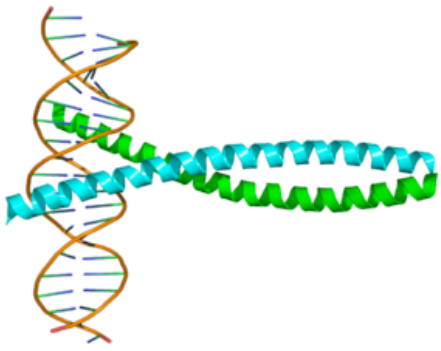


## Transcription Factor Binding Sites

1. Promoters and gene regulation in Eukaryotes
2. Position Weight Matrices (PWM)
3. PWM Databases
4. TFBS prediction using PWMs
5. Pattern Discovery: Finding unknown motifs
6. Exercise: Obtain mouse and human fosB promoters and predict TFBS with Match and JASPAR



## Transcription Factor Binding Sites

1. Promoters and gene regulation in Eukaryotes
2. Position Weight Matrices (PWM)
3. PWM Databases
4. TFBS prediction using PWMs
5. Exercise: Obtain mouse and human fosB promoters  
and predict TFBS with Match and JASPAR

# Non-coding sequences

## Protein binding sites:

Promoters, Enhancers, Silencers, Insulators → TFBS

## Other:

non-coding RNAs, etc

# TFBS: Detection methods

## in vivo

Functional analysis

ChIP

## in vitro on cloned fragment

Footprinting reactions

Exonuclease digests

Gel retardation (EMSA)

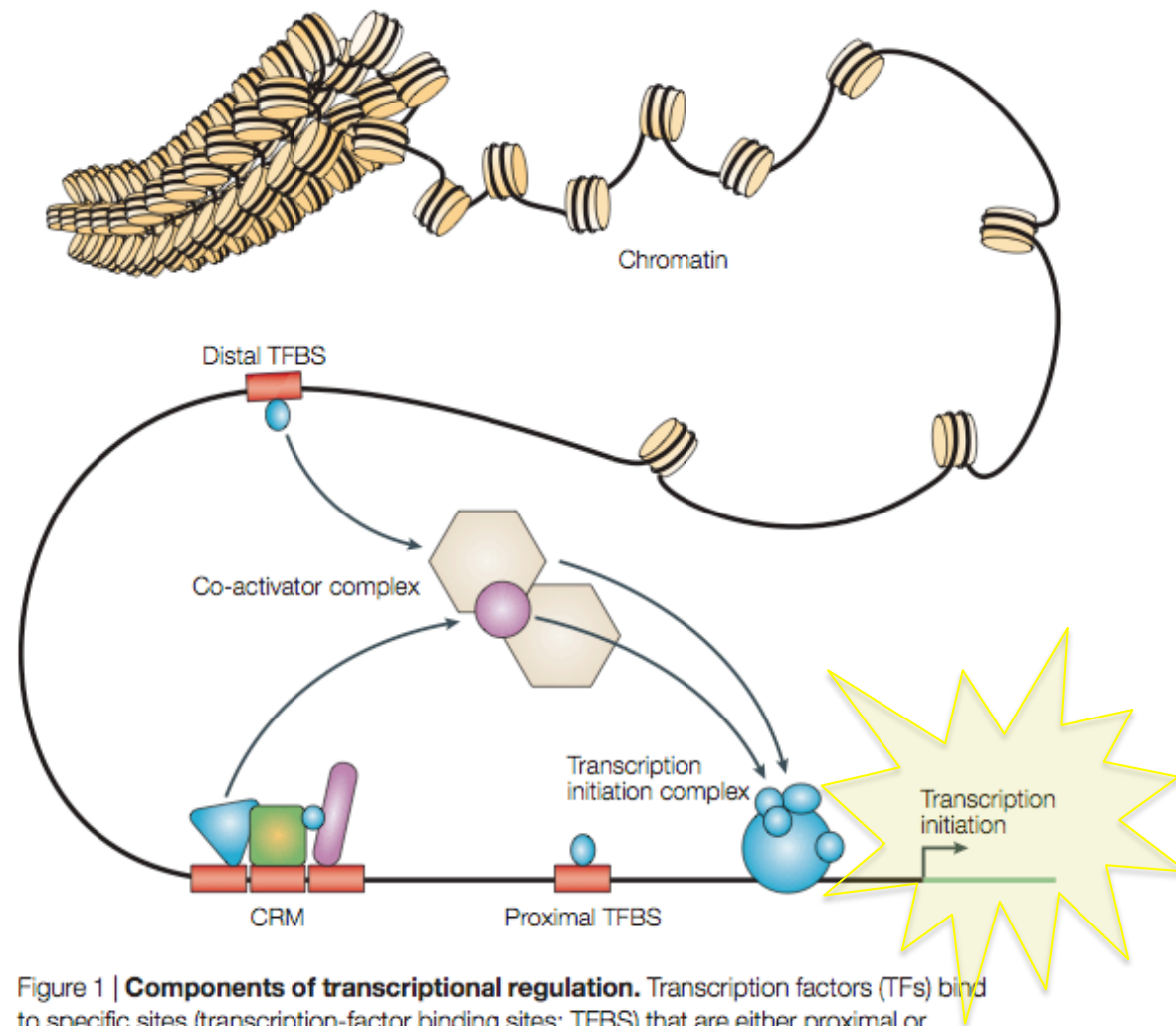
UV Crosslinking

## in vitro on artificial DNA:

SELEX: Systematic Evolution of Ligands by Exponential enrichment

# TF Binding Sites

- Problems:
  - often poorly defined consensus
  - Sequences not conserved within species, and even worse between species
  - Examples of enhancers functionally conserved but not sequence-conserved
  - Most of the TFBS sequence data comes from just a few species
  - Very often in vitro experiments
  - 2 completely different binding sites could be merged in the same matrix/consensus



