Bioinformatics of Gene Regulation

30 Years TRANSFAC

Göttingen, March 07-09, 2018

UNIVERSITÄTSMEDIZIN GÖTTINGEN : UMG

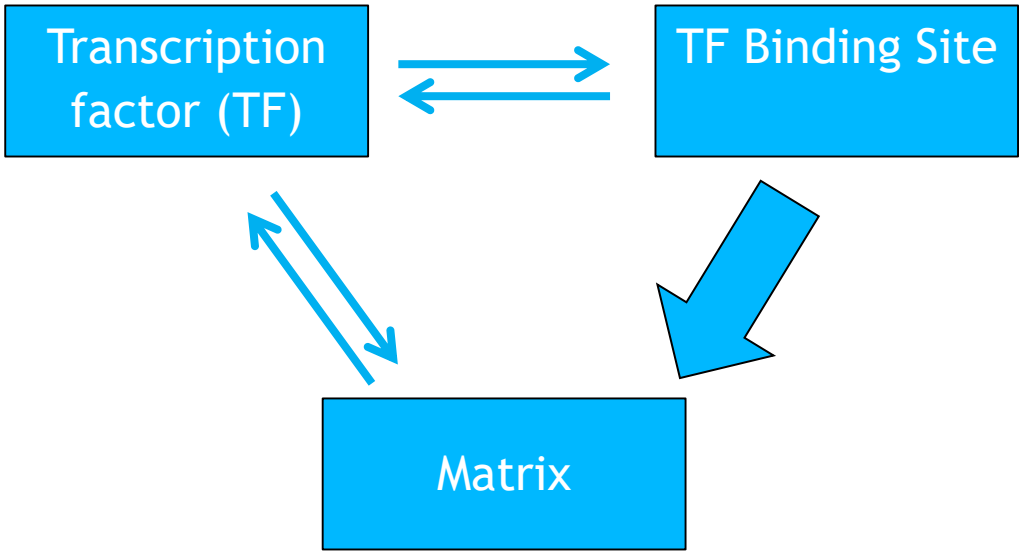
Georg August University, Göttingen, Germany

Volume 1(
Compilat
Edgar W
Gesellsch
Received

gene, gene product	species/tissue(a)	protein intera
α-actin	chicken/rat myocytes	-83 to -78
	/rat non-myocytes	"
β-actin	rat/HeLa	
actin 5C	Drosophila	-38 to +34
actin (cytoskeletal)	Xenopus laevis /HeLa	-94 to -75 (SF
Ad2MLP (adenovirus 2 major late prom.)	adenovirus/HeLa	-68 to -49
		-63 to -52
		-50 to -10
		-40 to +35
		-34 to -22
Adh (alcohol dehydrogenase)		
-distal prom.	Drosophila	- 85 to - 47
-proxim.prom.		-269 to -229
		-151 to -105
		- 98 to - 77
Adh1 (alcohol dehydrogenase)	maize	-190 to -186
		-145 to -138 (
		-120 to -117 (
		-108 to -100
ADH2 (alcohol dehydrogenase)	yeast	-257 to -216
albumin	rat/liver	-156 to -141
		-126 to -107
		-105 to - 89
		- 72 to - 35
A-MuLV (amphotrop. murine leukemia virus)	/F9, PCC4	-87 to -59
aP2 (adipocyte P2)	mouse/adipocytes	-124 to -108
BPV	/rec	7613 to 7639

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 (MSV)			125
	rat	tk (HSV)			125
CBF	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
PMID: 328


Transcription factor (TF)
TF Binding Site
Matrix


---------->V$FOXD3_01 (0.95)
---------->V$FOX_Q2 (0.95)
tctattatttatttttaaaagattttatttttatttttaatt
---------->V$HEB_Q6 (1.00)
gagtataggtggttgtgagccagctgatgtaggtgatgcc


