Transcription factor databases

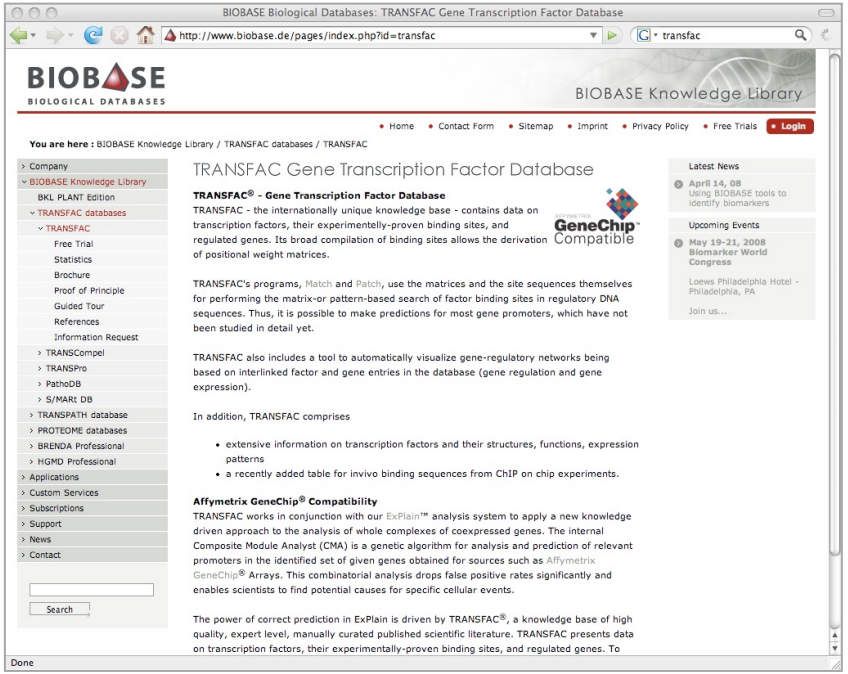
Jacques van Helden
https://orcid.org/0000-0002-8799-8584

Aix-Marseille Université, France
Theory and Approaches of Genome Complexity (TAGC)

Institut Français de Bioinformatique (IFB)
http://www.france-bioinformatique.fr

TRANSFAC - Gene transcription factor database

- Organisms
  - Eukaryotes
  - Particular emphasis on mammals (specially human, mouse, rat)
- Distribution
  - The public version is not updated anymore
  - Commercial version (TRANSFAC PRO)
  - Distributed by BioBaseTM
    - http://www.biobase.de/
- Data content
  - Transcription factors
  - Binding sites
    - **Evidences !**
    - **Publications !**
  - Position-specific scoring matrices
- Pattern matching tools (patch, match)



TRANSFAC – matrix format (V$SOX2_Q6)