**Figure 1 | Components of transcriptional regulation.** Transcription factors (TFs) bind to specific sites (transcription-factor binding sites; TFBS) that are either proximal or distal to a transcription start site. Sets of TFs can operate in functional *cis*-regulatory modules (CRMs) to achieve specific regulatory properties. Interactions between bound TFs and cofactors stabilize the transcription-initiation machinery to enable gene expression. The regulation that is conferred by sequence-specific binding TFs is highly dependent on the three-dimensional structure of chromatin.



# EPD

## The Eukaryotic Promoter Database

### Current Release 99



**The Eukaryotic Promoter Database** is an annotated non-redundant collection of eukaryotic POL II promoters, for which the transcription start site has been determined experimentally. Access to promoter sequences is provided by pointers to positions in nucleotide sequence entries. The annotation part of an entry includes description of the initiation site mapping data, cross-references to other databases, and bibliographic references. EPD is structured in a way that facilitates dynamic extraction of biologically meaningful promoter subsets for comparative sequence analysis. [[More details](#)].

**Current version is based on EMBL Release 99.**

#### Access to EPD

searching **EPD** using complete or partial AC, ID or documentation text  
(accepts one single query string!)

- [Browse the Eukaryotic Promoter Database](#)
- [Download promoter sequences](#)
- [BLAST search \[-10 to 6 kb relative to TSS in EPD\]](#)
- [Promoter Elements](#)
- [local EPD FTP](#)
- [SRS access to EPD](#)

#### Documents

- [EPD user manual](#)
- [List of genome assemblies used in EPD](#)
- [List of alternative promoters](#)
- [Keyword list](#)
- [Groups of homologous promoters](#)
- [Contact EPD developers](#)

<http://www.epd.isb-sib.ch/>



COLD SPRING  
HARBOR  
LABORATORY

Home

People

Databases and Software  
Tools

Collaborations

Courses

Seminars

Meetings and  
Workshops

Publications

Links

Positions Available

# Zhang Lab: Computational Biology and Bioinformatics

## Databases

- **AEDB**: Alternative Exon Database
- **AtProbe**: Arabidopsis thaliana promoter binding element database (public)
- **CEPDB**: C. elegans Promoter Database
- **CSEdb**: Conserved sequence elements database (public)
- **CSHLmpd**: Cold Spring Harbor Laboratory Mammalian Promoter database (public)
- **TRED**: Transcriptional Regulatory Element Database (public)
- **Drosophila Promoter and Gene Expression Database** (Coming soon).  
Try one of its components: **DBSD: Drosophila Binding Site Database**.
- **LSPD**: The Liver Specific Gene Promoter Database
- **SCPD**: Yeast Promoter Database (public)
- **VertPD**: Vertebrate Promoter Database (Internal to CSHL staff only)

## Software Tools

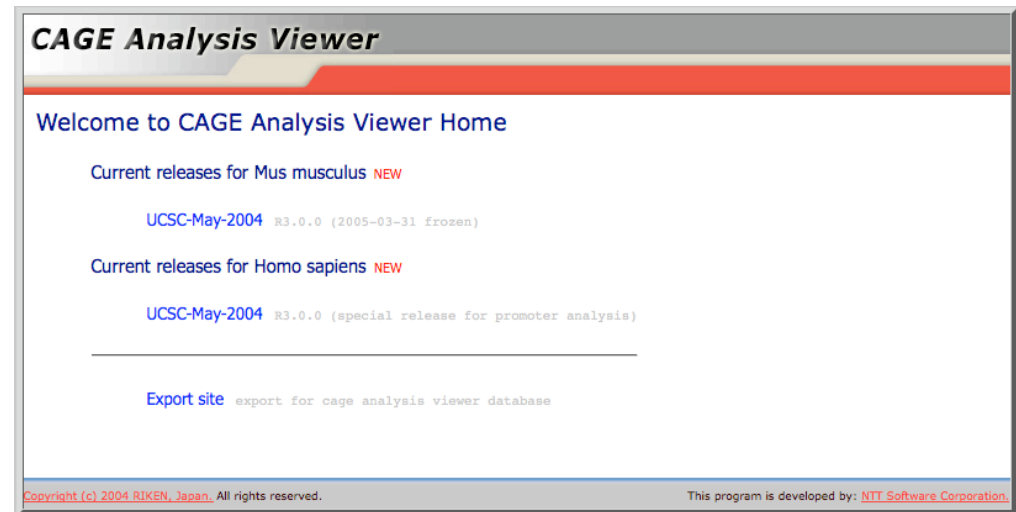
<http://rulai.cshl.edu/software/index1.htm>



