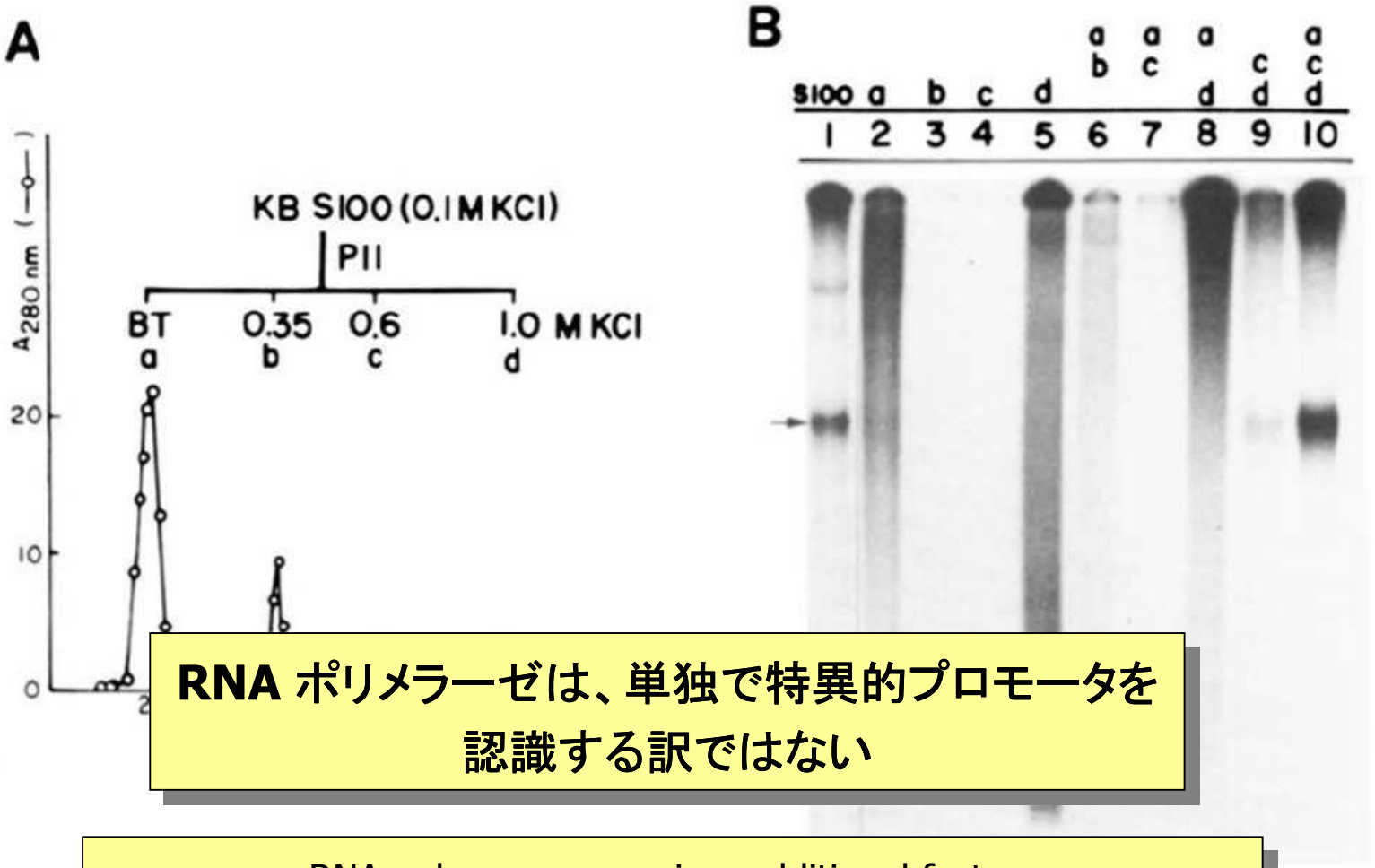
真核生物における特異的転写調節機構の理解

Goal #1:

Understand how the specificity of eukaryotic transcriptional regulation is achieved?

Specificity of transcriptional regulation

General transcription factors



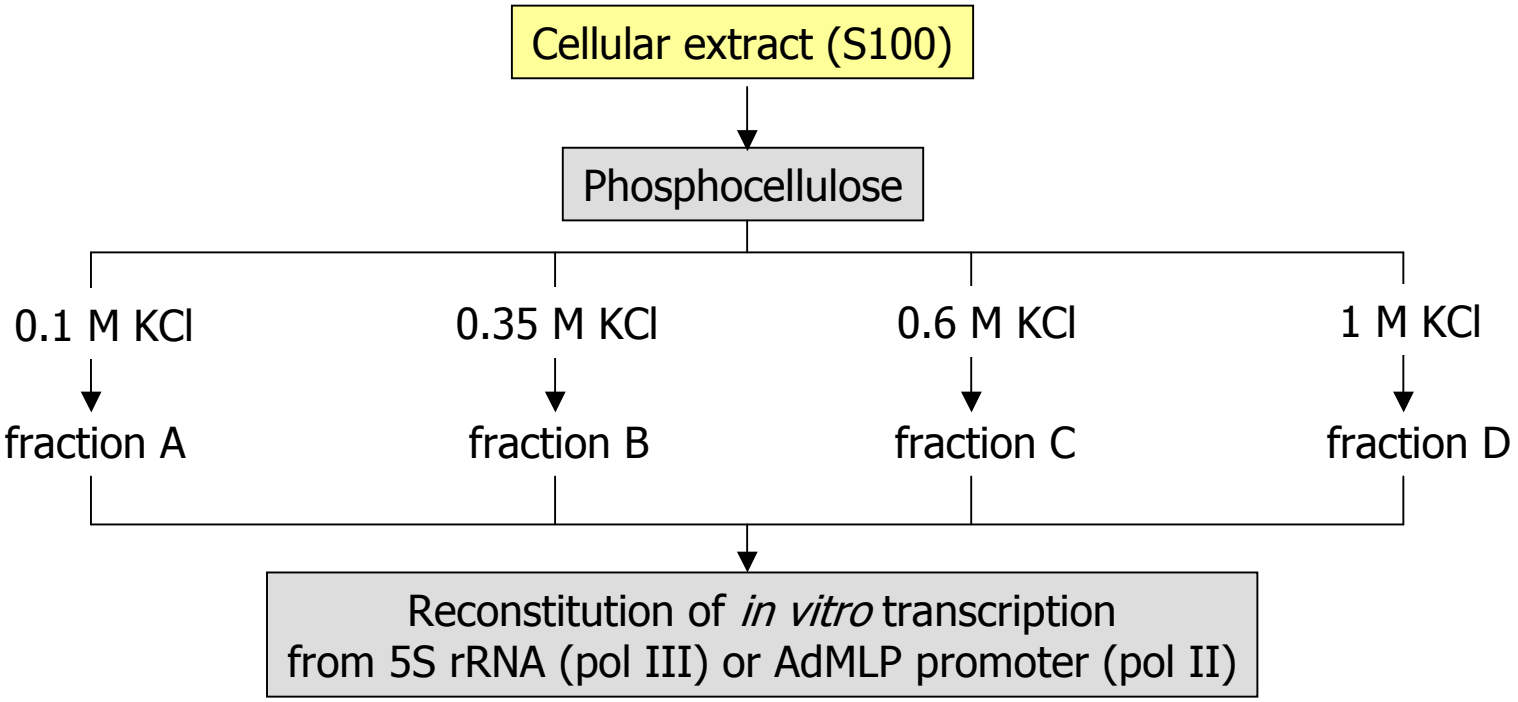
RNA ポリメラーゼは、単独で特異的プロモータを認識する訳ではない

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

Specificity of transcriptional regulation

General transcription factors

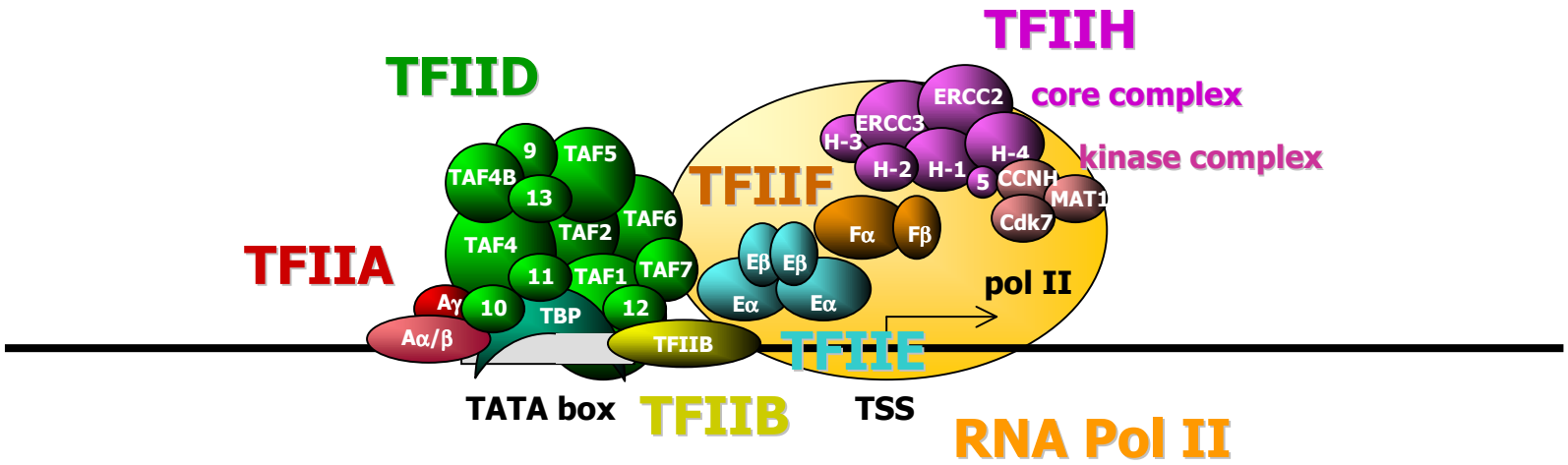


pol III:	TFIIIA	TFIIIB	TFIIIC	--
pol II:	TFIIA	--	TFIIB	TFIID

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

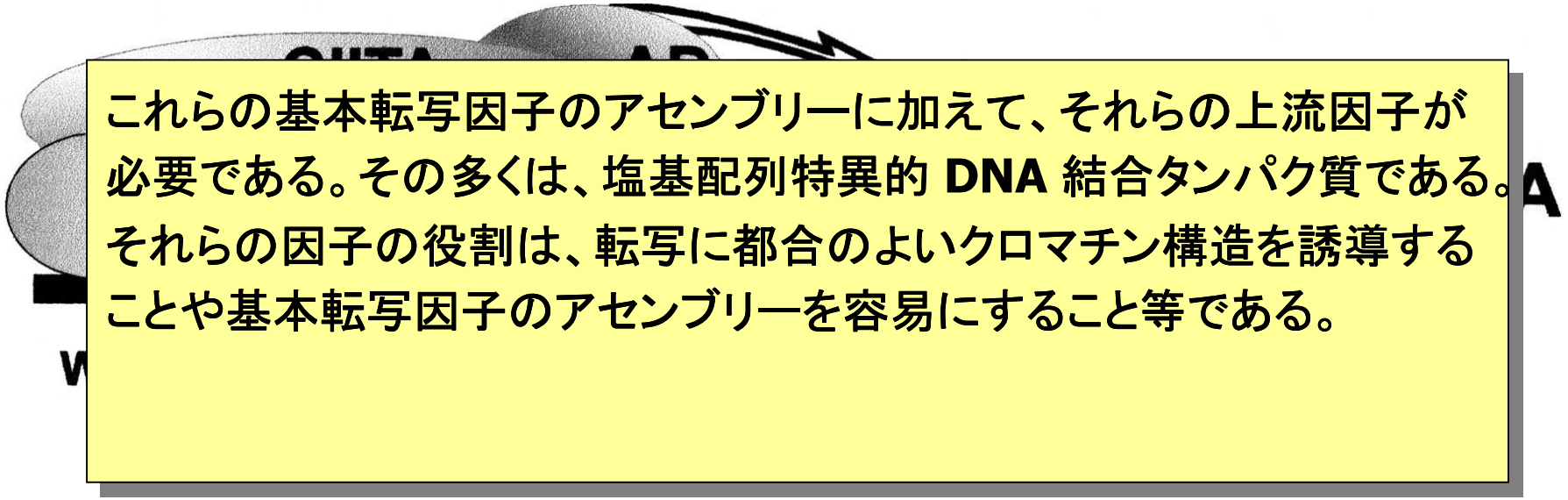
General transcription factors

The pre-initiation transcription complex



Specificity of transcriptional regulation

Enhanceosome



これらの基本転写因子のアセンブリーに加えて、それらの上流因子が必要である。その多くは、塩基配列特異的 **DNA** 結合タンパク質である。それらの因子の役割は、転写に都合のよいクロマチン構造を誘導することや基本転写因子のアセンブリーを容易にすること等である。

Masternak K et al., Genes Dev 2000 May 1;14(9):1156-66

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

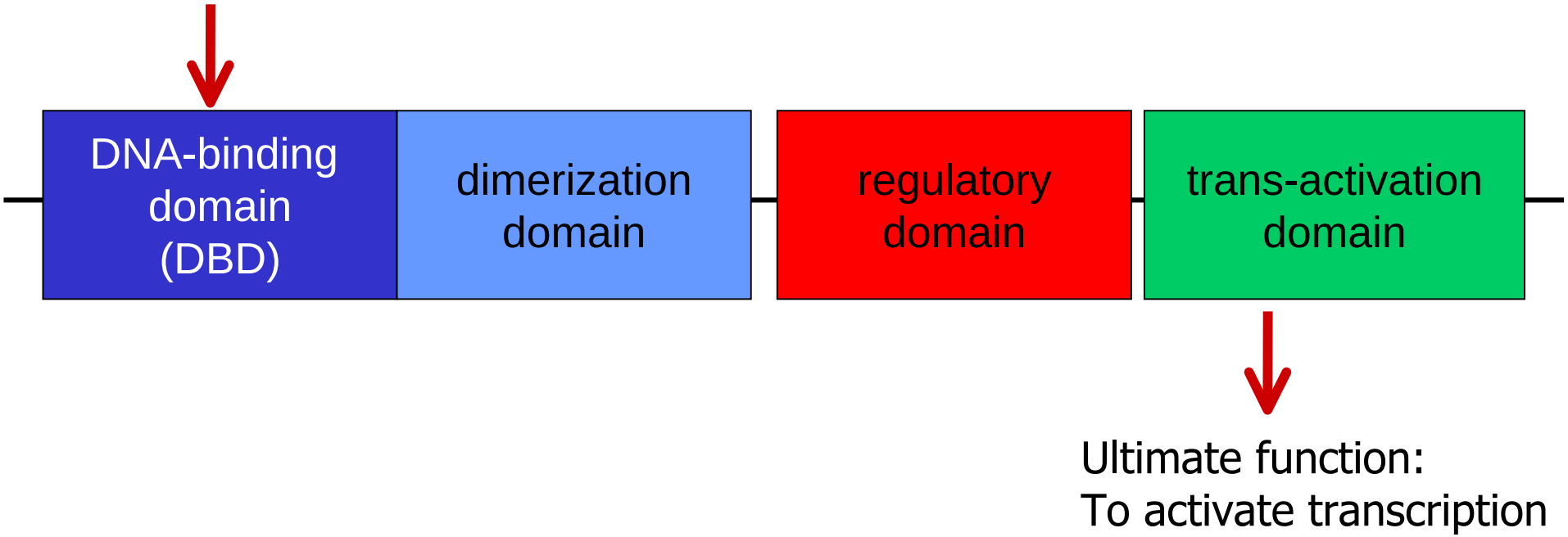
What is a transcription factor?

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



Goal #2:

ゲノム DNA 結合タンパク質の網羅的結合部位の地図作製

Goal #2:

To generate a comprehensive map of genome-bound proteins.