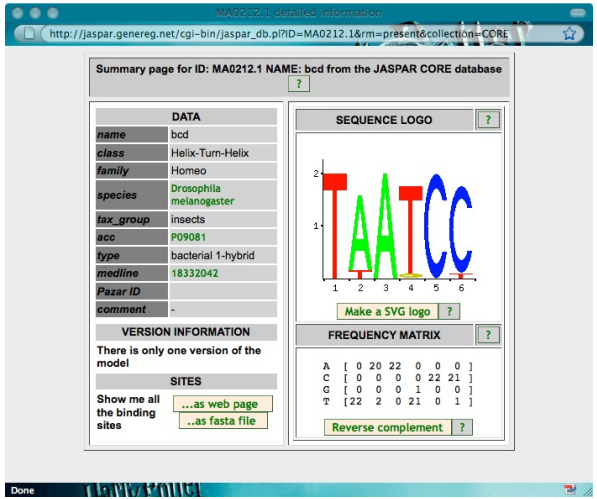
Field descriptions

AC	***Accession no.***
XX	***(field separator)***
ID	***Identifier***
DT	***Date; author***
NA	***Name of the binding factor***
DE	***Short factor description***
BF	***List of linked factor entries***
PO	***A C G T Position within the aligned sequences,***
01	***frequency of A, C, G, T residues, resp.;***
02	***last column: deduced consensus in***
03	***IUPAC 15-letter code***
BA	***Statistical basis***
BS	***Factor binding sites underlying the matrix***
BS	***(SITE accession no.; Start position for matrix sequence; length of sequence used;***
BS	***number of gaps inserted; strand orientation)***
CC	***Comments***
RX	***MEDLINE ID***
RN	***Reference no.***
RA	***Reference authors***
RT	***Reference title***
RL	***Reference data***
//	

```
AC  M01272
XX
ID  V$SOX2_Q6
XX
DT  08.07.2009 (created); dtc.
CO  Copyright (C), Biobase GmbH.
XX
NA  SOX2
XX
BF  T09507; Sox-xbb1; Species: mouse, Mus musculus.
BF  T01836; Sox2; Species: mouse, Mus musculus.
BF  T04915; Sox2; Species: human, Homo sapiens.
BF  T01837; Sox2; Species: chick, Gallus gallus.
BF  T10231; Sox2; Species: Mammalia.
BF  T09970; Sox2; Species: human, Homo sapiens.
BF  T10885; Sox2; Species: monkey, Cercopithecus aethiops.
XX
P0      A      C      G      T
01      6      2      4      4      N
02      7      2      3      4      N
03      4      6      2      4      N
04      4      5      4      3      N
05      2      9      1      4      C
06      0     12      0      4      C
07      8      0      0      8      W
08      0      0      0     16      T
09      0      0      0     16      T
10      0      0     16      0      G
11      0      0      0     16      T
12      2      2      2      9      T
13      7      2      0      6      W
14      0      2      2     11      T
15      1      0      9      5      K
16      4      6      3      2      N
XX
BA  16 compiled sequences
XX
BS  gccctcattgttatgc; R15133; 13; 16;; n.
BS  AAACTCTTTGTTTGGA; R15201; -1; 16;; p.
BS  ttcaccattgttctag; R15231; 11; 16;; n.
BS  GACTCTATTGTCTCTG; R15267; 11; 16;; p.
BS  GATATCTTTGTTTCTT; R16367; -4; 16;; p.
BS  tgcacctttgttatgc; R17099; 5; 16;; n.
BS  aattccattgttatga; R19276; 15; 16;; n.
BS  aaactctttgtttgga; R19367; 20; 16;; n.
BS  atggacattgtaatgc; R19510; 15; 16;; n.
BS  AGGCCTTTTGTCCTGG; R22342; 21; 16;; p.
BS  tgtgCTTTTGTnnnnn; R22344; 1; 16;; p.
BS  ctcaactttgtaattt; R22359; 13; 16;; n.
```

JASPAR

- http://jaspar.genereg.net/
- Public database
- Data content
  - PSSM
  - “sites” (i.e. sequences having served to build the matrix, but no genomic position)
  - Core: transcription factor-specific matrices
  - Collection: matrices for families of transcription factors
- Tools
  - Pattern matching, matrix randomization