# Experimental Transcription Start Sites (TSS)



<http://gerg01.gsc.riken.jp/cage/>

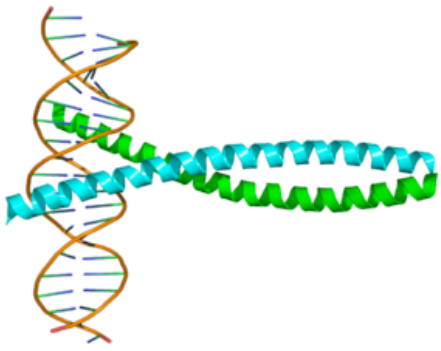


[http://gerg01.gsc.riken.jp/cage\\_analysis/](http://gerg01.gsc.riken.jp/cage_analysis/)

*D632–D636 Nucleic Acids Research, 2006, Vol. 34, Database issue  
doi:10.1093/nar/gkj034*

## **CAGE Basic/Analysis Databases: the CAGE resource for comprehensive promoter analysis**

Hideya Kawaji<sup>1</sup>, Takeya Kasukawa<sup>1,2</sup>, Shiro Fukuda<sup>2</sup>, Shintaro Katayama<sup>2,\*</sup>, Chikatoshi Kai<sup>2</sup>, Jun Kawai<sup>2,3</sup>, Piero Carninci<sup>2,3</sup> and Yoshihide Hayashizaki<sup>2,3</sup>



## Transcription Factor Binding Sites

1. Promoters and gene regulation in Eukaryotes
- 2. Position Weight Matrices (PWM)**
3. PWM Databases
4. TFBS prediction using PWMs
5. Pattern Discovery: Finding unknown motifs
6. Exercise: Obtain mouse and human fosB promoters and predict TFBS with Match and JASPAR

## Data collection

Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C
	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Source binding sites

B R M C W A W H R W G G B M

Consensus sequence

Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0

Probabilities can be calculated and corrected for background

Position weight matrix (PWM)

A	-1.93	0.79	0.79	-1.93	0.45	1.50	0.79	0.45	1.07	0.79	0.00	-1.93	-1.93	0.79
C	0.45	-1.93	0.79	1.68	-1.93	-1.93	-1.93	0.45	-1.93	-1.93	-1.93	-1.93	0.00	0.79
G	0.00	0.45	-1.93	-1.93	-1.93	-1.93	-1.93	-1.93	0.66	-1.93	1.30	1.68	1.07	-1.93
T	0.15	0.66	-1.93	-1.93	1.07	0.66	0.79	0.00	0.00	0.79	-1.93	-1.93	-0.66	-1.93

Also called position-specific scoring matrices (PSSMs). In log scale.

# From PFM to PWM/PSSM

Corrected probabilities of observing a given nucleotide can be calculated using equation 1.

Corrected probability calculation:

$$p(b,i) = \frac{f_{b,i} + s(b)}{N + \sum_{b' \in \{A,C,G,T\}} s(b')} \quad (1)$$

$f_{b,i}$  = counts of base  $b$  in position  $i$ ;  $N$  = number of sites;  $p(b,i)$  = corrected probability of base  $b$  in position  $i$ ;  
 $s(b)$  = pseudocount function

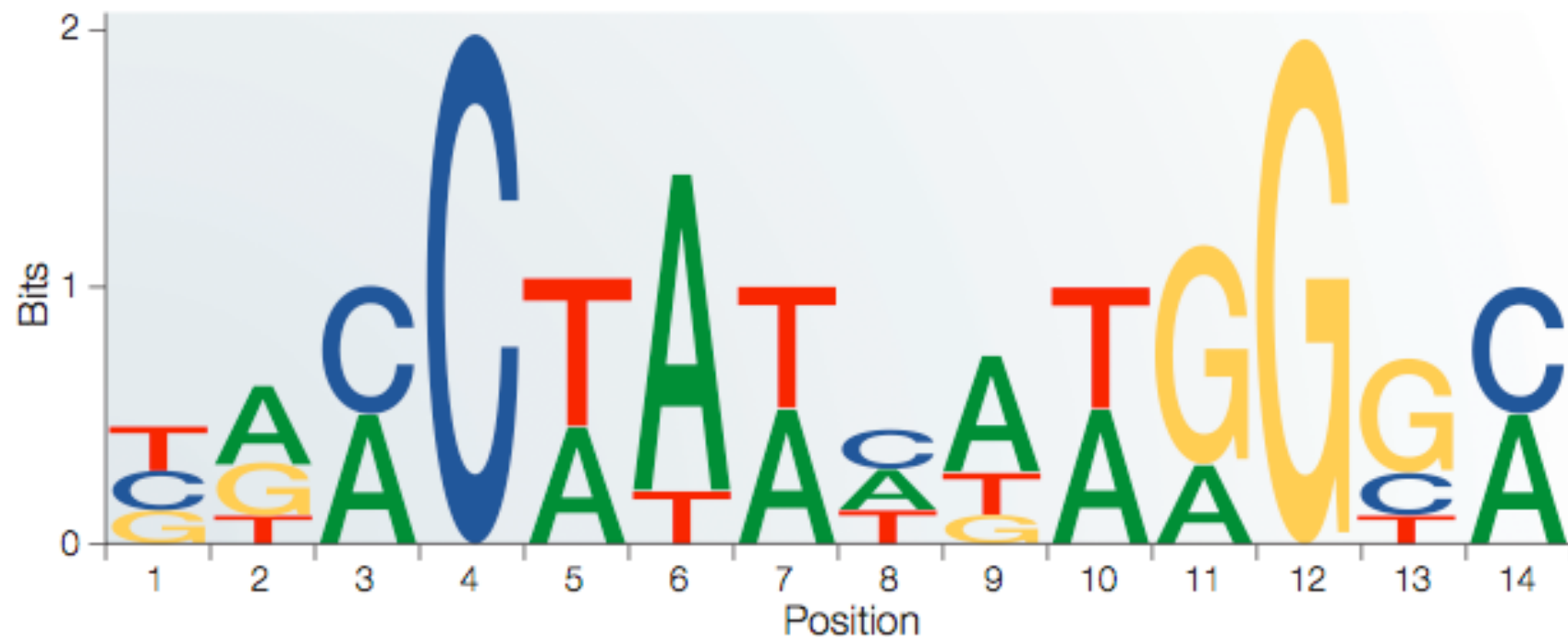
A position weight matrix (PWM) is constructed by dividing the nucleotide probabilities in (1) by expected background probabilities and converting the values to a log-scale (see equation 2).