The genome-wide map: TRANSFAC Site Table

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 genome-wide map: TRANSFAC Site Table

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. -proxim.prom.	Drosophila	- 85 to - 47	1a	GCTGCtGCTGCatcCGTCGaCGTCG	Adf-1	11
		-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

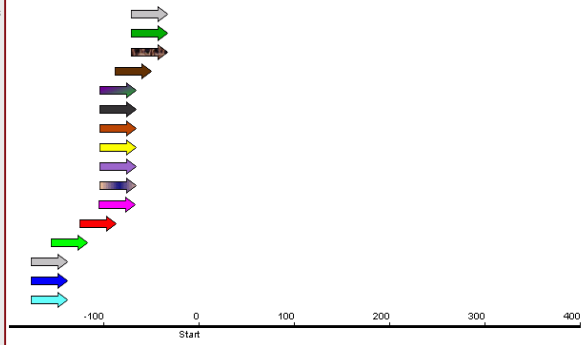
Gene Regulation [hide]

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 presen

A Site entry

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 HNF-3alpha(r) Quality:2 HNF-3alpha(h) Quality:5
Factor Source	rec(rat-COS-7); Rat; recombinant expression: rat factor has been expressed in COS-7 cells HepG2; Human;
Method	functional analysis, direct gel shift, gel shift competition
Comments	Positive regulatory element [1];

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

Genome-wide で迅速に結合部位地図が作製可能な新技術

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

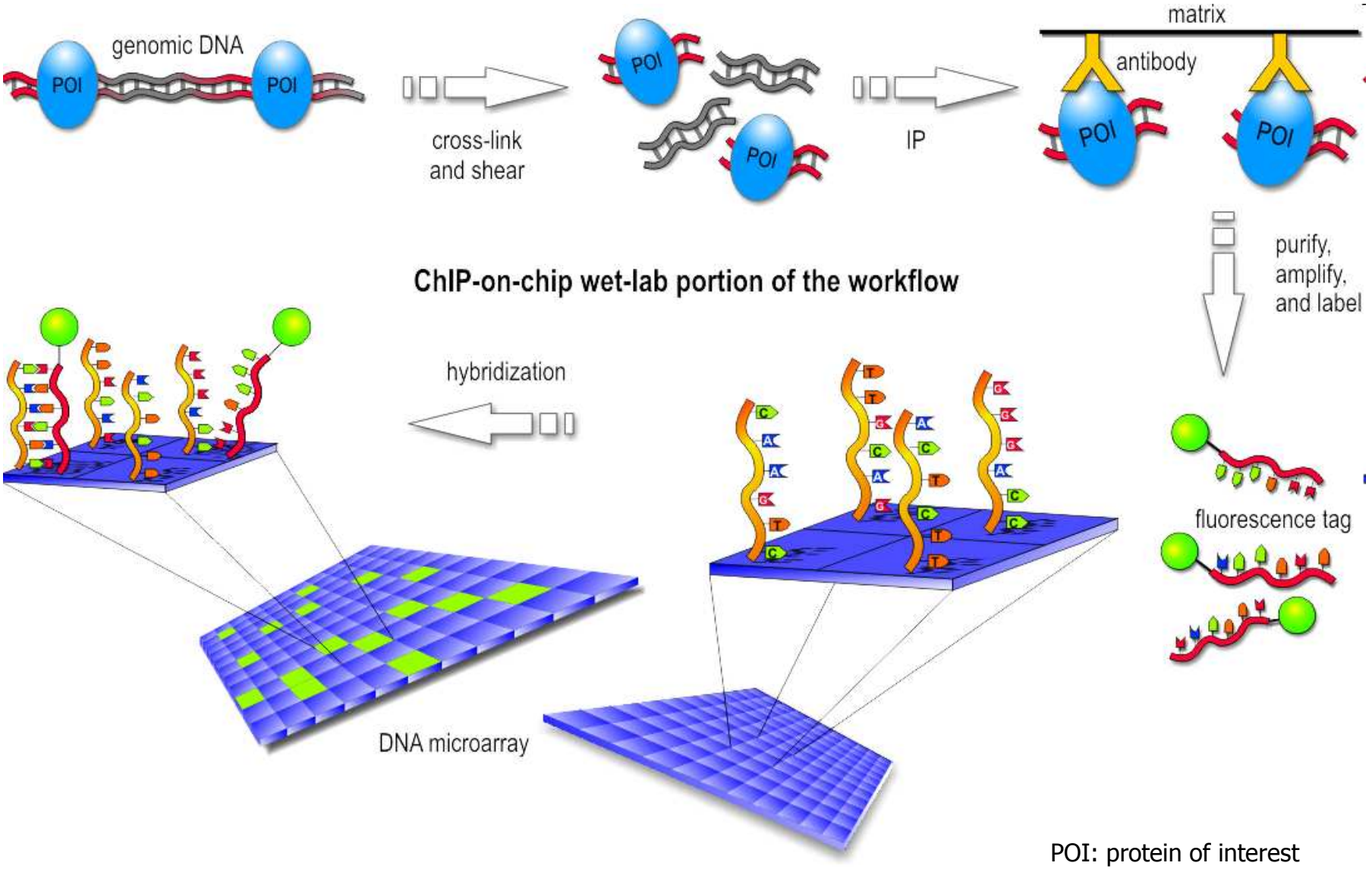
Additional 130,000 ChIP-Seq fragments coming up with the Summer release.

次回の update で、13万の ChIP-Seq 実験由来の結合部位情報が追加予定

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

The genome-wide map: TRANSFAC

High throughput approaches: ChIP-chip

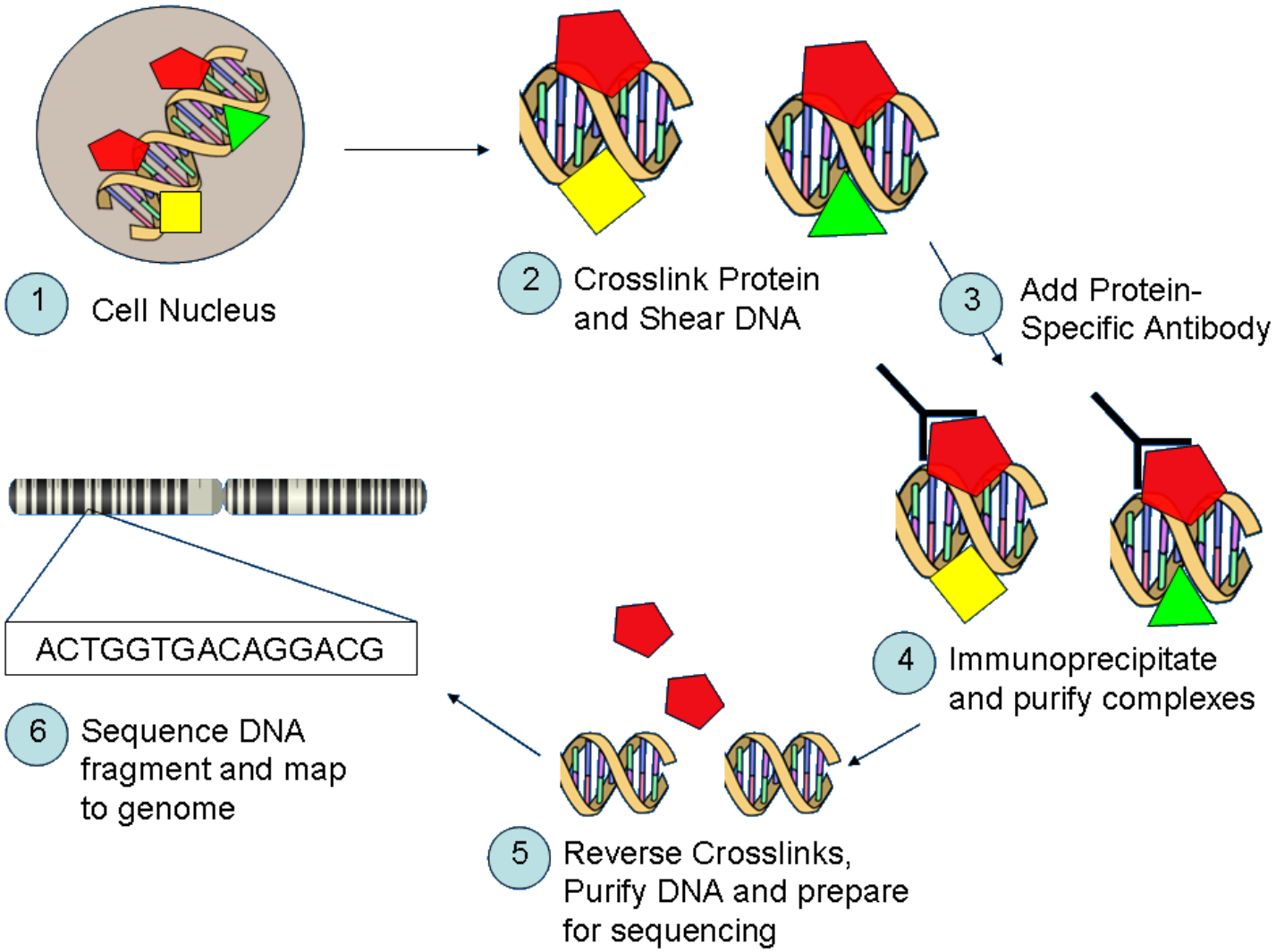


http://en.wikipedia.org/wiki/ChIP-chip

POI: protein of interest
IP: immunoprecipitation

The genome-wide map: TRANSFAC

High throughput approaches: ChIP-seq



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

The genome-wide map: TRANSFAC

The present status

次回の update で、13万の ChIP-Seq 実験由来の結合部位情報を追加予定

Current data sets (TRANSFAC 2009.1)

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

ChIP-chip

Available with release of Sep09

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**

ChIP-seq

interacting fragments for 17 TFs

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

The genome-wide map: TRANSFAC

The present status

2009 年中にトータルで 40 の転写因子の 約79万結合部位情報が追加される予定

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

殆どの結合部位に関する情報は、国際 ENCODE project に由来

Most of the fragments come from the international ENCODE project.

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

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 genome-wide map: TRANSFAC

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

The genome-wide map: TRANSFAC

UMG

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 (≥24 in human) and
- under all environmental conditions (--> ∞).

しかしながら、これらの新規 **high-throughput** 技術 を用いても完全な、いわゆる **promotome** を明らかにすることはできない。

理由１: 多数の転写因子 （約 **1800** のヒト転写因子）

理由２: 多数の細胞種 （約 **300** のヒト培養細胞）

理由３: 多数の組織 （約**14000** の形態的に識別可能な組織構造）

理由４: それら組織の異なる発生段階や環境下での結合部位情報

このような無限の条件下での実験は不可能

私たちが実施可能なことは、観察された実験結果からルールを見出し、新規結合配列予測システムを開発することである

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

Goal #3:
転写因子の新規結合部位の予測

Goal #3:
To enable prediction of new transcription factor binding sites.

Prediction of TFBSs

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.

結合配列情報とマトリックスは、転写因子結合部位の予測に使用される

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

Matrix construction

TRANSFAC

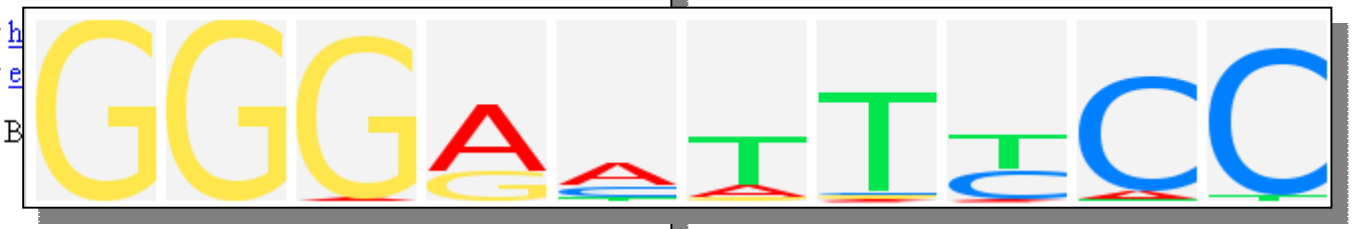
A	3	3	1	11	0	0	0	0	2	2	0	0
C	6	2	1	0	12	0	0	0	7	3	2	4
G	1	7	3	1	0	12	0	12	3	6	8	4
T	2	0	7	0	0	0	12	0	0	1	2	4

GCCCTACGTGCTGTCTCA
CAGGCAACGTGCAGCCGGA G
CAGTGCATACGTGGGCTCCA
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	13.04.1995 by h
Updated	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. T00588; NF-kappaB; Species: mouse, Mus musculus. T00590; NF-kappaB; Species: human, Homo sapiens. T00593; p50; Species: human, Homo sapiens. T00594; RelA-p65; Species: human, Homo sapiens.



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).

UMG

Validation of potential TFBSs by comparative genome analysis:

一般的なアイデア：

機能的に重要な転写因子結合部位は、進化の過程で保存されると考えられるので、異なる種間のゲノムを比較した場合、保存されている（系統的 **footprinting**)。

この解析手法は、転写因子結合部位の予測において パターンサーチとは独立しており、精度を上げる可能性がある。

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.

UMG

Validation of potential TFBSs by comparative genome analysis:

では、保存されるのは **DNA** 配列なのかそれともパターンなのか？

However:

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

UMG

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
```

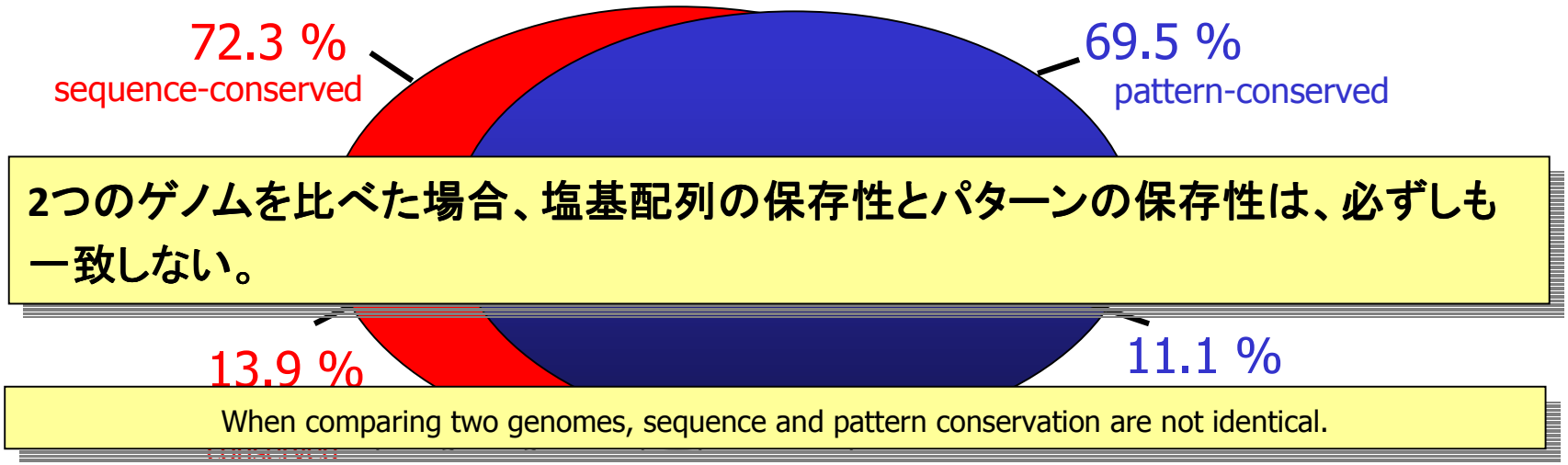
Matrix V$AP2ALPHA_02, M01045

UMG

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 %

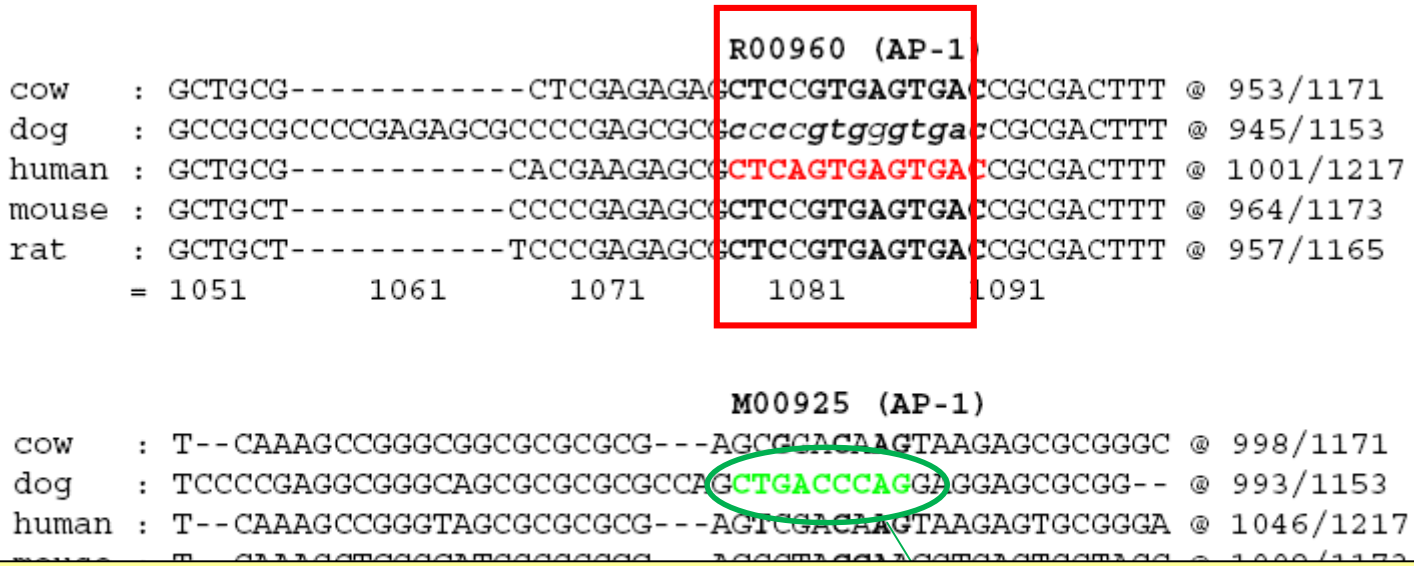


Sauer et al., Bioinformatics 22:403 (2006)

UMG

Validation of potential TFBSs by comparative genome analysis:

human c-*jun* gene



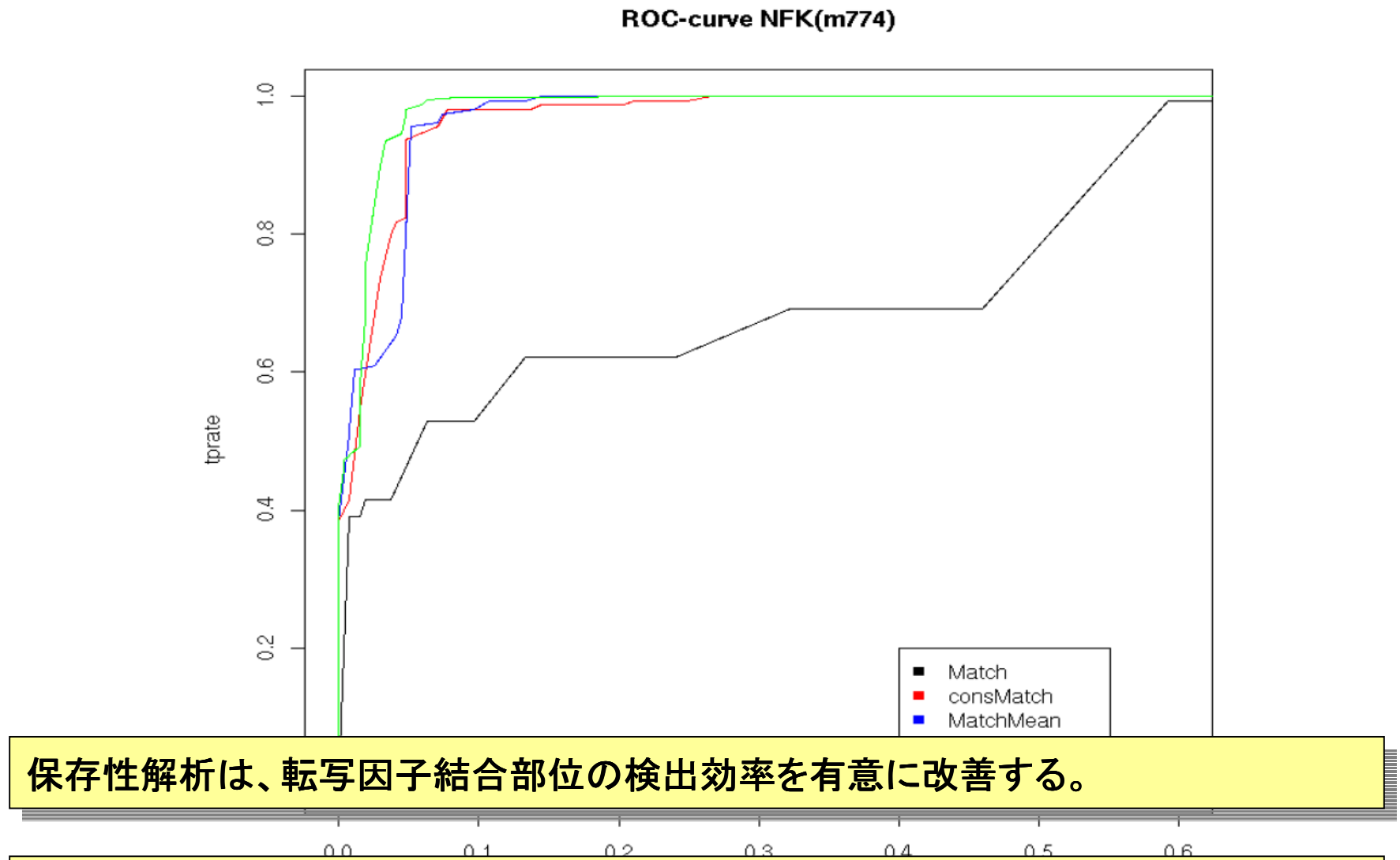
複数のゲノム比較は、保存された結合部位をより明らかにでき、いわゆる surrogate 部位を検出する。

surrogate site

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

UMG

Validation of potential TFBSs by comparative genome analysis:



保存性解析は、転写因子結合部位の検出効率を有意に改善する。

Conservativity analysis significantly improves detection of single TFBSs.

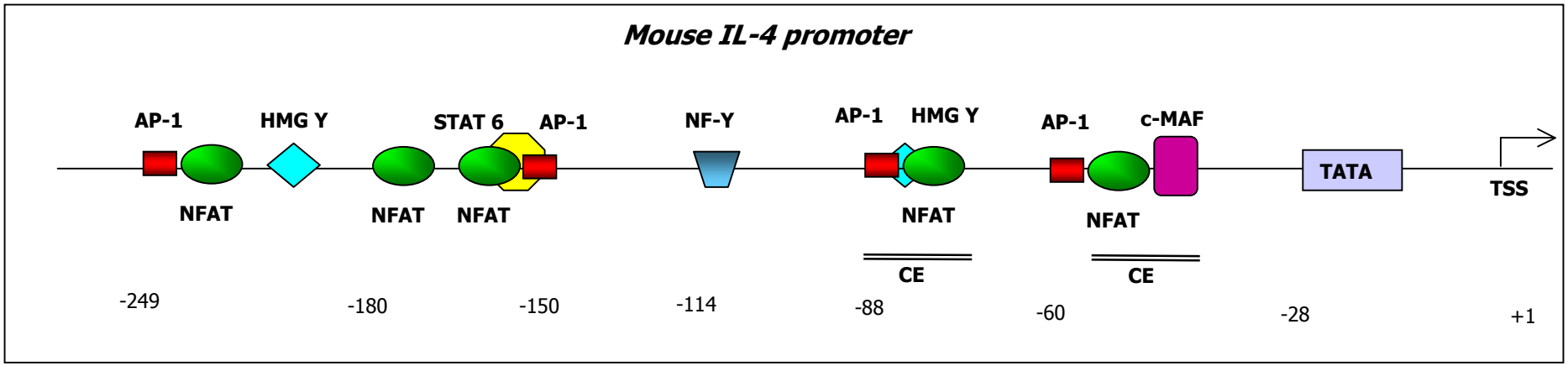
Goal #4:

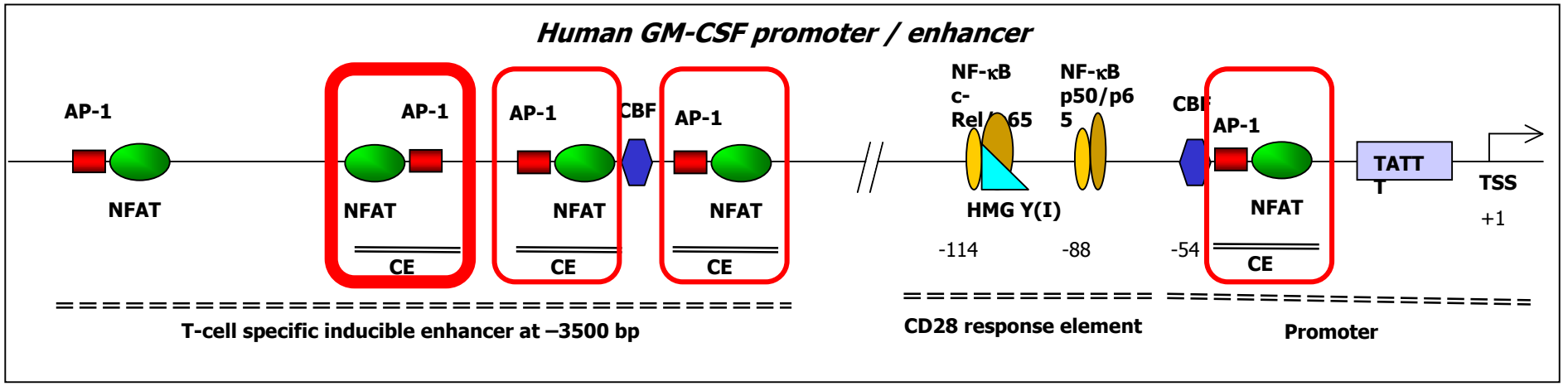
プロモータやエンハンサーといった転写調節領域の複雑な構造の解明

Goal #4:

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

Promoters and enhancers: examples



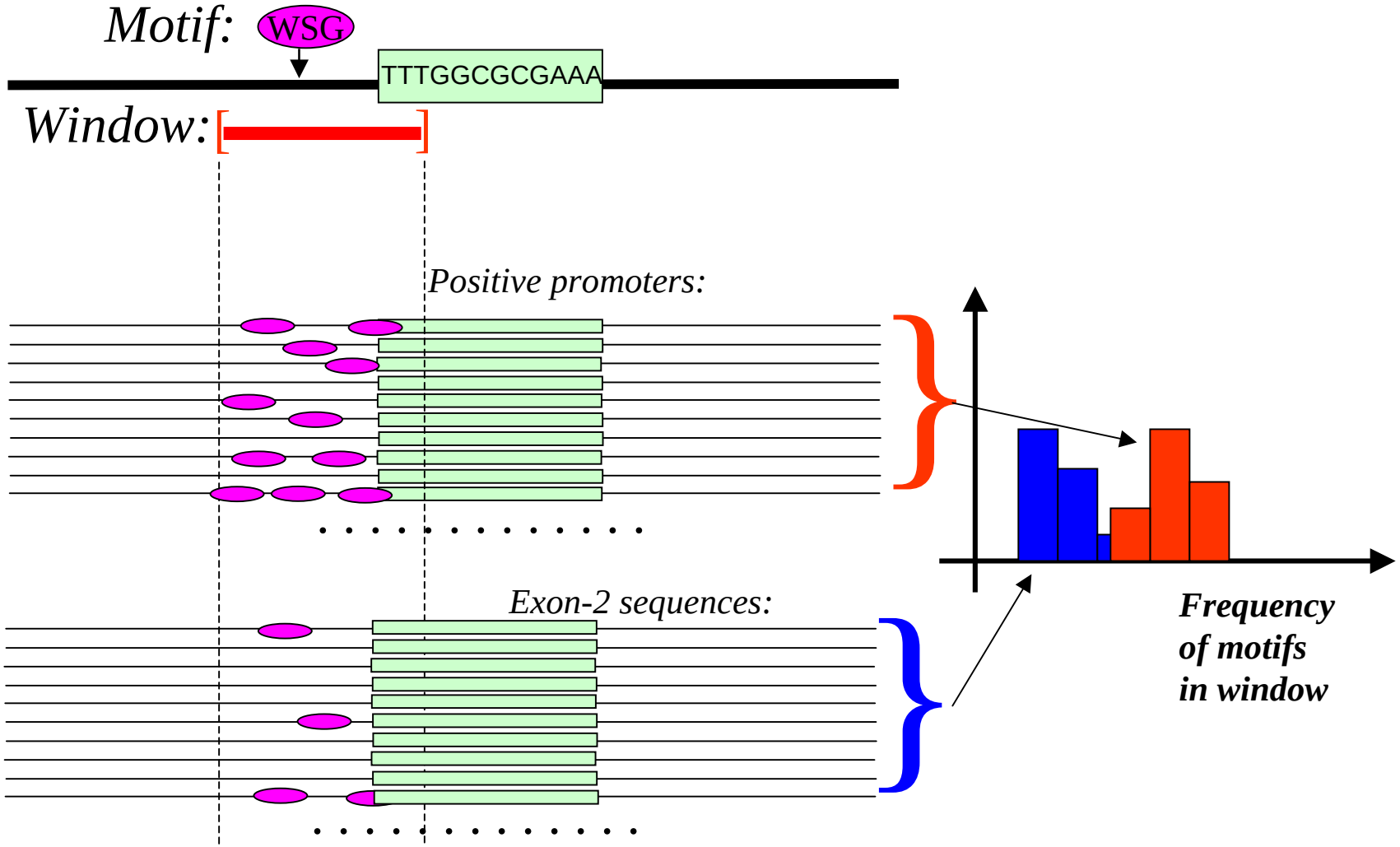


Local context analysis

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.

UMG

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 λ in window $w=[t_1, t_2]$ of sequence S

		N	Motif (λ)	Window $(w)^{1)}$	$\hat{f}^{Y} / \hat{f}^{N}$ 2)	Utility	α_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:

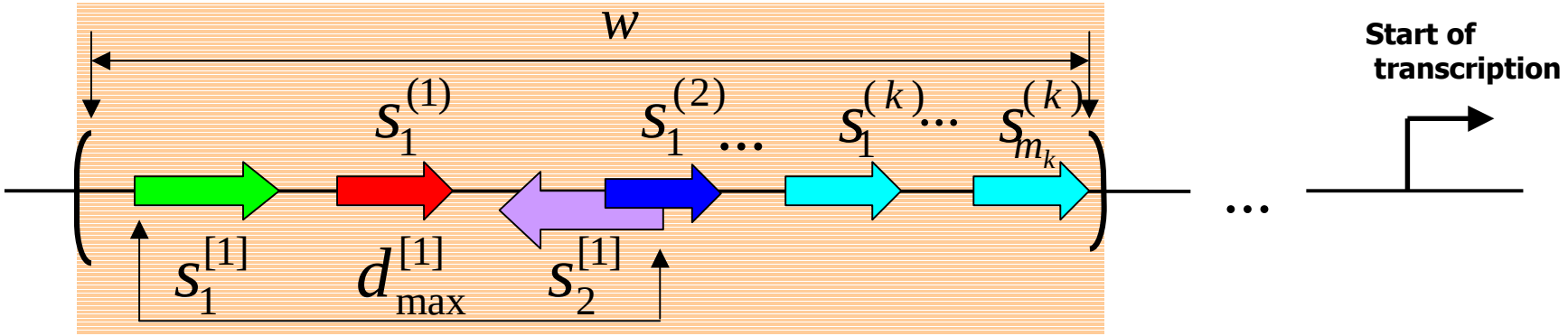
Finding (a) combination(s) of TFBS („Composite Modules“) that is/are characteristic for a given set of promoters.

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)

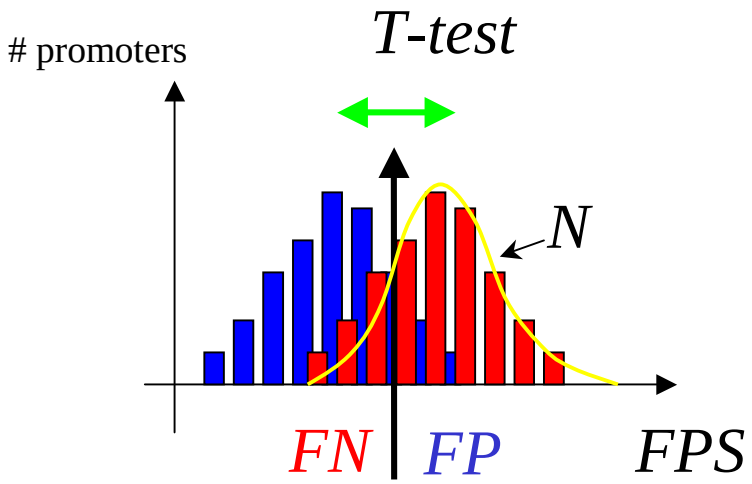
$$\begin{matrix} d_{\max}^{[1]} & d_{\max}^{[1]} & \cdots & d_{\max}^{[R]} \\ q_{cut-off}^{(1)} & q_{cut-off}^{(2)} & \cdots & q_{cut-off}^{(k)} \\ \phi^{(1)} & \phi^{(2)} & \cdots & \phi^{(k)} \end{matrix}$$

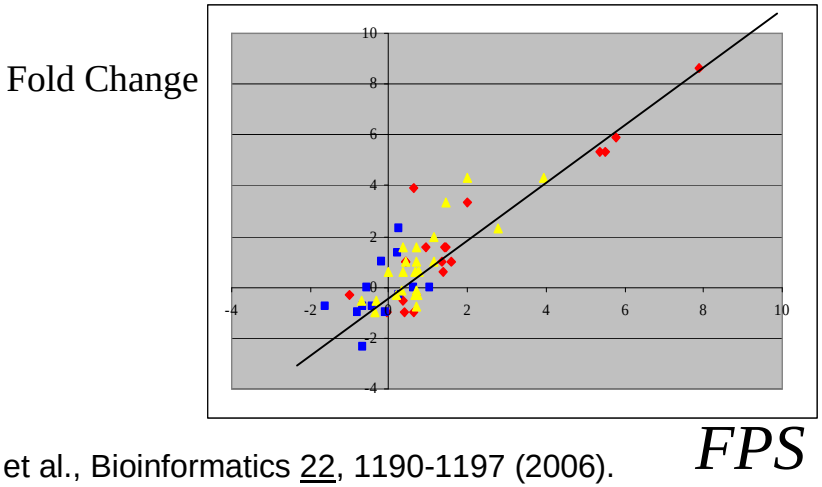
Parameters of the model to be estimated by a Genetic Algorithm