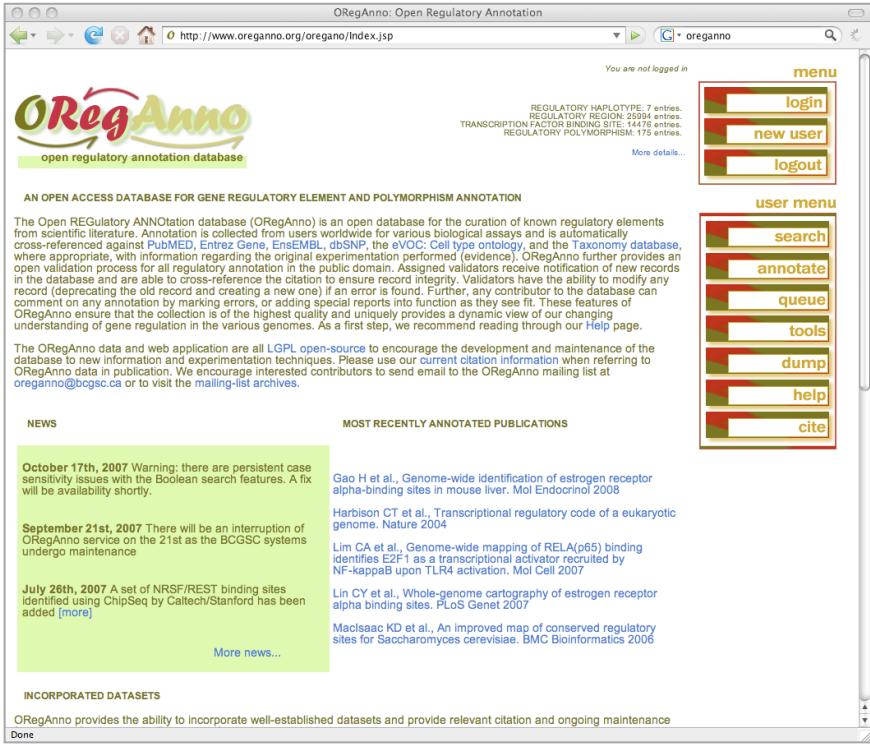
Sequences for model MA0212.1

Site	Occurences
tgt**TAATCC**c	1
tg**GGATTA**ta	1
tta**TAATCC**	1
gct**TAATCC**g	1
ggt**TAATCC**g	1
agc**TTATCC**	1
gaga**TAATCC**	1
gtcc**TAATCC**	1
cgt**TAATCT**c	1
at**GGATTA**ga	2
cgct**TAATCC**	1
cggg**TAATCC**	1
GGCTTAagcc	1
tgt**TAATCC**g	1
tgt**TAATCC**	1
tct**TAATCC**c	1
gg**TTATCC**g	1
gcg**TAATCC**a	1
ggg**TAATCC**	1