PWM conversion:

$$W_{b,i} = \log_2 \frac{p(b,i)}{p(b)} \quad (2)$$

Position frequency matrix (PFM)

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0



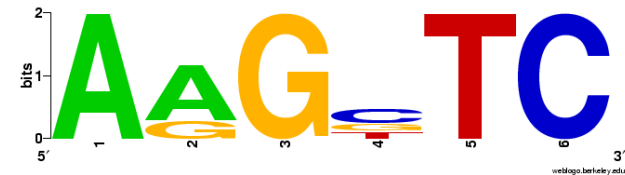
SEQUENCE LOGOS: The information content of a matrix column ranges from 0 (no base preference) and 2 (only 1 base used).

# Summary

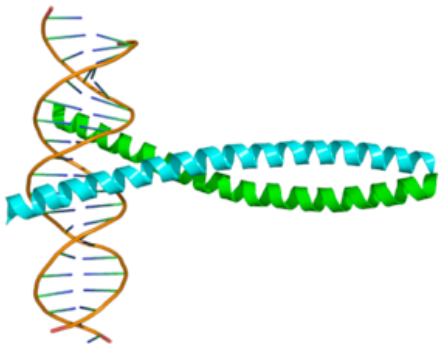
AAGTTC  
AAGCTC  
AGGCTC  
AAGGTC



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



Consensus: ARGBTC



## Transcription Factor Binding Sites

1. Promoters and gene regulation in Eukaryotes
2. Position Weight Matrices (PWM)
- 3. PWM Databases**
4. TFBS prediction using PWMs
5. Pattern Discovery: Finding unknown motifs
6. Exercise: Obtain mouse and human fosB promoters and predict TFBS with Match and JASPAR

**Transfac:** not free, 848 matrices, loads of information and references, quality score based on methods used

**Jaspar:** open sources, 123 matrices, minimal information, majority based on SELEX method (80%)



# TRANSFAC®

**gene-regulation.com**

Sponsored by BIOBASE

> Home  
> **Databases**  
> Free Trials  
> Programs  
> Publications  
> Commercial Offers  
> Events (BB Website)  
> News (BB Website)  
> Links  
> About Us

## Make the switch and get the rest of the year free!

For the rest of 2008, if you sign up for any of our professional products, we'll give you the rest of the year for free.

[Sign up now and save!](#)

## Public Databases for Academic and Non-profit Organizations

**TRANSFAC® 7.0 Public 2005** contains data on transcription factors, their experimentally-proven binding sites, and regulated genes. Its broad compilation of binding sites allows the derivation of positional weight matrices.

- [TRANSFAC Professional](#)
- [Subscription benefits](#)
- [Search TRANSFAC Public](#)
- [Classification](#)
- [Documentation](#)
- [Fungal Metabolic](#)
- [Pax factors in TRANSFAC®](#)
- [The green site of TRANSFAC®](#)
- [Quality Management in TRANSFAC®](#)
- [TfBlast: Search Tool for Sequence Search in the TRANSFAC® Factor Table](#)

## Database Login

Not logged in.

Name

Password

Log In

Password [forgotten?](#)

[New User?](#) [Help?](#)

## Feedback

[Contact us](#)

<http://www.gene-regulation.com/pub/databases.html>

# Transfac example: WT1

TRANSFAC MATRIX TABLE, Release 12.1 - licensed - 2008-03-31, (C) Biobase GmbH					
Statistics	Number of binding factors 9 Number of references 1				
<a href="#">Accession Number</a>	M01118				
<a href="#">Accession numbers, secondary</a>	M00709				
<a href="#">Identifier</a>	V\$WT1_Q6				
<a href="#">Created</a>	02.06.2006 by <a href="#">dtc</a> .				
<a href="#">Updated</a>					
Copyright	Copyright (C), Biobase GmbH.				
<a href="#">Name</a>	WT1				
<a href="#">Binding factors</a>	<a href="#">T00899</a> ; WT1; Species: human, Homo sapiens. <a href="#">T02351</a> ; WT1; Species: mouse, Mus musculus. <a href="#">T02352</a> ; WT1; Species: rat, Rattus norvegicus. <a href="#">T01839</a> ; WT1 -KTS; Species: human, Homo sapiens. <a href="#">T01840</a> ; WT1 I; Species: human, Homo sapiens. <a href="#">T00900</a> ; WT1 I -KTS; Species: human, Homo sapiens. <a href="#">T01842</a> ; WT1 I-del2; Species: human, Homo sapiens. <a href="#">T01841</a> ; WT1-del2; Species: human, Homo sapiens. <a href="#">T09249</a> ; WT1-isoform1; Species: mouse, Mus musculus.				
<a href="#">Binding Matrix</a>	A	C	G	T	Consensus
	1	12	8	0	S
	6	11	4	1	M
	0	23	0	0	C
	4	8	2	9	N
	0	23	0	0	C
	4	19	0	0	C
	1	9	6	6	N
	3	8	8	2	S
	1	15	4	1	C