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


...gcccTACGTGCTgtctca...
human epo gene 3 enhancer

Tab. 4: Base frequencies around Spl sites

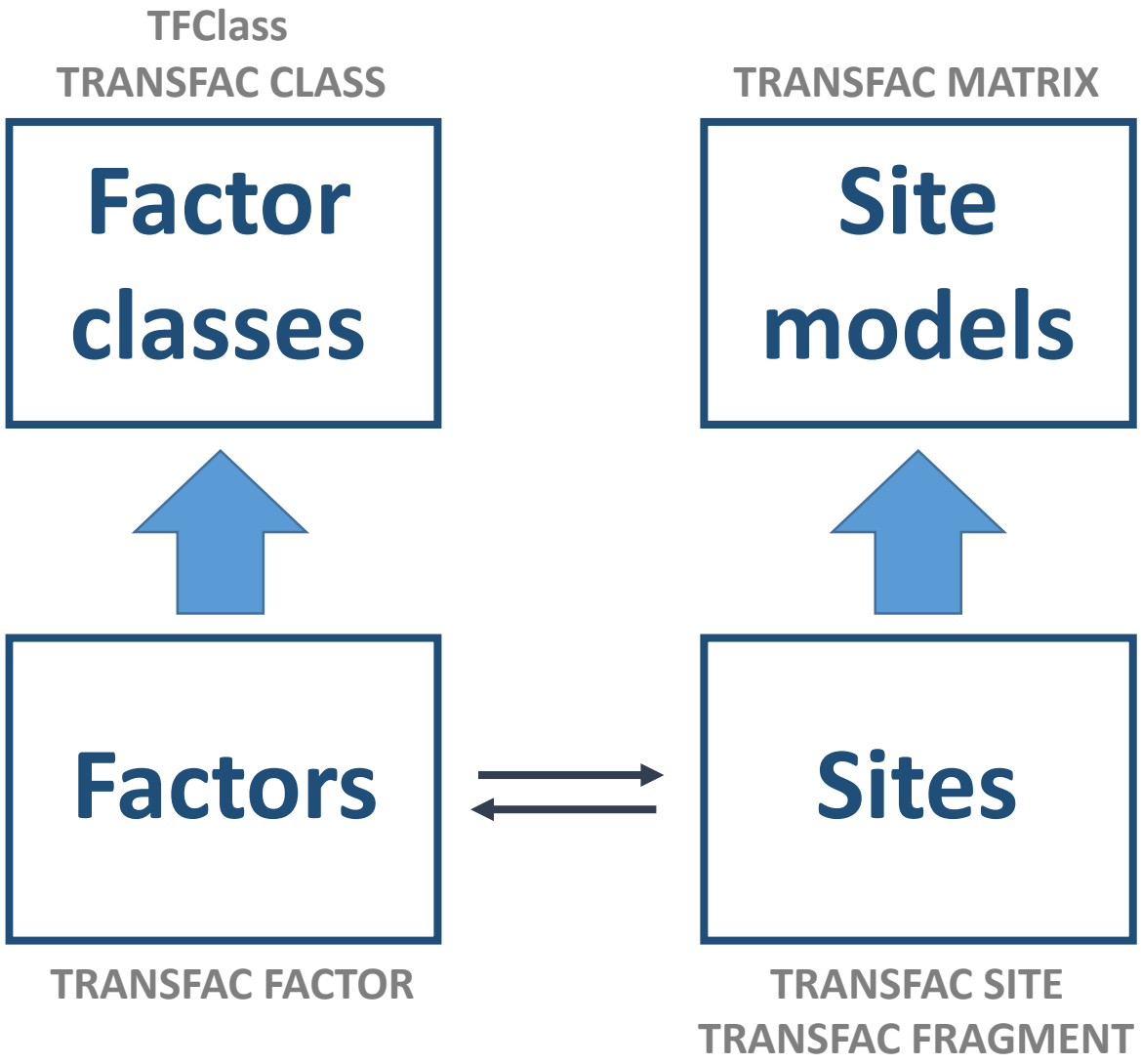
A:	5	4	6	7	5	8	5	8	11	2	3	0	0	0	2	0	9	7	2	3	6	6	7	4	4	7
C:	9	14	19	10	14	9	9	7	11	3	0	0	0	40	0	0	2	1	28	14	8	7	6	5	16	8
G:	21	17	11	18	17	21	19	14	11	26	37	41	41	1	39	40	27	31	4	10	12	19	21	28	18	18
T:	6	6	5	6	5	3	8	12	8	10	1	0	0	0	0	1	3	2	7	14	15	9	7	4	3	8
	g	g	c	g	g	g	g	g	a c g	g	G	G	G	C	G	G	g	g	c	c t	t	g	g	g	g	g