Promoter structure recognition

Fitness function of the Genetic Regression Algorithm (GRA)

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





R – linear regression

FN – false negatives

FP – false positives

T – T-test (difference between mean values)

N – normal likeness

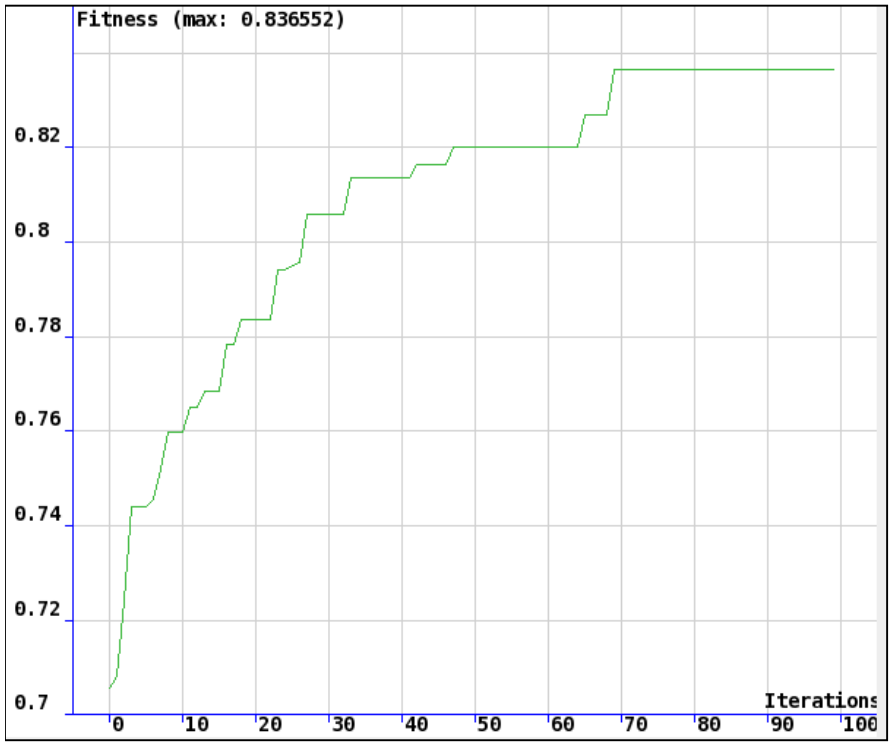
k – number of free parameters

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

Composite module analyzer (CMA)

Promoter model generation

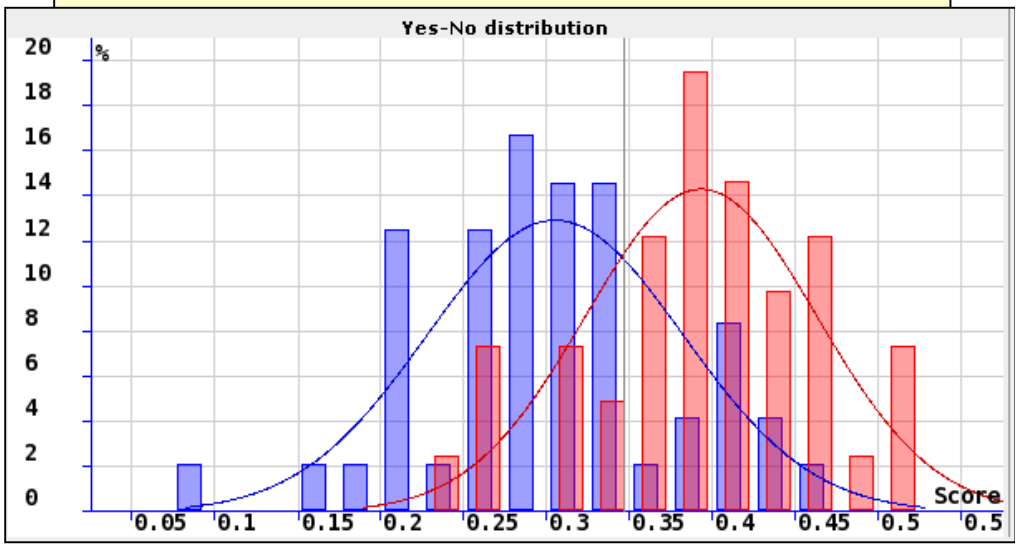
Increase of Fitness function with number of iterations



Composition of the promoter model

V$OCT1_Q5_01			V$SP1_01
V$AP2_Q3			V$EGR2_01
V$HIF1_Q5	or		V$FOXM1_01
V$USF_Q6			V$CREB_Q2_01
<-V$USF_C-> [3...30] <-V$AR_01->			<-V$GR_Q6-> [3...30] <-V$EGR_Q6->

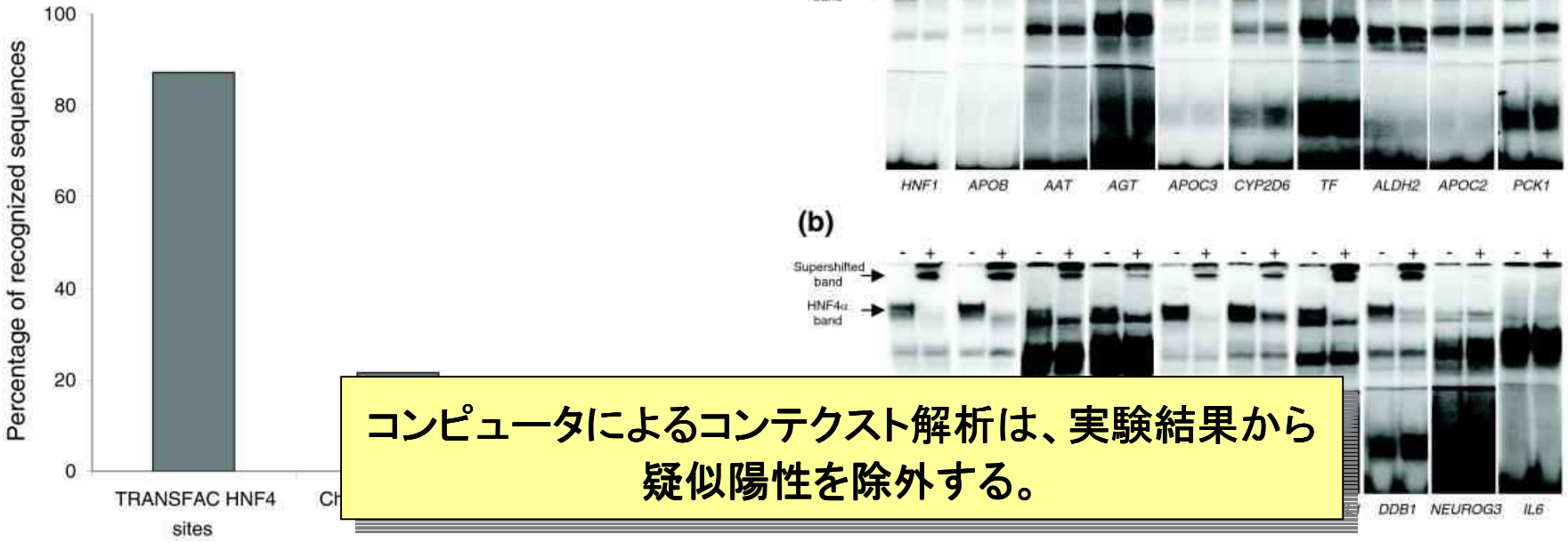
Sequences of YES and NO sets are well separated by the selected promoter model



Composite module analyzer (CMA)

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

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



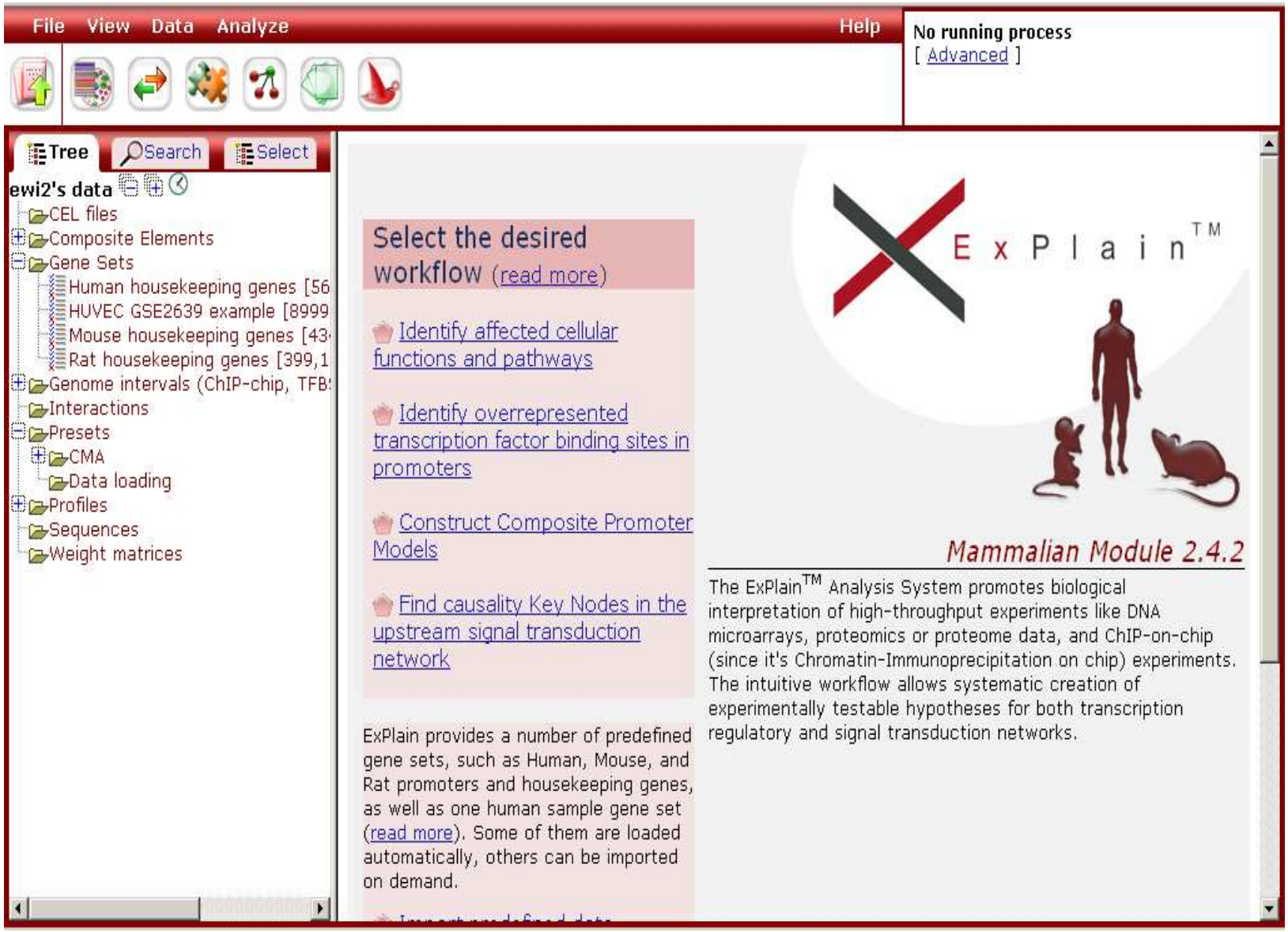
コンピュータによるコンテクスト解析は、実験結果から疑似陽性を除外する。

Computational context analysis can sort out experimental false positives.

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

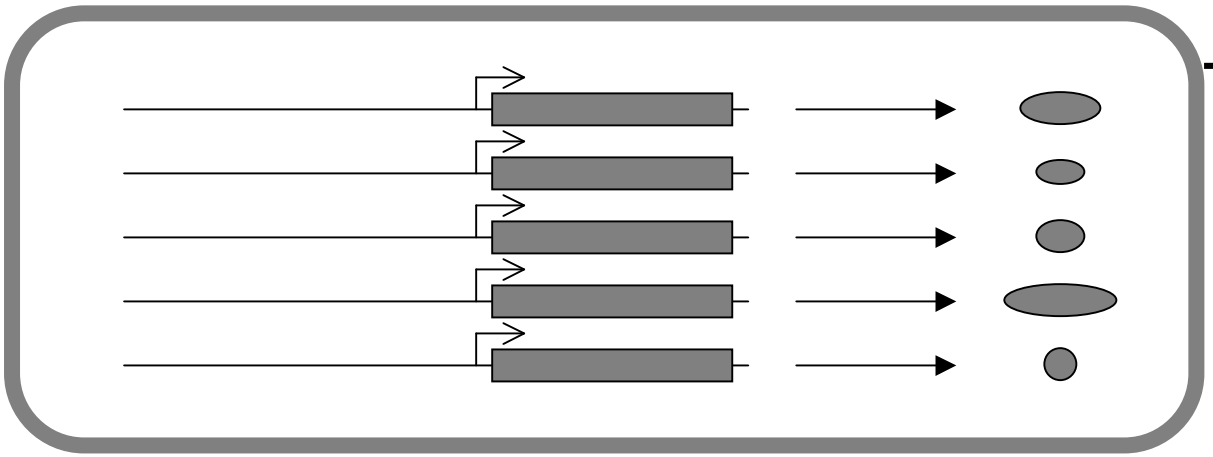
Promoter structure recognition

The ExPlain™ System



UMG

The ExPlain™ System



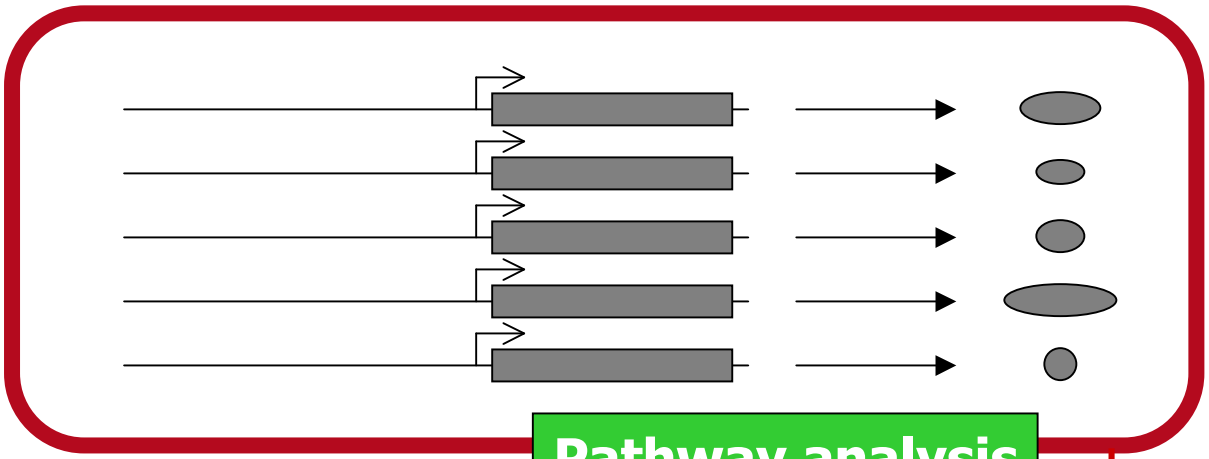
Disease terms

Deduce the gene products and map them to functional categories

Functional analysis

UMG

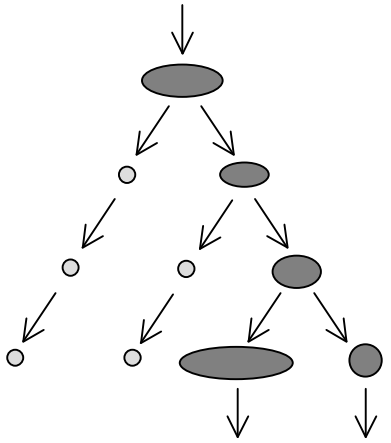
The ExPlain™ System

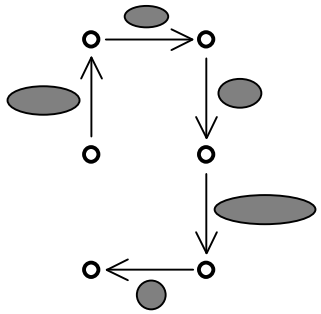


Pathway analysis

TRANSPATH

Downstream analysis:
map the induced gene products to the network of regulatory pathways





Biological effects

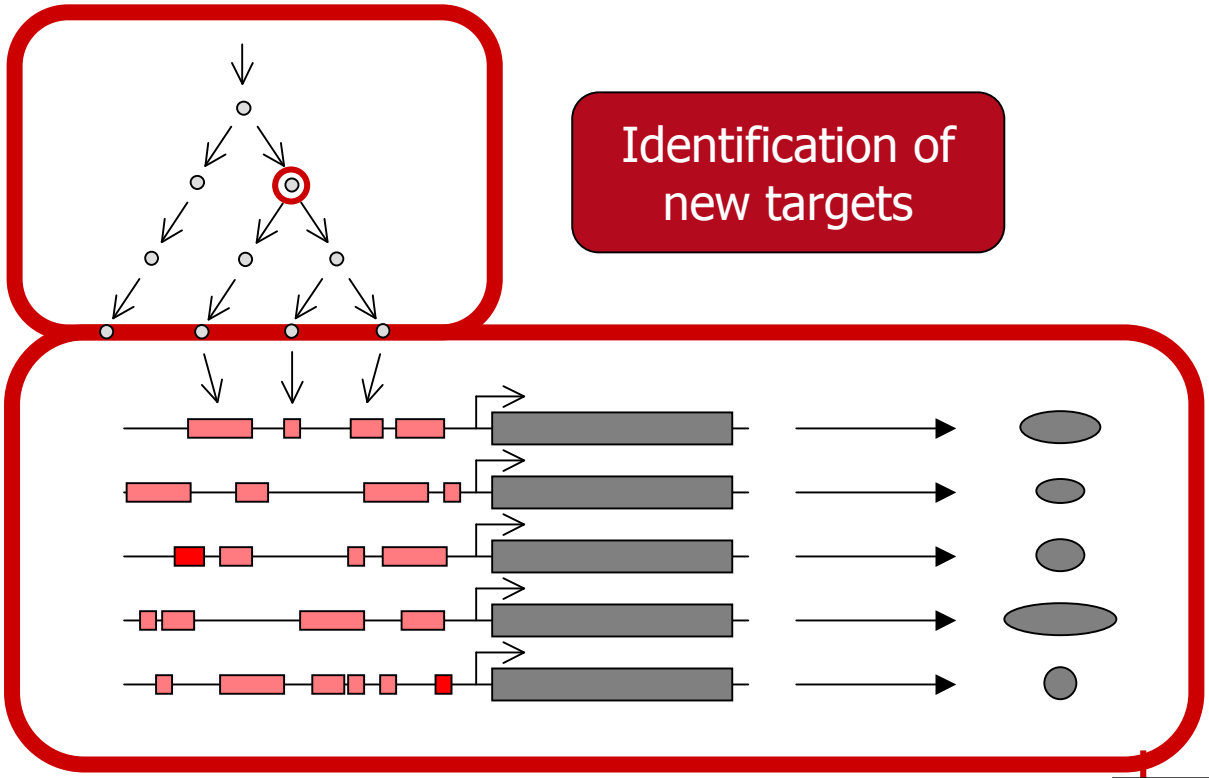
Biochemical effects

Promoter structure recognition

The ExPlain™ System