# Transfac example: WT1

<a href="#">Basis</a>	23 compiled sequences					
<a href="#">Binding sites</a>	Sequence	Derived from	Start	Length	Gaps	Orientation
	gcctcacnn	<a href="#">R02307</a>	-2	9		n.
	nnCTCCCTC	<a href="#">R02308</a>	-2	9		p.
	cacacannn	<a href="#">R02309</a>	-3	9		n.
	cacacatac	<a href="#">R02310</a>	2	9		n.
	cacaccctc	<a href="#">R02311</a>	3	9		n.
	CACTCCAGG	<a href="#">R02312</a>	7	9		p.
	CGCCCCCGC	<a href="#">R02313</a>	1	9		p.
	gcccccgca	<a href="#">R02314</a>	-1	9		n.
	CGCCCCCGC	<a href="#">R02315</a>	1	9		p.
	cccaccgc	<a href="#">R04858</a>	-4	9		n.
	agcccacgc	<a href="#">R04859</a>	1	9		n.
	gcccccgcg	<a href="#">R04860</a>	-1	9		n.
	gcccccgcg	<a href="#">R04861</a>	-1	9		n.
	CACGCCCGC	<a href="#">R04862</a>	-2	9		p.
	ccctcctcc	<a href="#">R04863</a>	-4	9		n.
	ggctccggc	<a href="#">R04864</a>	-5	9		n.

# Transfac example: WT1

<a href="#">Accession Number</a>	R04859
<a href="#">Identifier</a>	RAT\$IGF1R_02
<a href="#">Created</a>	05.03.1998 by <a href="#">ili</a>
<a href="#">Updated</a>	08.11.2000 by <a href="#">vma</a>
<a href="#">Copyright</a>	Copyright (C), Biobase GmbH
<a href="#">Sequence type</a>	DNA
<a href="#">Description</a>	IGF-1 receptor (insulin-like growth factor I receptor); Gene: <a href="#">G001118</a>
<a href="#">Species</a>	rat, Rattus norvegicus
<a href="#">Taxonomic classification</a>	eukaryota; animalia; metazoa; chordata; vertebrata; tetrapoda; mammalia; eutheria; rodentia; myomorpha; muridae; murinae
<a href="#">Sequence</a>	GCGTGGGCT
<a href="#">First position of element</a>	-250
<a href="#">Last terminating position</a>	-242
<a href="#">Binding factors</a>	<a href="#">T00899</a> ; WT1; Quality: 2; Species: human, Homo sapiens. <a href="#">T01839</a> ; WT1 -KTS; Quality: 2; Species: human, Homo sapiens. <a href="#">T01840</a> ; WT1 I; Quality: 2; Species: human, Homo sapiens. <a href="#">T00900</a> ; WT1 I -KTS; Quality: 2; Species: human, Homo sapiens.
<a href="#">Matrices</a>	<a href="#">M01118</a> ; V\$WT1_Q6
<a href="#">Cellular factor source</a>	<a href="#">0306</a> ; rec(human-E.coli).
<a href="#">Method(s)</a>	DNase I footprinting
<a href="#">Comments</a>	conflict: EMBL #M37807 (168:176) is gcgtgggcG; some weak protection by the WT1 zinc finger region containing the KTS insertion was shown, but it was argued against it in <a href="#">[1]</a>
<a href="#">External database links</a>	EMBL: <a href="#">M37807</a> ; RNIGFIRC (168:176). TRANSPRO: <a href="#">RNO_10607</a> .
<a href="#">Reference number</a>	[1]; <a href="#">RE0006666</a> . PUBLISHED: <a href="#">8175666</a> .
<a href="#">Author(s), Title, Journal</a>	Werner H., Rauscher III F. J., Sukhatme V. P., Drummond I. A., Roberts jr C. T., LeRoith D. Transcriptional repression of the insulin-like growth factor I receptor (IGF-I-R) gene by the tumor suppressor WT1 involves binding to sequences both upstream and downstream transcription start site J. Biol. Chem. 269:12577-12582 (1994).