Wingender, E., Heinemeyer, T. and Lincoln, D., **Regulatory DNA sequences: predictability of their function.**
Genome Analysis - From Sequence to Function; BioTechForum- Advances in Molecular Genetics (J. Collins, A.J. Driesel, eds.) 4, 95-108 (1991)

```
ID   HS$6-16_1          STANDARD;  DNA;  14 BP.
XX
AC   R00001;
XX
DT   29-JUL-1990   (DATA ENTRY)
XX
DE   6-16.
XX
OS   HUMAN (HOMO SAPIENS, MAN, HOMME, MENSCH).
OC   EUCARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMALIA;
OC   EUTHERIA; PRIMATES.
XX
RN   [1]  (aa)
RA   Dale T.C., Ali Imam A.M., Kerr I.M., Stark G.R.;
RT
RL   Proc. Natl. Acad. Sci. USA 86:1203-1207(1989).
RN   [2]  (a)
RA   Porter A.C.G., Chernajowski Y., Dale T.C., Gilbert C.S.,
RA   Stark G.R., Kerr I.M.;
RT
RL   EMBO J. 7:85-92(1988).
RN   [3]  (ba)
RA   Dale T.C., Rosen J.M., Guille M.J., Lewin A.R., Porter A.G.C.,
RA   Kerr I.M., Stark G.R.;
RT
RL   EMBO J. 8:831-839(1989).
XX
CC   EWI.
CC   Data edited (30-JUL-1990) by Thomas Heinemeyer.
XX
DR   CELLINE       human/HeLA+IFN; Bristol 88+; HFF+.
DR   METHODS       3, 4a, 4b, 4d, 1f.
XX
KW   Protein interacting regions.
XX
FH   Key        From     To       Binding factor
FH
FT   TRANSFAC   -127     -89       E factor.
SQ   SEQUENCE    14 BP; 7 A; 1 C; 4 G; 2 T;
     gGGAAAaTGA AACT
```

Fig. 4: The first entry of the transcription factor database.

Wingender, E., Heinemeyer, T. and Lincoln, D., **Regulatory DNA sequences: predictability of their function.**
Genome Analysis - From Sequence to Function; BioTechForum- Advances in Molecular Genetics (J. Collins, A.J. Driesel, eds.) 4, 95-108 (1991)



Wingender, E., Kel, A., Krull, M., **Transcription factor databases.** Encyclopedia of Bioinformatics and Computational Biology (M. Cannataro et al., eds.), Chapter 20216 (2018)