tcta**TAATCC**	1
ggtt**TAATCC**	1

ORegAnno

http://www.oreganno.org/oregano/Index.jsp

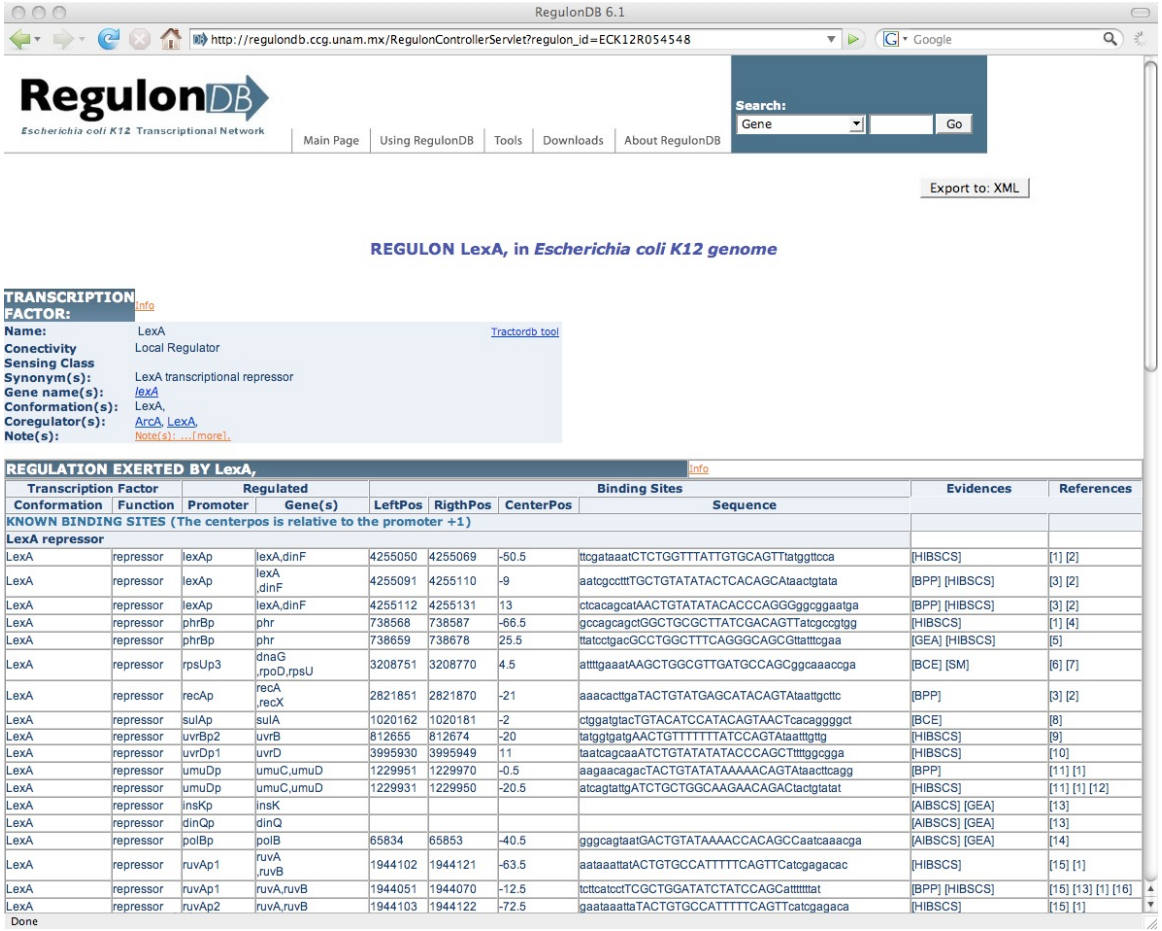
- Also available from the UCSC genome browser
  - http://genome.ucsc.edu/
- Community-based annotation (Jamboree)
- Data content
  - Transcription factor binding sites
  - Mapping on the genomes
  - NO MATRICES
- Scope: all organisms (with specific focus on metazoan)