# Transfac example: WT1

<a href="#">Accession Number</a>	R16085
<a href="#">Identifier</a>	RAT\$EGFR_01
<a href="#">Created</a>	02.12.2004 by abd2.
<a href="#">Updated</a>	17.03.2005 by oke.
Copyright	Copyright (C), Biobase GmbH.
<a href="#">Sequence type</a>	DNA
<a href="#">Description</a>	EGFR (epidermal growth factor receptor); Gene: <a href="#">G009653</a> .
<a href="#">Species</a>	rat, Rattus norvegicus
<a href="#">Taxonomic classification</a>	eukaryota; animalia; metazoa; chordata; vertebrata; tetrapoda; mammalia; eutheria; rodentia; myomorpha; muridae; murinae
<a href="#">Gene region</a>	promoter
<a href="#">Sequence</a>	cTCCTCCTCCacttagtcccTCCTCCTCCcgcccaacctccccacgtcccgaccaggg
<a href="#">Element</a>	NGF-responsive
<a href="#">First position of element</a>	-318
<a href="#">Site terminating position</a>	-260
<a href="#">Binding factors</a>	<a href="#">T02352</a> ; WT1; Quality: 3; Species: rat, Rattus norvegicus.
<a href="#">Matrices</a>	<a href="#">M01118</a> ; V\$WT1_Q6
<a href="#">Cellular factor source</a>	<a href="#">0925</a> ; PC12+NGF. <a href="#">1729</a> ; PC-12.
<a href="#">Method(s)</a>	direct gel shift functional analysis gel shift competition supershift (antibody binding)
<a href="#">Comments</a>	TCC TCCTCC repeat sequences are required for down-regulation of rat EGFR by NGF in PC12 cells and is mediated through WT1 <a href="#">[1]</a> <a href="#">[2]</a>
<a href="#">External database links</a>	EMBL: <a href="#">AF142153</a> ; AF142153 (2407:2465). TRANSPRO: <a href="#">RNO_3204</a> .
<a href="#">Reference number</a>	[1]; <a href="#">RE0025181</a> .    PUBMED: <a href="#">11071895</a> .
<a href="#">Author(s), Title, Journal</a>	Liu X. W., Gong L. J., Guo L. Y., Katagiri Y., Jiang H., Wang Z. Y., Johnson A. C., Guroff G. The Wilms' tumor gene product WT1 mediates the down-regulation of the rat epidermal growth factor receptor by nerve growth factor in PC12 cells

# Transfac quality assignment

Score	Meaning
1	Functionally confirmed transcription factor binding site
2	Binding of pure protein (purified or recombinant)
3	Immunologically characterized binding activity of a cellular extract
4	Binding activity characterized via a known binding sequence
5	Binding of uncharacterized extract protein to a bona fide element
6	No quality assigned



# Jaspar example: Pax6

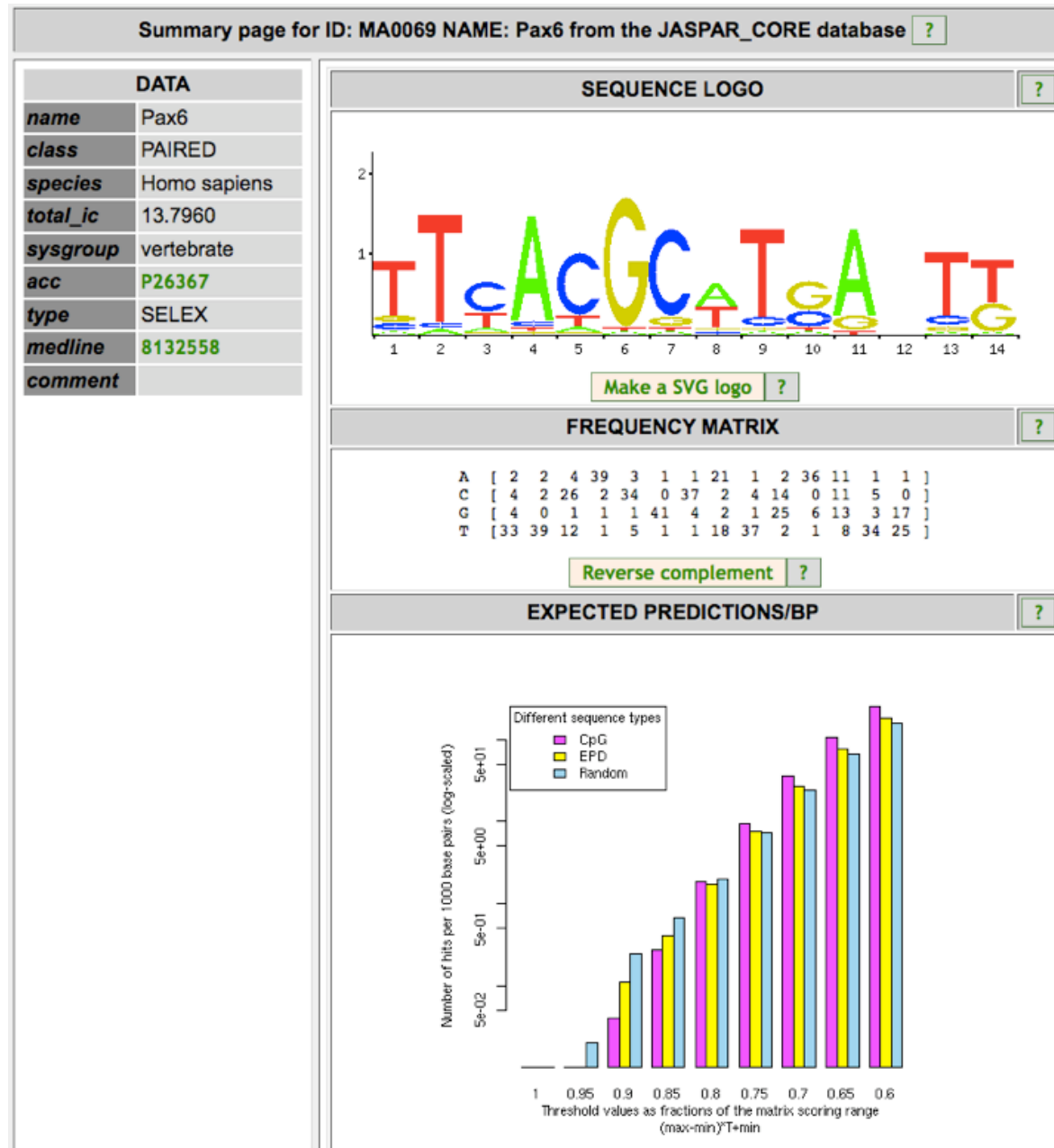
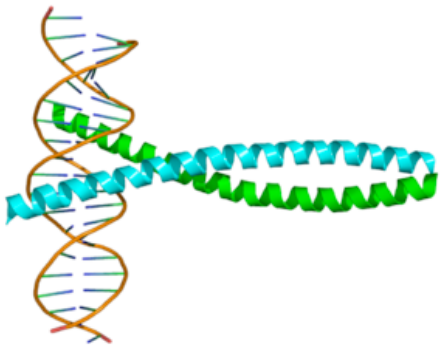