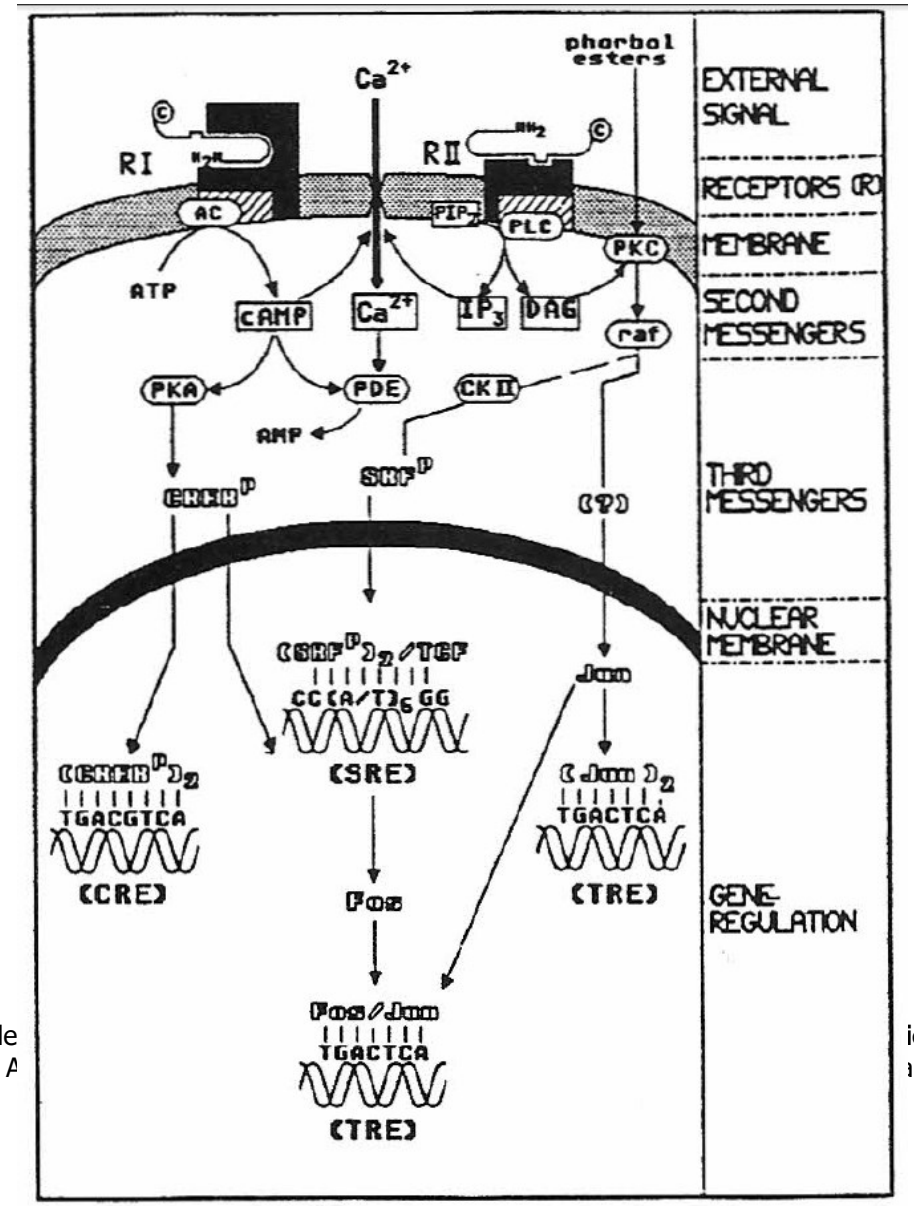
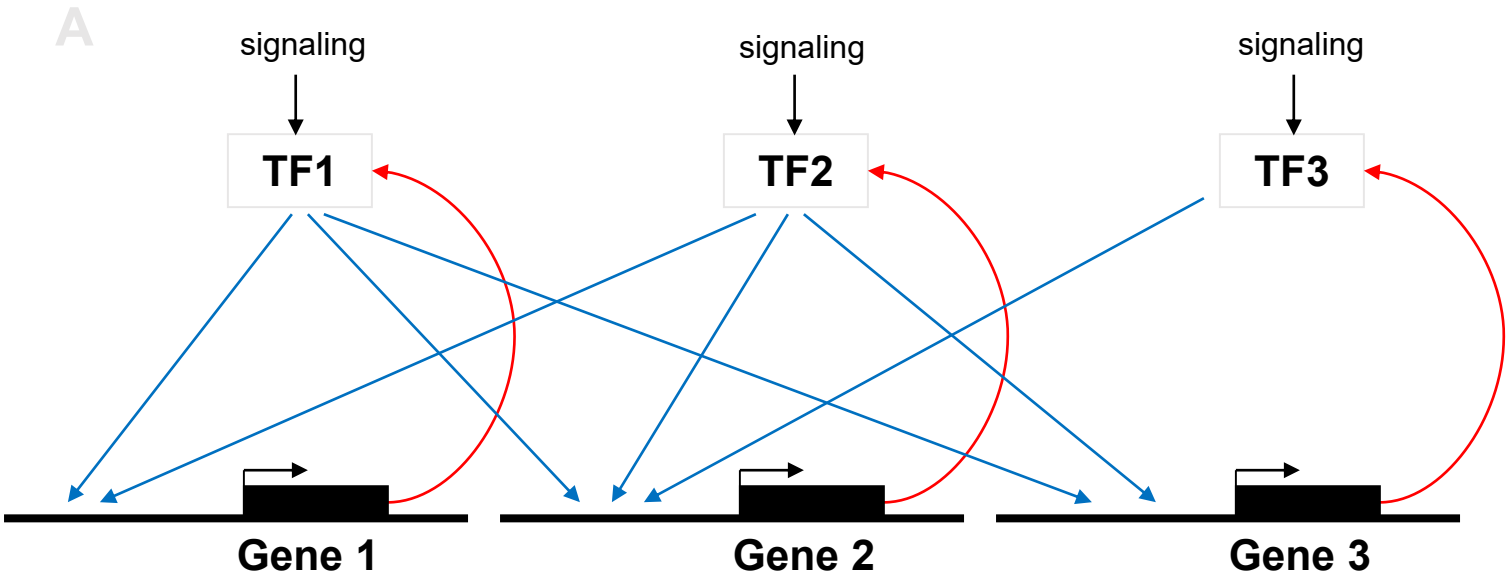
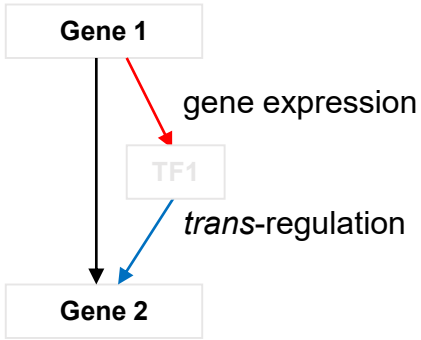