RegulonDB – Transcriptional regulation in Escherichia coli

- RegulonDB Web site
  - http://regulondb.ccg.unam.mx/
- Model organism: Escherichia coli
- Data content
  - Transcription factors
  - Transcription factor binding sites (TFBS)
  - Position-specific scoring matrices (PSSM)
  - Promoters
  - Operons
- Collaboration with EcoCyc
  - EcoCyc is the reference database about metabolism in Escherichia coli
  - RegulonDB is integrated in the EcoCyc database

Example of regulon in RegulonDB



Transcription Factor		Regulated		Binding Sites				Evidences	References
Conformation	Function	Promoter	Gene(s)	LeftPos	RigthPos	CenterPos	Sequence		
KNOWN BINDING SITES (The centerpos is relative to the promoter +1)									
LexA repressor									
LexA	repressor	lexAp	lexA,dinF	4255050	4255069	-50.5	ttcgataaatCTCTGGTTTATTGTGCAGTTtatggttcca	[HIBSCS]	[1] [2]
LexA	repressor	lexAp	lexA ,dinF	4255091	4255110	-9	aatcgcctttTGCTGTATATACTCACAGCAtaactgtata	[BPP] [HIBSCS]	[3] [2]
LexA	repressor	lexAp	lexA,dinF	4255112	4255131	13	ctcacagcatAACTGTATATACACCCAGGGggcggaatga	[BPP] [HIBSCS]	[3] [2]
LexA	repressor	phrBp	phr	738568	738587	-66.5	gccagcagctGGCTGCGCTTATCGACAGTTatcgccgtgg	[HIBSCS]	[1] [4]
LexA	repressor	phrBp	phr	738659	738678	25.5	ttatcctgacGCCTGGCTTTCAGGGCAGCGttatttcgaa	[GEA] [HIBSCS]	[5]
LexA	repressor	rpsUp3	dnaG ,rpoD,rpsU	3208751	3208770	4.5	attttgaaatAAGCTGGCGTTGATGCCAGCggcaaaccga	[BCE] [SM]	[6] [7]
LexA	repressor	recAp	recA ,recX	2821851	2821870	-21	aaacacttgaTACTGTATGAGCATACAGTAtaattgcttc	[BPP]	[3] [2]
LexA	repressor	sulAp	sulA	1020162	1020181	-2	ctggatgtacTGTACATCCATACAGTAACTcacaggggct	[BCE]	[8]
LexA	repressor	uvrBp2	uvrB	812655	812674	-20	tatggtgatgAACTGTTTTTTTTATCCAGTAtaatttgttg	[HIBSCS]	[9]
LexA	repressor	uvrDp1	uvrD	3995930	3995949	11	taatcagcaaATCTGTATATATACCCAGCTtttggcgga	[HIBSCS]	[10]
LexA	repressor	umuDp	umuC,umuD	1229951	1229970	-0.5	aagaacagacTACTGTATATAAAAACAGTAtaacttcagg	[BPP]	[11] [1]
LexA	repressor	umuDp	umuC,umuD	1229931	1229950	-20.5	atcagtattgATCTGCTGGCAAGAACAGACtactgtatat	[HIBSCS]	[11] [1] [12]
LexA	repressor	insKp	insK					[AIBSCS] [GEA]	[13]
LexA	repressor	dinQp	dinQ					[AIBSCS] [GEA]	[13]
LexA	repressor	polBp	polB	65834	65853	-40.5	gggcagtaatGACTGTATAAAACCACAGCCaatcaaacga	[AIBSCS] [GEA]	[14]
LexA	repressor	ruvAp1	ruvA ,ruvB	1944102	1944121	-63.5	aataaattatACTGTGCCATTTTTCAGTTCatcgagacac	[HIBSCS]	[15] [1]
LexA	repressor	ruvAp1	ruvA,ruvB	1944051	1944070	-12.5	tcttcatcctTCGCTGGATATCTATCCAGCatttttttat	[BPP] [HIBSCS]	[15] [13] [1] [16]
LexA	repressor	ruvAp2	ruvA,ruvB	1944103	1944122	-72.5	gaataaattaTACTGTGCCATTTTTCAGTTcatcgagaca	[HIBSCS]	[15] [1]