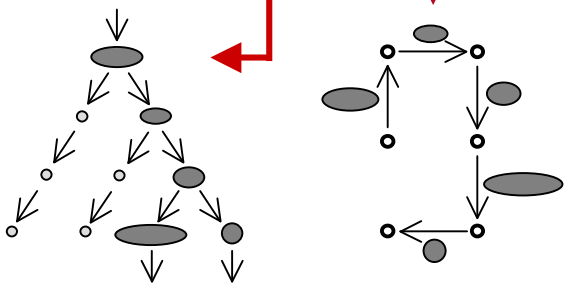
Pathway analysis

Promoter analysis



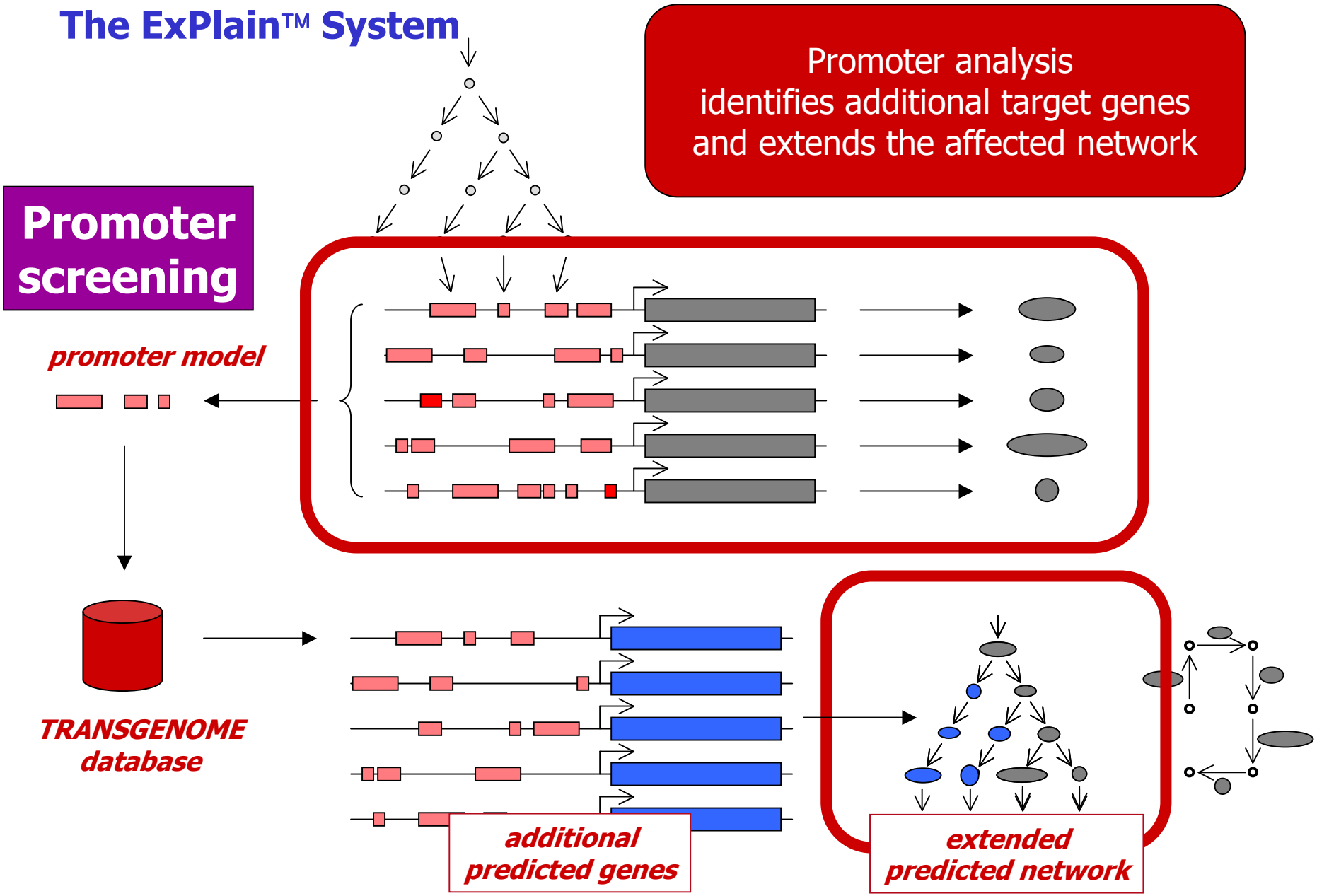
Identification of new targets

TRANSPATH



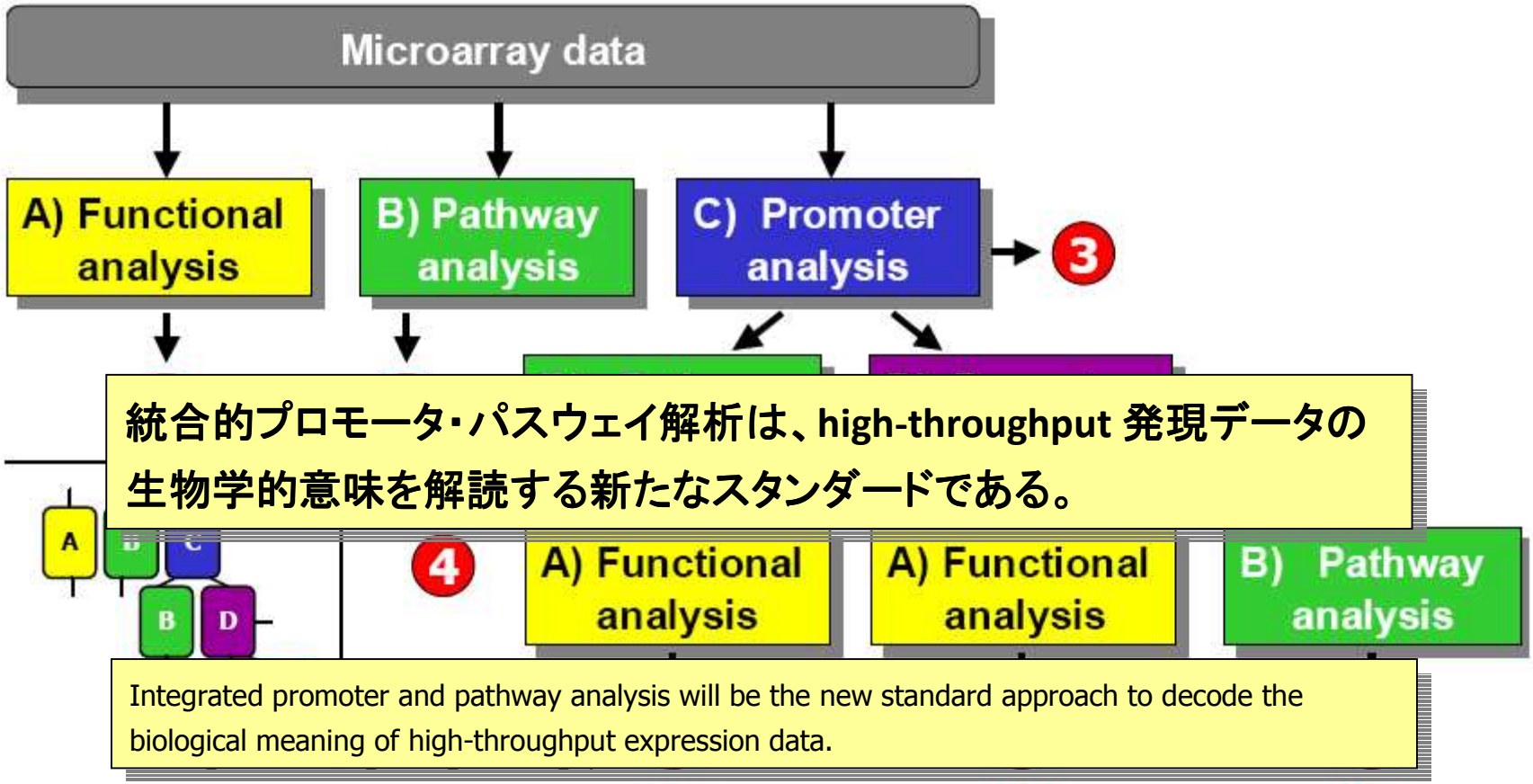
Reasoning
of experimental findings:
promoter analysis of induced genes
connected to network mapping

Promoter structure recognition



UMG

The ExPlain™ System: workflow example



Goal #5:

新規転写因子の DNA 結合特異性を予測するためにタンパク質･DNA 相互作用機構の解明

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

Zinc-finger 構造：
バクテリアの Helix-turn-helix 構造の真核生物版？

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	α-globin		CTF ?	51
	human	hsp70 ?			79
	rat	LTR (HSV)			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	α_2(I) collagen			20
CCAAT box bind.protein	mouse, rat, human(HeLa)	α_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[a] (ref.)	position	finger sequences
TFIIIA	1	M G E K A L
(210, 211)	7	P V V Y K R Y I C S F A D C G A A Y N K N W K L Q A H L C K H
	38	T G E K P F P C K E E G C E K G F T S L H H L T R H S L T H
	68	T G E K N F T C D S D G C D L R F T T K A N M K K H F N R F H
	99	N I K I C V Y V C H F E N C G K A F K K H N Q L K V H Q F S H
	130	T Q Q L P Y E C P H E G C D K R F S L P S R L K R H E K V H
	160	A G Y P C K K D D S C S F V G K T W T L Y L K H V A E C H
	189	Q D L A V C D V C N R K F R H K D Y L R D H Q K T H
	215	E K E R T V Y L C P R D G C D R S Y T T A F N L R S H I Q S F H
	247	E E Q R P F V C E H A G C G K C F A M K K S L E R H S V V H
Xfin	103	S A K K S H I C S H T G K L F S C T A A V V R H Q R M H
(212)	131	Q L Q K S H H C P H C K K S F V Q R S D F I K H Q R T H
	159	T G E R P Y Q C V E C Q K K F T E R S A L V N H Q R T H
	187	T G E R P Y T C L D C Q K T F N Q R S A L T K H R R T H
	215	T G E R P Y R C S V C S K S F I Q N S D L V K H L R T H
	243	T G E K P Y E C P L C V K R F A E S S A L M K H K R T H
	271	S T H R P F R C S E C S R S F T H N S D L T A H M R K H