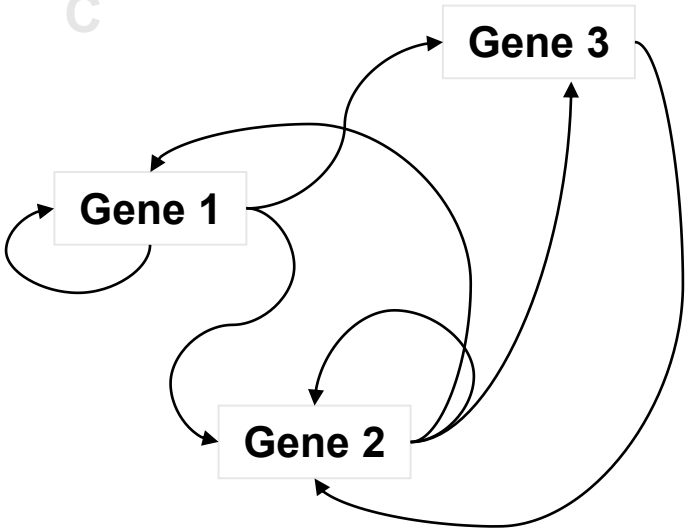
Fig. 3: Some signal transduction pathways and their chromosomal targets.

Wingende … **ictability of their function.**
Genome A… ar Genetics (J. Collins, A.J. Driesel, eds.) 4, 95-108 (1991)

The transcription core network







The choice of a logarithmic base corresponds to the choice of a unit for measuring information. If the base 2 is used the resulting units may be called binary digits, or more briefly *bits,* a word suggested by J. W. Tukey. A device with two stable positions, such as a relay or a flip-flop circuit, can store one bit of information. N such devices can store N bits, since the total number of possible states is 2^N and $\log_2 2^N = N$. If the base 10 is used the units may be called decimal digits. Since

$$\begin{aligned}\log_2 M &= \log_{10} M / \log_{10} 2 \\ &= 3.32 \log_{10} M,\end{aligned}$$

$$H = -K \sum_{i=1}^{n} p_i \log p_i$$

where K is a positive constant.

This theorem, and the assumptions required for its proof, are in no way necessary for the present theory. It is given chiefly to lend a certain plausibility to some of our later definitions. The real justification of these definitions, however, will reside in their implications.

Quantities of the form $H = -\sum p_i \log p_i$ (the constant K merely amounts to a choice of a unit of measure) play a central role in information theory as measures of information, choice and uncertainty. The form of H will be recognized as that of entropy as defined in certain formulations of statistical mechanics[8] where p_i is the probability of a system being in cell i of its phase space. H is then, for example, the H in Boltzmann's famous H theorem. We shall call $H = -\sum p_i \log p_i$ the entropy of the set of probabilities $p_1, \ldots, p_n$. If x is a chance variable we will write $H(x)$ for its entropy; thus x is not an argument of a function but a label for a number, to differentiate it from $H(y)$ say, the entropy of the chance variable y.

C. E. Shannon, *The Bell System Technical Journal,* 27:379–423, 623–656 (1948)

$$\text{mat\_sim} = [\sum_{j=1}^{n} C_i(j) \times \text{score}(b, j)] / [\sum_{j=1}^{n} C_i(j) \times \text{max\_score}(j)] \quad (3)$$

$$0 \leq \text{mat\_sim} \leq 1$$

where $C_i(j)$ is the consensus index value of position j, n is the length of the consensus matrix, score(b,j) is the matrix value for base b at position j and max_score(j) is:

$$\max_{b \in A, C, G, T} \{\text{score}(b, j)\}$$

$$C_i(i) = (100/\ln 5) \times [\sum_{b \in A,C,G,T,gap} P(i, b) \times \ln P(i, b) + \ln 5] \quad (1)$$

$$0 \leq C_i \leq 100$$

$$mSS = \frac{Current - Min}{Max - Min} \quad 1$$

$$Current : \sum_{i=1}^{L} I(i) f_{i,b_i}$$

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