PSSM in RegulonDB

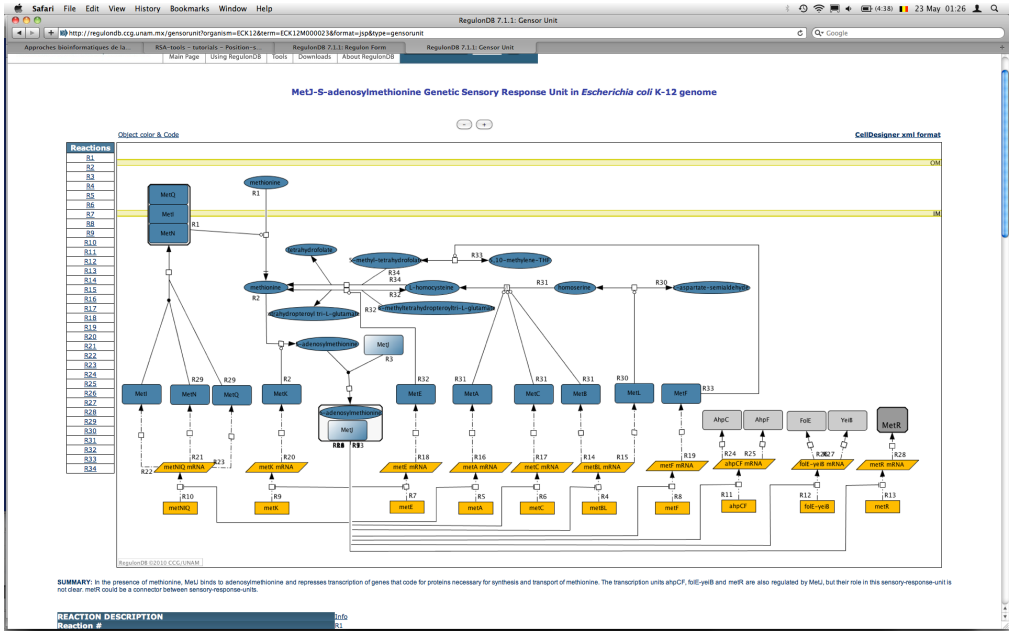
- RegulonDB contains a collection of PSSM built
  aligning annotated binding sites.
- This collection can be used to scan genomes an
  new TFBS.

```
...
------------------------------

Transcription Factor Name      LexA
Total of uniq binding sites    23

Matrix
A   12   0   0   0   1  12   1  12   6  10   7  13   4  12   0  23   0   1  12   6  11
C    3  22   0   0   2   3   5   2   2   5   5   2   4   7  23   0   0   8   2   2   3
G    5   0   0  23   6   3   2   4   0   2   0   3   3   2   0   0  23   1   3   2   1
T    3   1  23   0  14   5  15   5  15   6  11   5  12   2   0   0   0  13   6  13   8

Alignment       Score
ACTGTATAAAACCACAGCCAA          12.05
GCTGCGCTTATCGACAGTTAT          8.48
CCTGGCTTTCAGGGCAGCGTT          7.51
ACTGTTTTTTTATCCAGTATA          16.18
ATTGGCTGTTTATACAGTATT          12.01
CCTGTTAATCCATACAGCAAC          10.7
ACTGTACATCCATACAGTAAC          14.66
TCTGCTGGCAAGAACAGACTA          3.36
ACTGTATATAAAAACAGTATA          17.23
GCTGGATATCTATCCAGCATT          15.55
GCTGGATATCTATCCAGCATT          15.55
ACTGTGCCATTTTTCAGTTCA          8.61
ACTGTGCCATTTTTCAGTTCA          8.61
ACTGTATATAAAACCAGTTTA          16.16
ACTGTACACAATAACAGTAAT          12.47
ACTGTATGAGCATACAGTATA          14.73
GCTGGCGTTGATGCCAGCGGC          4.27
ACTGTTTATTTATACAGTAAA          16.67
TCTGTATATATACCCAGCTTT          14.73
TCTGGTTTATTGTGCAGTTTA          9.97
GCTGTATATACTCACAGCATA          15.05
ACTGTATATACACCCAGGGGG          9.28
CCTGAATGAATATACAGTATT          12.9

------------------------------
....
```

http://regulondb.ccg.unam.mx/data/Matrix_AlignmentSet.txt

“Gensor units” (RegulonDB)

- RegulonDB defines the concept of “***Gensor unit***” as “*a unit that initiates with the signal, continues with the signal transduction to the core of regulation to modify expression of the affected set of target genes, and ends with an adequate response*.”
- Example: MetJ-S-adenosylmethionine Genetic Sensory Response Unit in Escherichia coli K-12 genome



Other databases

- PAZAR http://www.pazar.info/
  - Unification of independent collection of transcription factor binding sites and motifs.
- YeasTract http://www.yeastract.com/
  - Yeast-specific database. Factors, binding sites and motifs + tools.
- FlyReghttp://www.flyreg.org/
  - Drosophila DNase I Footprint Database
- PlantCARE http://bioinformatics.psb.ugent.be/webtools/plantcare/html/
  - Plant Cis-Acting Regulatory Elements

Practical – Transcription factor databases

- Take some matrix from either of those database
  - JASPAR (http://jaspar.genereg.net/)
    - TRANSFAC public version (http://www.gene-regulation.com/)
- Open a connection to RSAT Metazoan server (https://metazoa.rsat.eu).
- Use the tool *convert-matrix* to obtain information on the matrix
  - Display the logo. How do you interpret
    - the information content in each column?
    - the error bars?
  - Redo the conversion, but set the option “Multiply counts” to 10. How does it affect the logo?
  - Redo the conversion, but set the option “Multiply counts” to 0.1. How does it affect the logo?
  - Convert counts to weights
    - How do you justify your choice of the background model?
  - Explore the statistical parameters
- Tips:
  - don’t forget to specify the input format
  - pay a particular attention to the choice of the background model.
- Examples:
  - TRANSFAC e.g. V$OCT1_01
  - JASPAR: compare the two matrices for the factor Klf4 (identifiers MA0039.1 and MA0039.2 resp.)