	299	T E F R ...
	320	N V A S S P Y S C S K C R K T F K R W K S F L N H Q Q T H
	349	S R E K P Y L C S H C N K G F I Q N S D L V K H F R T H
	377	T G E R P Y Q C A E C H K G F I Q K S D L V K H L R T H
	405	T G E K P F K C S H C D K K F T E R S A L A K H Q R T H
	433	T G E K P Y K C S D C G K E F T Q R S N L I L H Q R I H

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

UMG

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

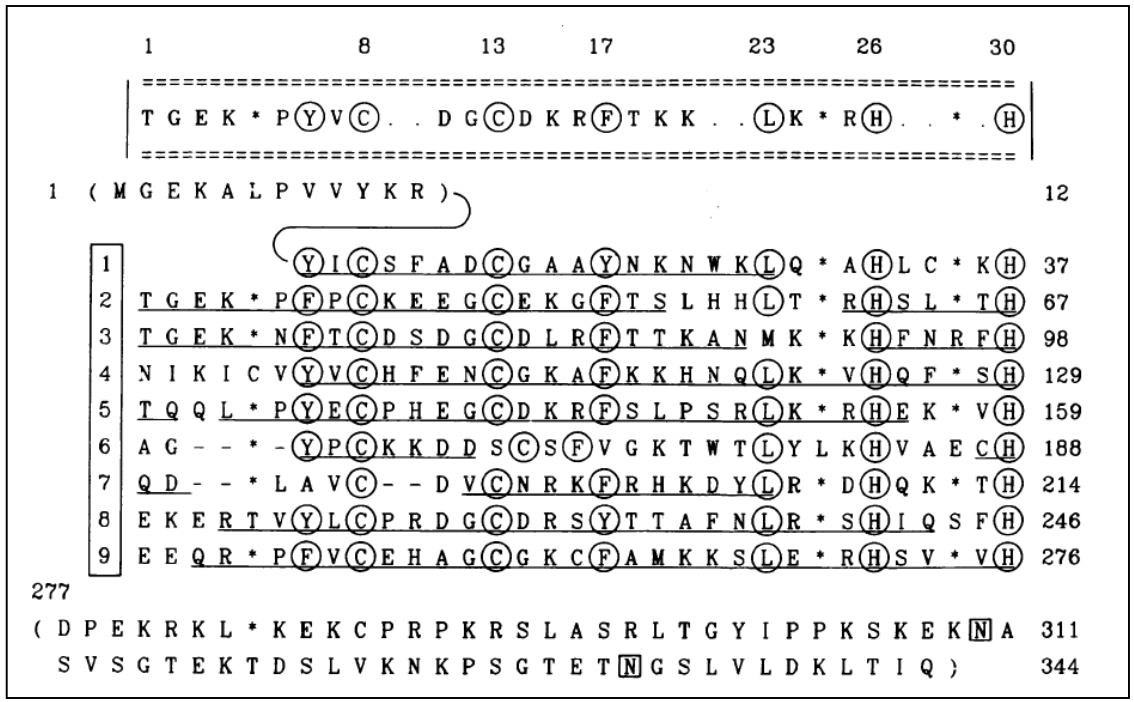
PDB entry
1CGP

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:

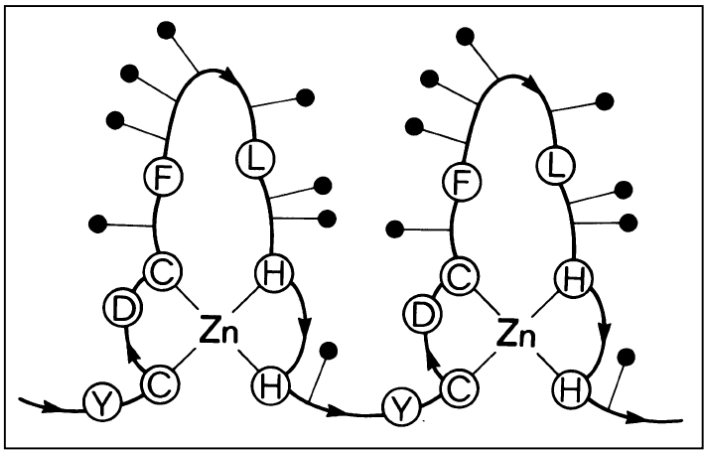
McKay, D.B., Steitz, T.A. (1981) Structure of catabolite gene activator protein at 2.9 A resolution suggests binding to left-handed B-DNA. Nature **290**:744-749.

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



TFIIIA で発見された9つの Cys_2-His_2 モチーフが、Zinc finger の正体である。

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

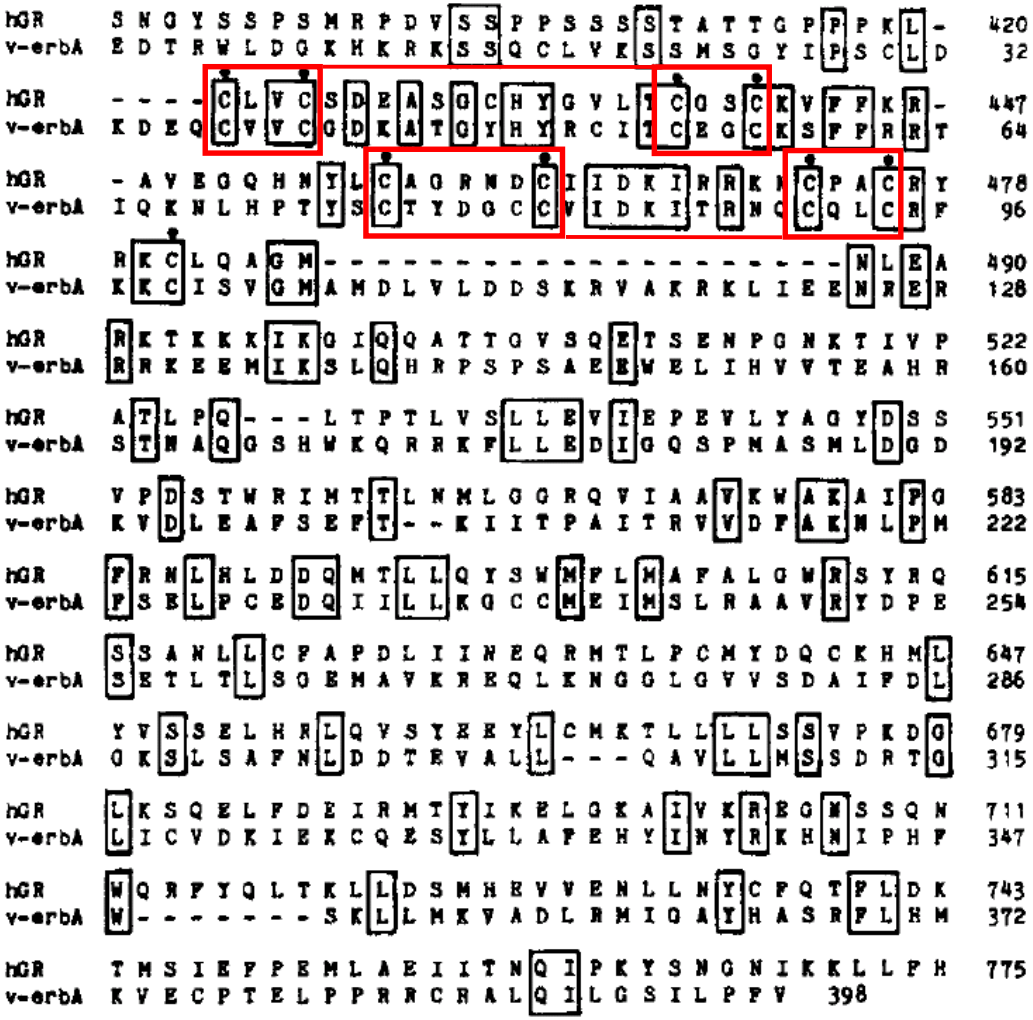
Foster, M.P., Wuttke, D.S., Radhakrishnan, I., Case, D.A., Gottesfeld, J.M., Wright, P.E. (1997) Domain packing and dynamics in the DNA complex of the N-terminal zinc fingers of TFIIIA. Nat.Struct.Biol. **4:** 605-608

UMG

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

GR zinc finger

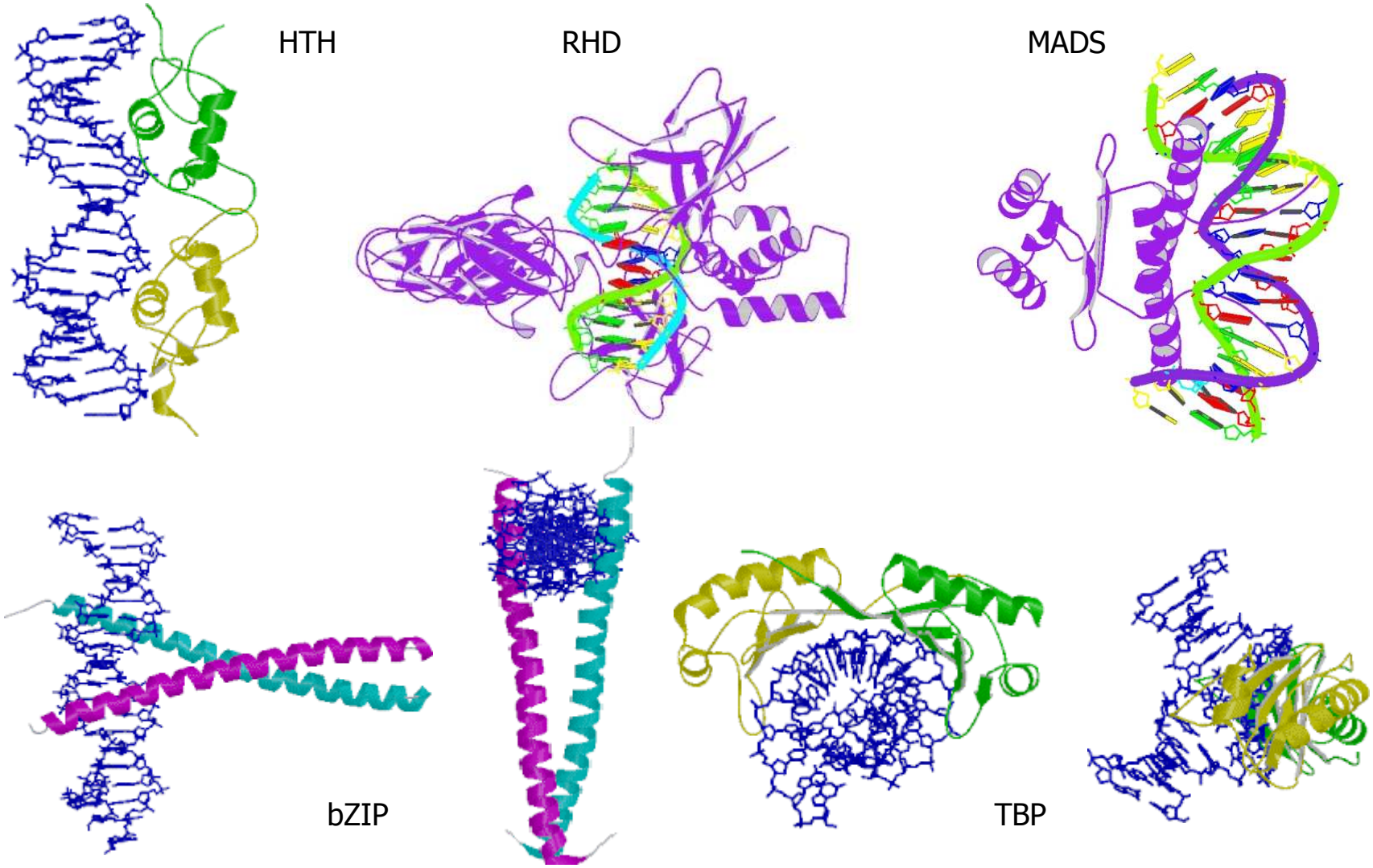
Similarity with TFIIIA zinc finger structure suggested for the first time.



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

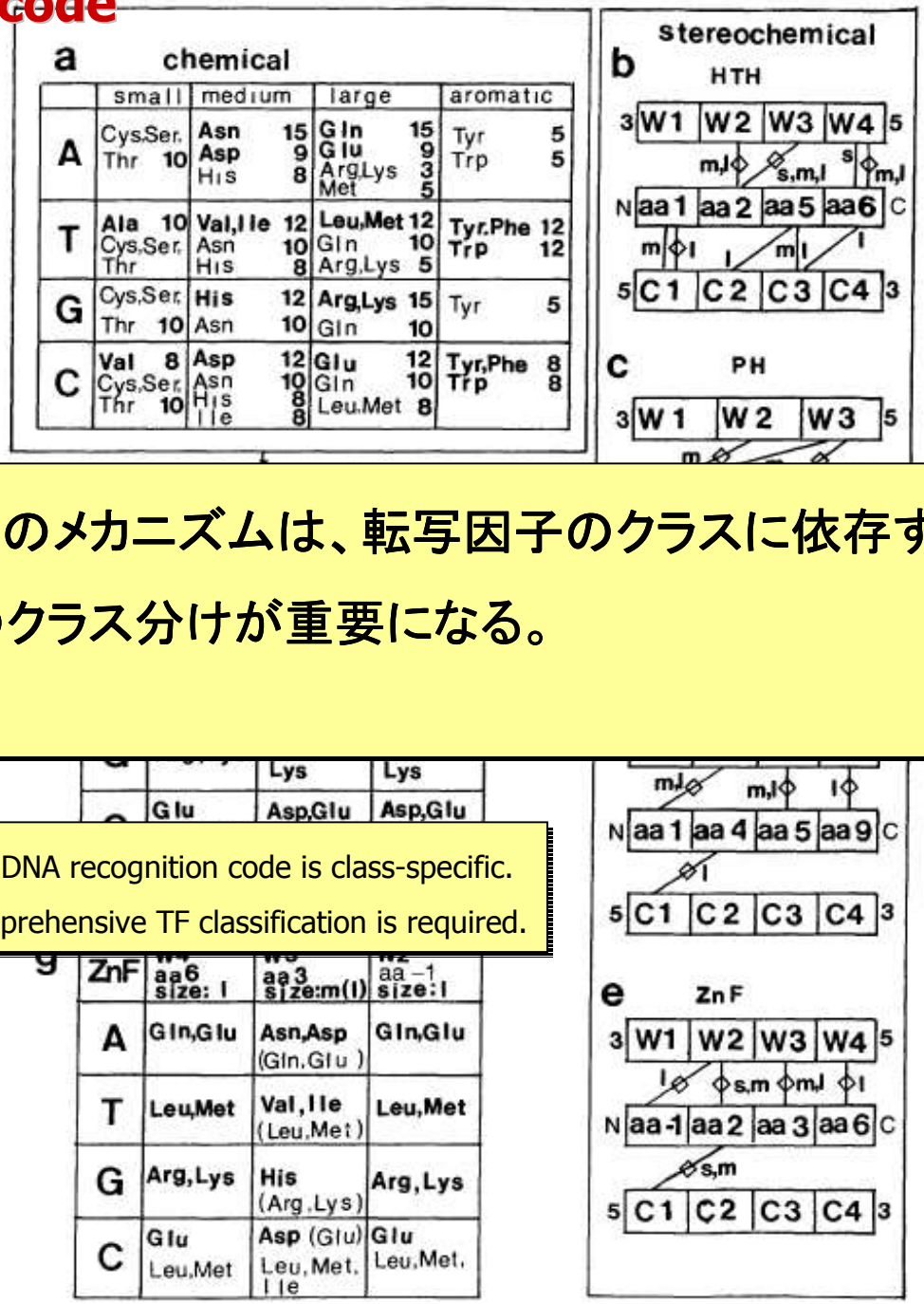
UMG

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



Protein-DNA interaction code

Protein-DNA recognition code?



タンパク質・DNA 相互作用のメカニズムは、転写因子のクラスに依存する。

故に、網羅的な転写因子のクラス分けが重要になる。

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-α, RAR-β
6	N.N.N.N.N.N	factor "species"	initiation/ splice/ processing variants	RAR-α1, RAR-α2

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