Table of number of hits per 1000 base pairs for each sequence type

Threshold	CpG	EPD	Random
1	0	0	0
0.95	0	0	0.01
0.9	0.03	0.1	0.23
0.85	0.26	0.39	0.67
0.8	1.85	1.71	1.94
0.75	9.36	7.44	7.19
0.7	35.2	26.42	23.61
0.65	105.31	76.35	67.07
0.6	253.49	183.08	159.38





## Transcription Factor Binding Sites

1. Promoters and gene regulation in Eukaryotes
2. Position Weight Matrices (PWM)
3. PWM Databases
- 4. Pattern Matching: TFBS prediction using PWMs**
5. Pattern Discovery: Finding unknown motifs
6. Exercise: Obtain mouse and human fosB promoters and predict TFBS with Match and JASPAR

## Programs

- [AliBaba2](#)
- [BOXSHADE](#)
- [ClustalW](#)
- [Dialign2](#)
- [F-Match](#)
- [Match](#)
- [MatrixCatch](#)
- [m2transfac](#)
- [Composite Module Analyst \(CMA\)](#)
- [PolyAScan](#)
- [ReadSeq](#)
- [SignalScan](#)
- [molwSearch](#)
- [P-Match](#)
- [Patch](#)
- [SbBlast](#)
- [SnpFind](#)
- [TfBlast](#)

Database Login

User: [dricorod](#) [Logout](#)

Name

dricorod

Password

\*\*\*\*\*

Log In

Password [forgotten?](#)

[New User?](#) [Help?](#)

Feedback

<http://www.gene-regulation.com/cgi-bin/pub/programs/match/bin/match.cgi>

## Match - 1.0 Public

Match<sup>TM</sup> is designed for searching potential binding sites for transcription factors (TF binding sites) nucleotide sequences. Match<sup>TM</sup> uses a library of mononucleotide weight matrices from TRANSFAC<sup>®</sup> 6.0

**Authors:** Alexander Kel - BIOBASE GmbH, Ellen Goessling - BIOBASE GmbH

**License:** free for non-commercial use only

[Use Match on this site](#)

your login name: dricorod

Select a previous search result:

default.out

and

VIEW

it

DELETE

Select a previously stored sequence:

default.seq

and

DELETE

it

[A Match<sup>TM</sup> version with additional functionalities is included in the ExPlain<sup>TM</sup> Analysis Platform](#)
[Get help](#)
[Goto Match Profiler](#)

MATCH<sup>TM</sup> public version 1.0

Matrix Search for Transcription Factor Binding Sites

Please enter a name for your search:

default

Sequence Selection:

Select one of your stored sequences:

default.seq


OR take an example

OR take a new sequence and enter a name for it:

default

Please [enter your sequence](#) or several sequences (you can use cut & paste):

Allowed formats are: RAW, FASTA, TRANSFAC, EMBL, GenBank, IG



Matrix or Profile Selection:

Matrices:

Group of matrices:

all, vertebrates, fungi

☒ use high quality matrices only

Cut-off selection for matrix group:

☐ to minimize false positives

☒ to minimize false negatives

☐ to minimize the sum of both error rates

0.7 and 0.75 as mat. sim. and core sim. cut-off

Predefined Profiles:

our profiles

your profiles

muscle\_specific.prf

Submit the form

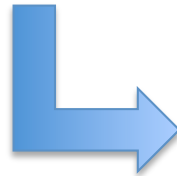
Reset the form

Mail to [info@biobase.de](mailto:info@biobase.de)



## The high-quality transcription factor binding profile database

Browse the JASPAR\_CORE database right away!



Click here to select  
all TFBS

The JASPAR database

http://jaspar.genereg.net/cgi-bin/jaspar\_db.pl

Most ▾ cnio INT FTP GP 4.0 Fati Ast ID BiC R hhp Map q Mail wavi hancock

SEARCH  AND  AND  SEARCH ?

JASPAR matrix models:					
	ID	name	species	class	Sequence logo
<input type="checkbox"/>	MA0001	AGL3	Arabidopsis thaliana	MADS	
<input type="checkbox"/>	MA0002	RUNX1	Arabidopsis thaliana	RUNT	
<input type="checkbox"/>	MA0003	TFAP2A	Homo sapiens	AP2	
<input type="checkbox"/>	MA0004	Amt	Mus musculus	bHLH	
<input type="checkbox"/>	MA0005	Agamous	Arabidopsis thaliana	MADS	
<input type="checkbox"/>	MA0006	Arnt-Ahr	Mus musculus	bHLH	
<input type="checkbox"/>	MA0007	Ar	Rattus rattus	NUCLEAR RECEPTOR	

ANALYZE selected matrix models:

CLUSTER ? selected models using STAMP

Create RANDOM matrix models based on selected models

Number of matrices: 200 Format: Raw RANDOMIZE ?