真核生物転写因子の **DNA** 結合領域は、少なくとも**4**つのスーパークラスに属する**30**のクラスに分類できる。

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

3.1 Homeo domain
3.5 Tryptophan clusters
3.6 TEA domain

最近の研究では、約**13**のスーパークラスに分類できると考えられる。殆どのものは、**4**つの旧スーパークラスから細分化されたものである。

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

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

Protein-DNA interaction code

Transcription factor classification

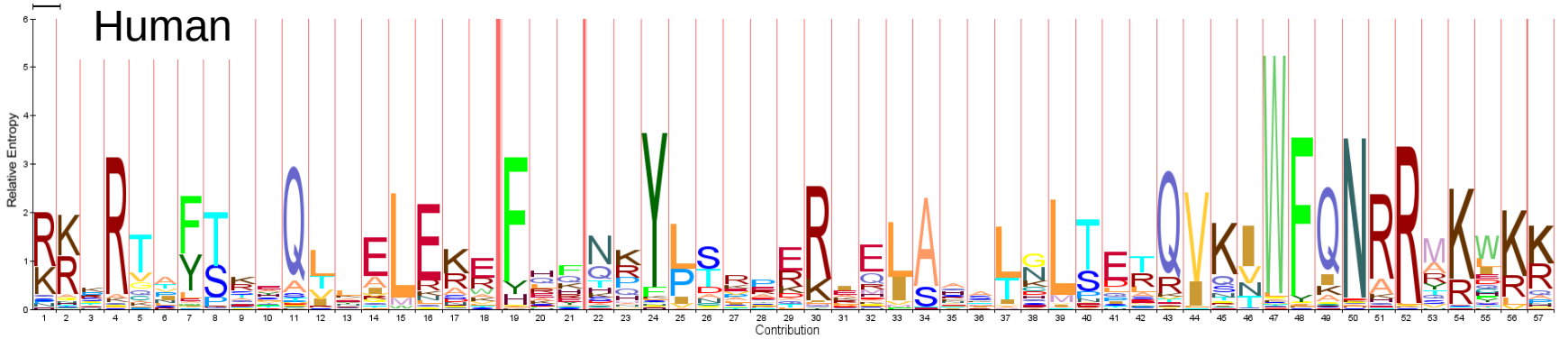
1 Basic Domains
2 Zinc-coordinating DNA-binding domains
2.1 Cys4 zinc finger of nuclear receptor type
2.1.1 Steroid hormone receptors (NR3)
2.1.1.1 Corticoid receptors (NR3C)
2.1.1.1.1 **GR (NR3C1):** GR GR GR GR GR GR GR
2.1.1.1.1.1 **GR-alpha :** GR-alpha
2.1.1.1.1.2 **GR-beta:** GR-beta
2.1.1.1.2 **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

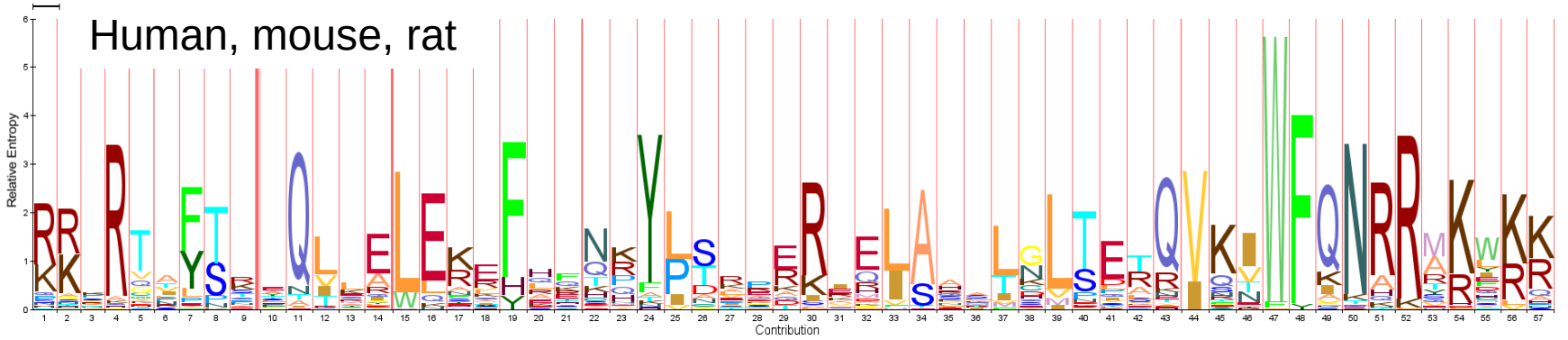
HomeoboxMensch

Human



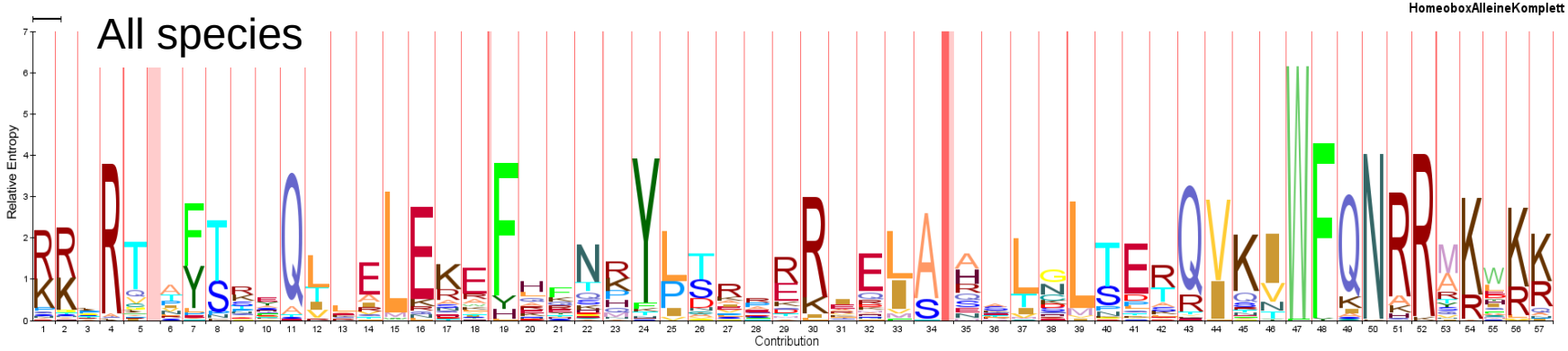
HomeoboxMenschMausRatte

Human, mouse, rat

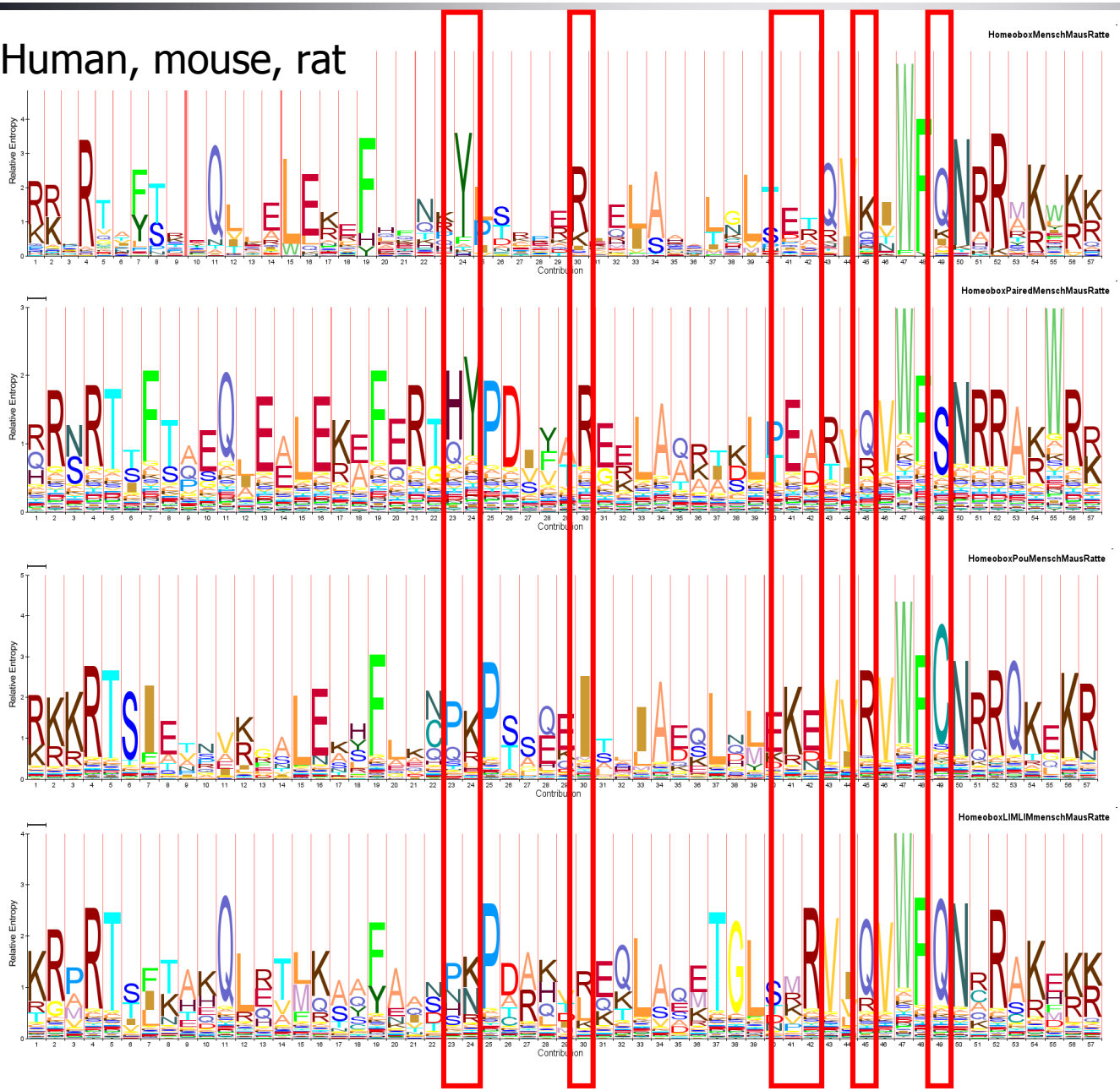


HomeoboxAlleineKomplett

All species



Protein-DNA interaction code



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

次のステップは、転写因子のDNA 結合ドメイン全体または、いくつかの重要アミノ酸残基のどちらが、特異的な相互作用に重要であるのかを調べることである。

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:

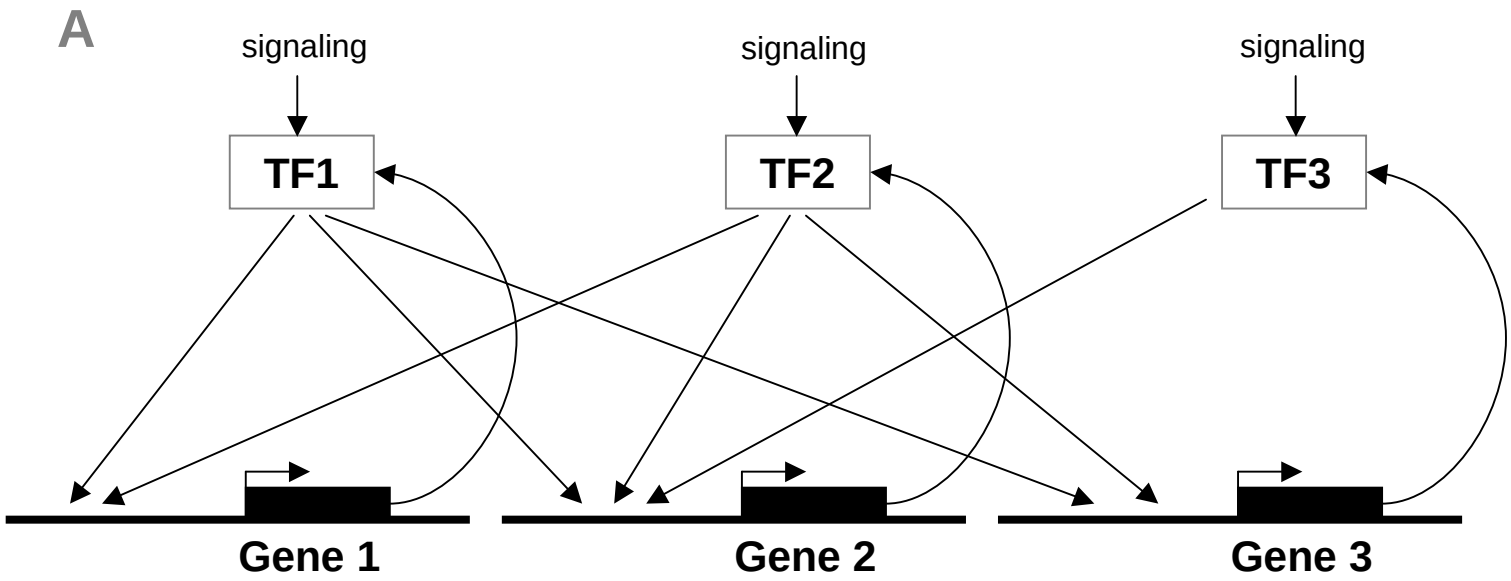
細胞や生物における完全な転写ネットワークの解析と利用

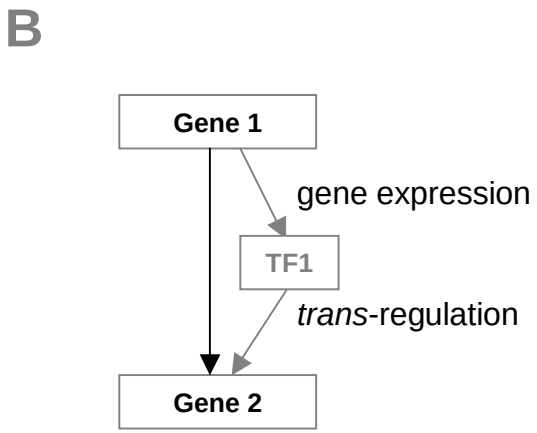
Goal #6:

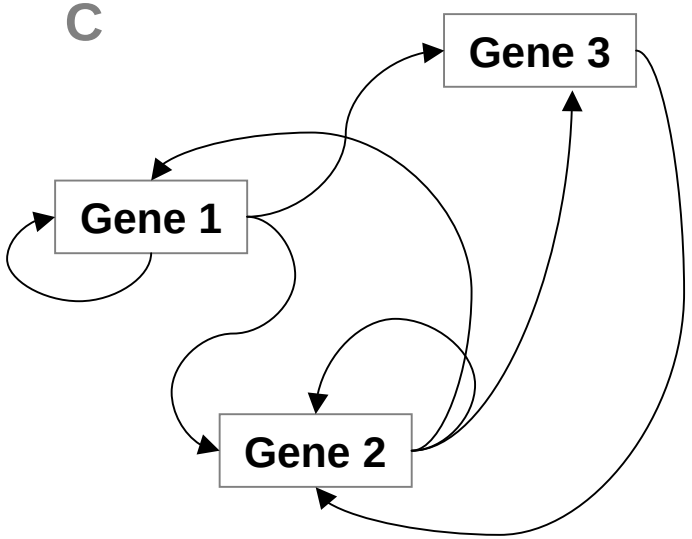
Re-engineering and analyzing the complete transcription network of a cell / an organism.

Transcription network

Definition







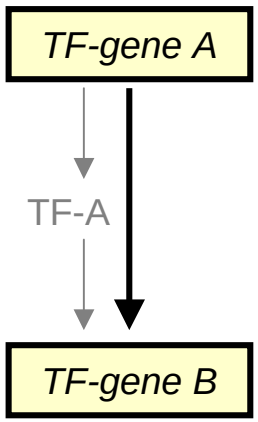
Transcription network

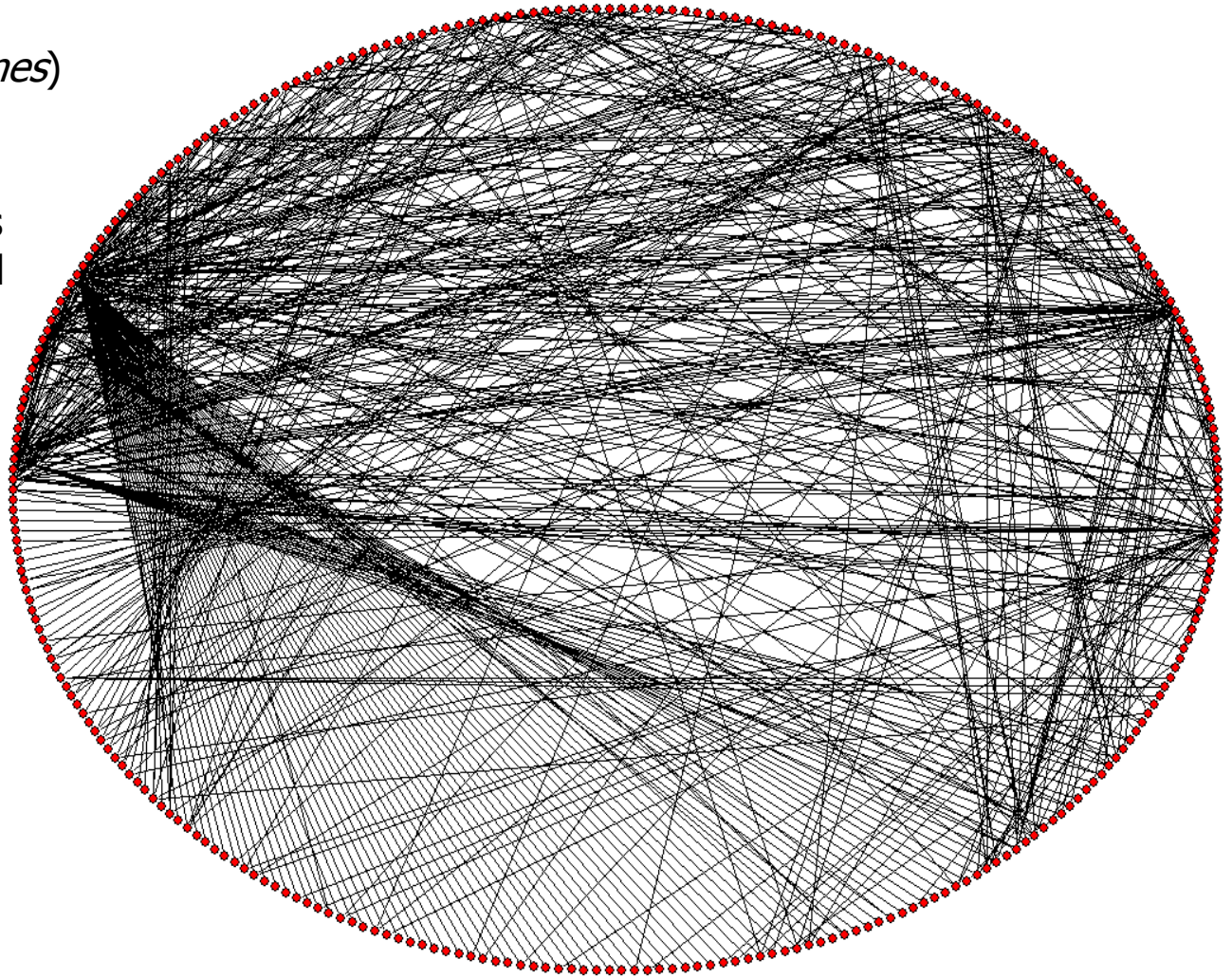
Mammalian network of transcription factor genes (TFG_RN)

298 vertices (*TF-genes*)

638 edges

Each edge combines gene expression and *trans*-regulation





UMG

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.

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

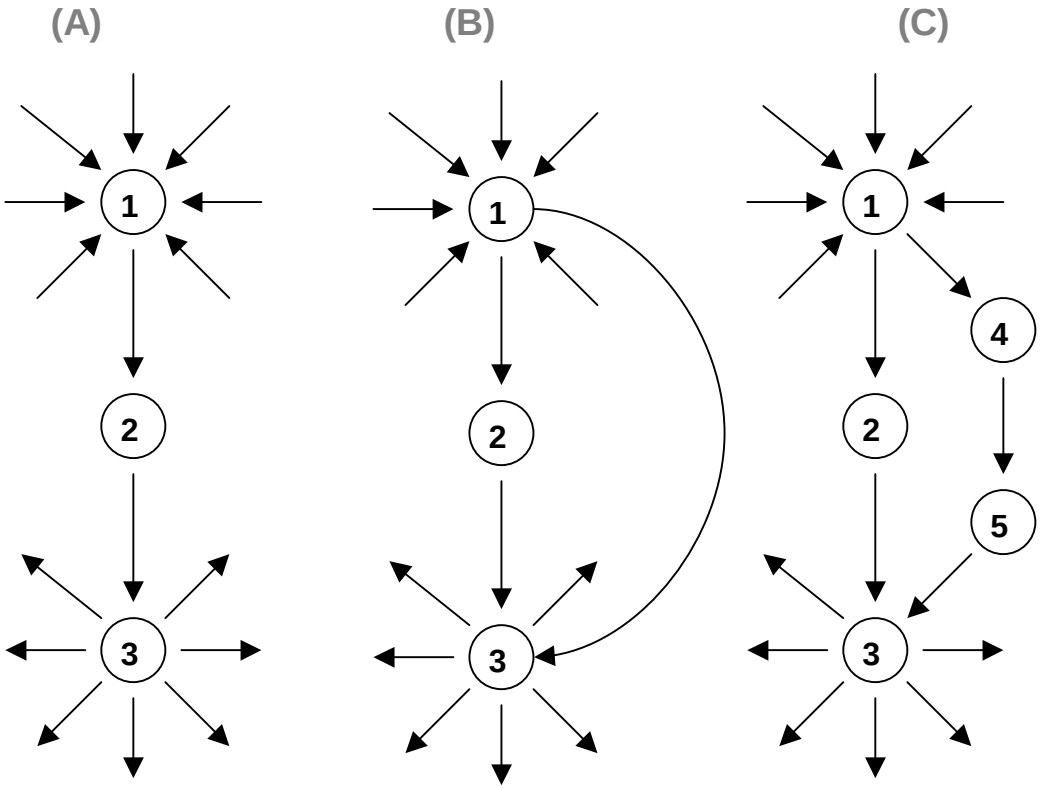
Fraction of all-pairs connections for which vertex ***v*** is crucial: there is no other parallel path(s) between vertices of such a pair and these vertices will not be connected anymore if vertex ***v*** is deleted.

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

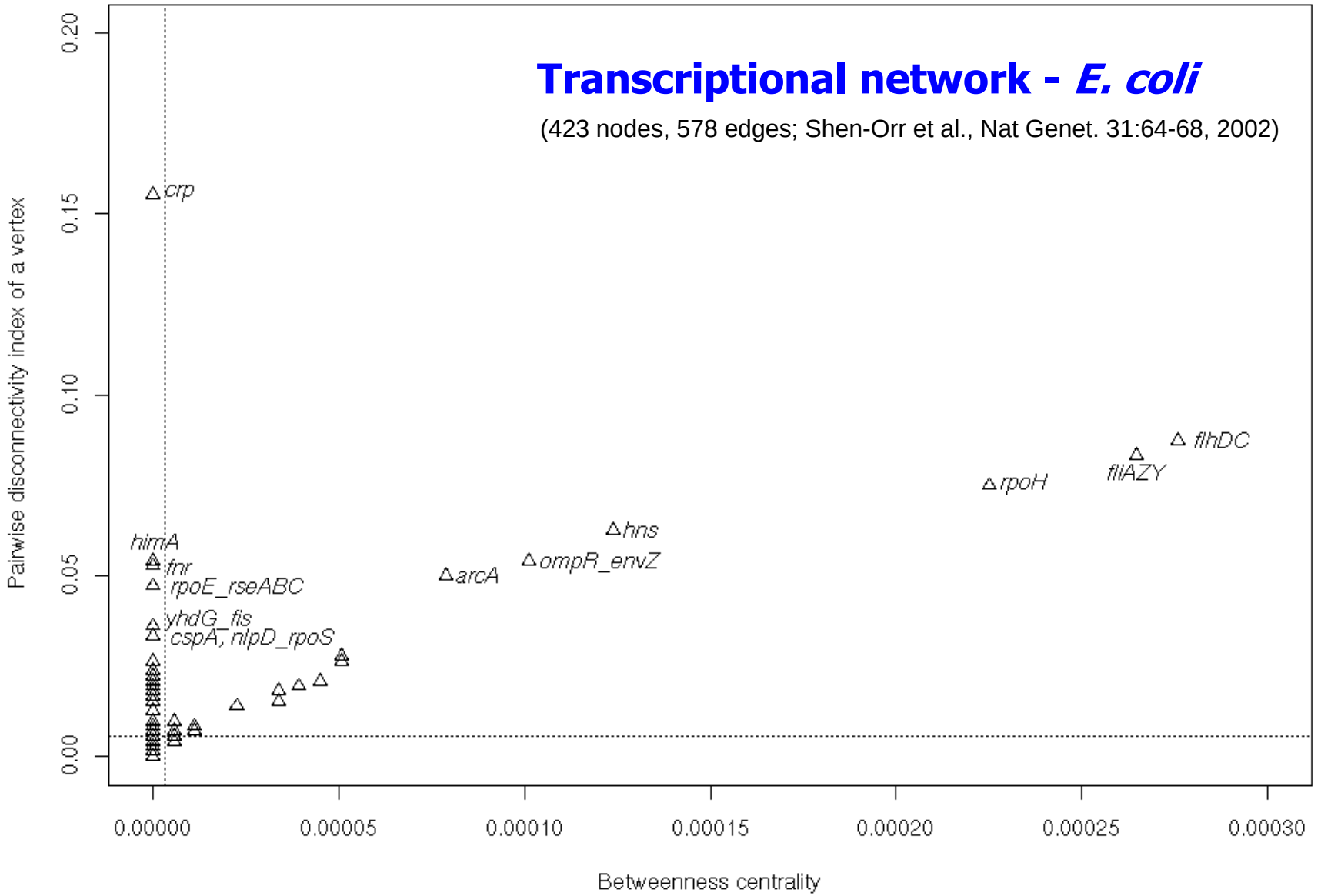
Disconnectivity sensitive towards bypasses



B(*v*2):	high	low	high
Dis(*v*2):	high	low	low

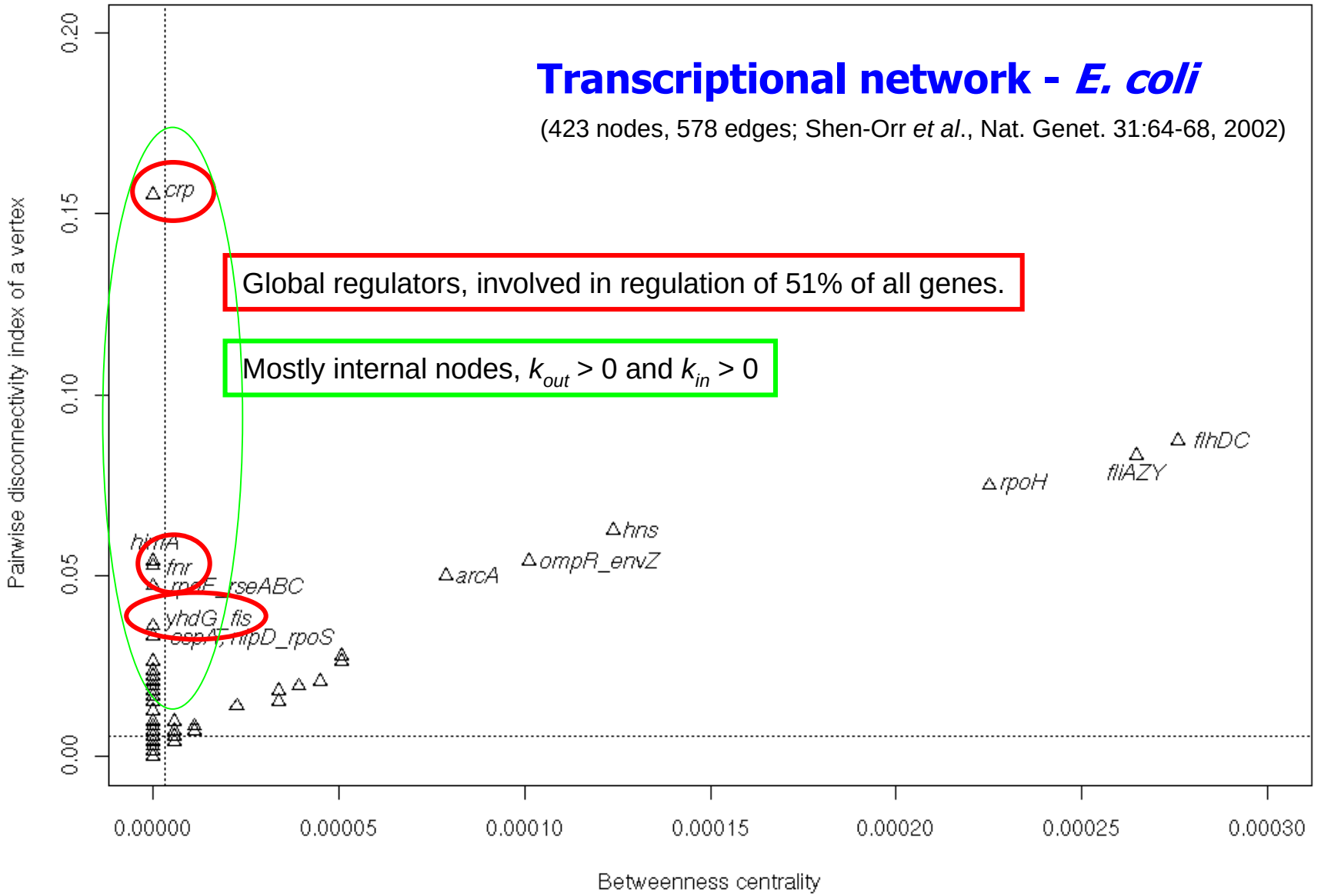
Potapov et al., BMC Bioinformatics, 9, 227 (2008)

Transcription network



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

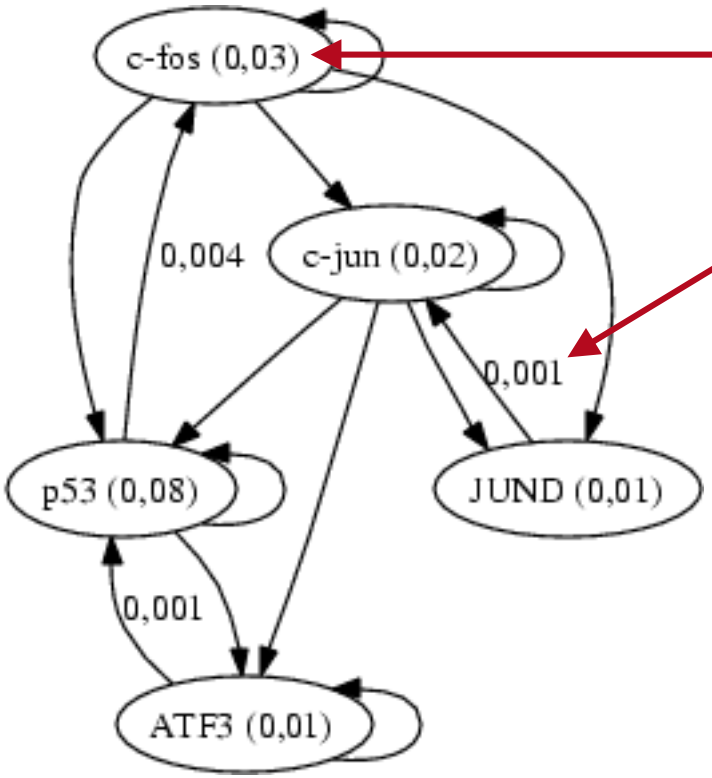
Transcription network



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

Transcription network

‘p53’ strongly connected sub-graph in TFG_RN



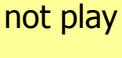
Disconnectivity indices

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 D

		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
	102	-	-	-	1	15.91	0.028	3	0.35	0.0224
										37
										51
										58
										09
	238	-	-	-	-	-	-	1	-	0.0073

転写因子ネットワークで、解析した全てのネットワークで見られる **3-vertex motif** は、**feed-forwad loop** である。

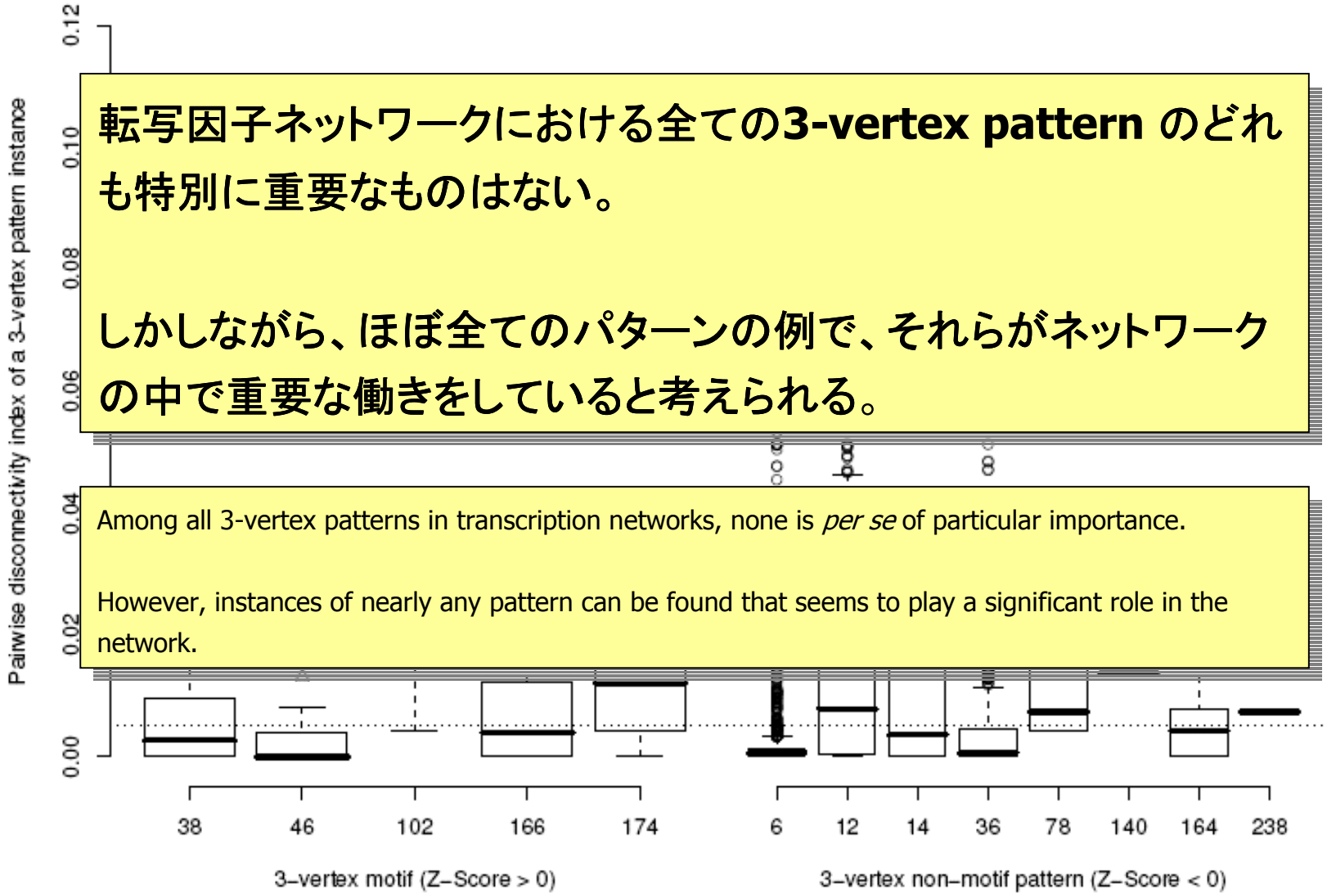
しかしながら、このモチーフそのものは、モチーフのないパターンと比べて、ネットワークの結合においてより重要な機能はしていない。

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



転写因子ネットワークにおける全ての**3-vertex pattern** のどれも特別に重要なものはない。

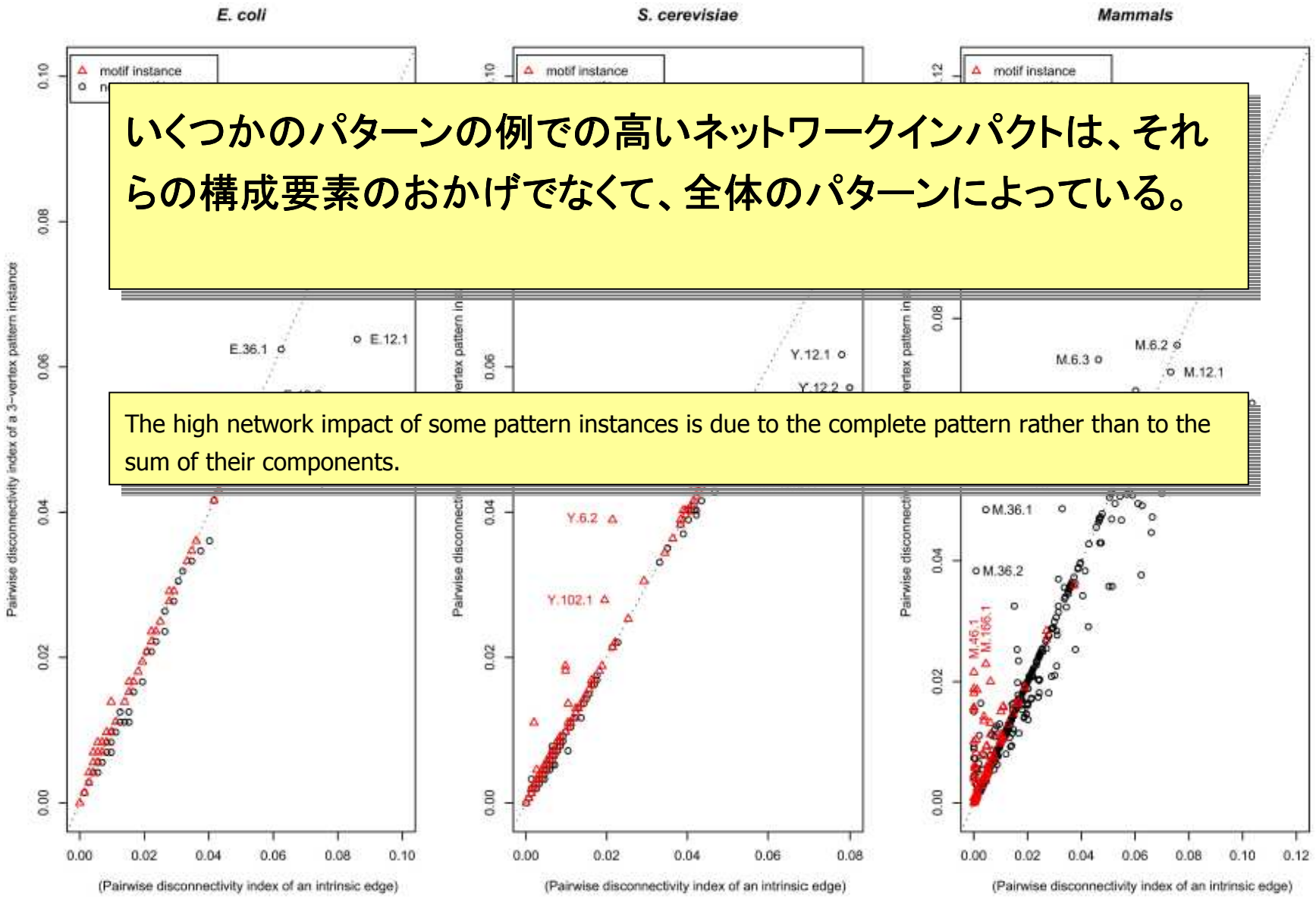
しかしながら、ほぼ全てのパターンの例で、それらがネットワークの中で重要な働きをしていると考えられる。

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



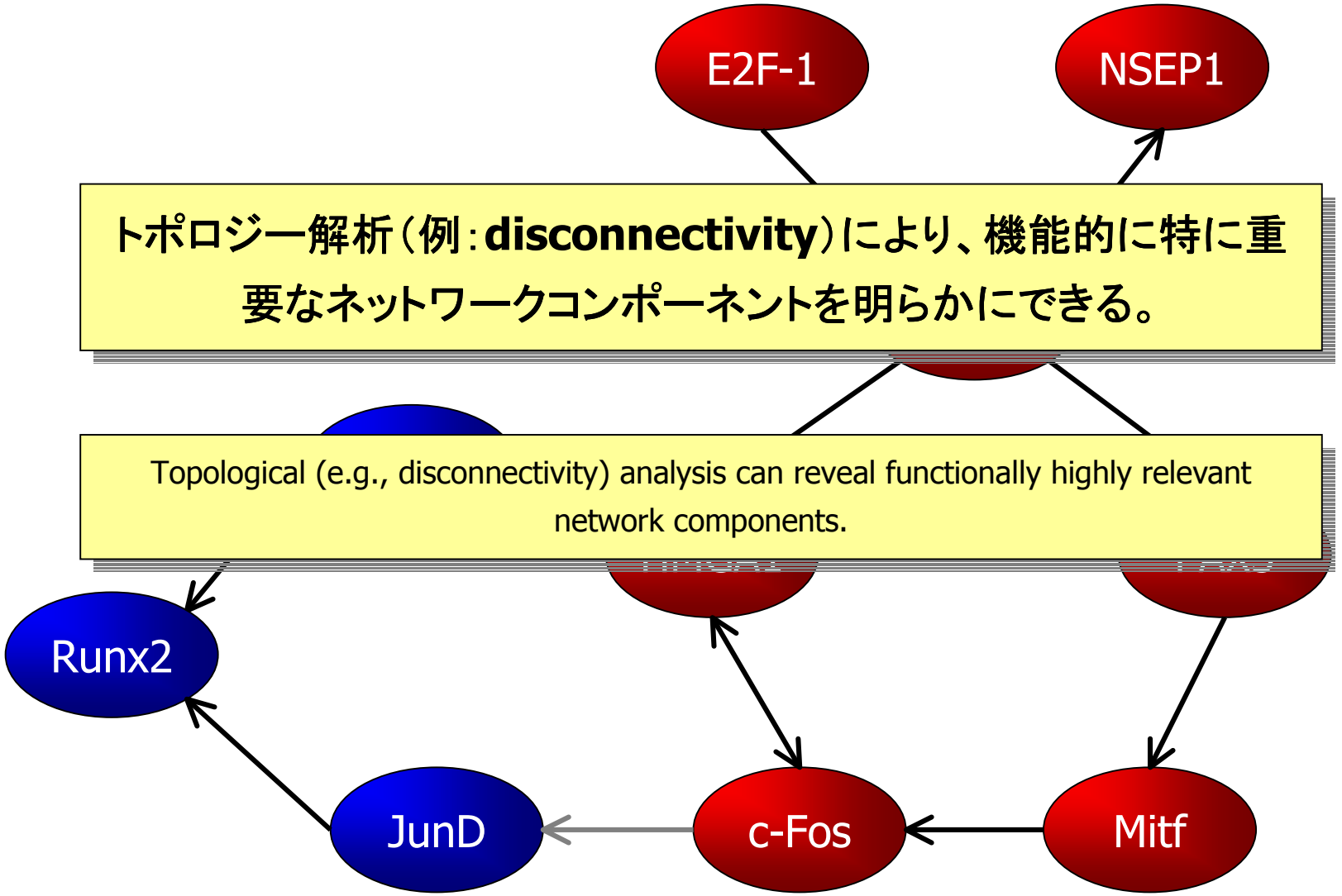
いくつかのパターンの例での高いネットワークインパクトは、それらの構成要素のおかげでなくて、全体のパターンによっている。

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)

UMG

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



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.