Create models with PERMUTED columns from selected:

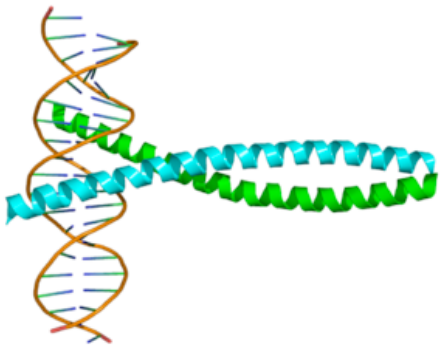
Type: Within each matrix Format: Raw PERMUTE ?

SCAN this (fasta-formatted) sequence with selected matrix models

```
>Fosb:chr7:13752431 [-1500..299](-) [mouse, Mus musculus]
tgcattgtggagatcagagaaacatttttaggagtcatttttctcttcca
ccgttgagtcacagggatgggactcaggttgacaggtatgtgtagaagat
ccttcaactgtgaacacatttgcgtgggctgatacactgcgtgatgetaaa
cacatgocacagattatattatgcactagttttttttttttttttttt
cgagacaggtttctctgtgtagccctggctgtccctggaactcactctgt
agaacaagctggtttcgactcagagatctccctgctctgtctgggat
taagatgtgtgctgtgcgcagctgcacacacacacacacacacacaa
ttattattgtcaattattattgtatacactttctttctgtctctctg
agctaagtgggaaggagaggttccacacacacacacacacacacacac
ccgggtgtgggttccactcctcctcctcctcctcctcctcctcctcctc
ggacccggaaaagccacactccttctgttagctatgagtaactgaagtt
tagatcaggaggagacacattttccaaagctcaccaggtcacacagatc
cgtgacaaagctagtggggaggagactgcactcctacacacacacacac
tggagatatcacagactctccaatctcctcctcctcctcctcctcctcct
tagactgattataccctctcgcagaaatctgcactggggacacctgctct
tttccacagcaggggcgtgcacccgtttggggagggtggggtccacagg
gggtataagcagacactgggactctggagtgcacacctccacacccgggtca
gcaggggcctctcaggagggttttaggcgcctgtcaattcagcctccggg
acagcgtggaactgcgcaggcgcgggcgggttcgcacacgcgcacacagc
ggcgcgcgcaggagcagggttccctcctcagcgaattgtcaggtatcca
```

Relative profile score threshold 80 %

SCAN ?



## Transcription Factor Binding Sites

1. Promoters and gene regulation in Eukaryotes
2. Position Weight Matrices (PWM)
3. PWM Databases
4. Pattern Matching: TFBS prediction using PWMs
- 5. Pattern Discovery: Finding unknown motifs**
6. Exercise: Obtain mouse and human fosB promoters and predict TFBS with Match and JASPAR

# Pattern discovery

## Reference Genome

## Sequences of interest

Seq. oligo	expected frequency	Seq. oligo	observed frequency	
AAAAAA	0.00024	AAAAAA	0.00023	
AAAAAC	0.00030	AAAAAC	0.00031	
AAAAAG	0.00031	AAAAAG	0.00125	***
AAAAAT	0.00024	AAAAAT	0.00018	
AAAACC	0.00028	AAAACC	0.00026	
...		...		

#### MEME Suite Menu

- + Submit A Job
- + Documentation
- + Downloads
- + User Support
- + Alternate Servers
- + Authors
- + Citing

# The MEME Suite

## Motif-based sequence analysis tools

Previous version: [Meme 4.0.0](#)

The MEME Suite allows you to:

- discover motifs using [MEME](#) or [GLAM2](#) on groups of related DNA or protein sequences,
- [search](#) sequence databases using motifs,
- [compare](#) a motif to all motifs in a database of motifs, and
- [associate](#) motifs with Gene Ontology terms via their putative target genes.

To submit a query, click on one of the logos below or select "Submit A Job" from the menu at the left.



Maintenance and development of the MEME Suite is funded by the National Center for Research Resources grant NIH/NCRR R01 RR021692. The MEME Suite web server is funded by the [National Biomedical Computation Resource](#).

Developed and maintained by:




Version 4.1.0

Please send comments and questions to: [meme@nbcrc.net](mailto:meme@nbcrc.net)

Powered by [Opal](#)

<http://meme.sdsc.edu/meme/>

# Looking at conservation of over-represented motif (pattern) discovery



[cisRED Home](#) | [Databases & Methods](#) | [Documents](#) | [Software](#) | [Publications](#) | [Acknowledgements](#)

[Inicio](#) > [Home](#)

## Databases of genome-wide regulatory module and element predictions

Database	Assembly	Search regions	Search region type	Nbr. of input species	Conserved motifs	Discovery p-value threshold	Ensembl compatibility	Release date
<a href="#">Human 9</a>	NCBI v36b	18.7k	promoter	41	236k	0.01	Build 38-49	26 Jul. 2007
<a href="#">Mouse 4</a>	NCBI m37	17.5k	promoter	38	223k	0.1	Build 47-49	26 Sep. 2007
<a href="#">Mouse 3.1</a>	NCBI m35	17.5k	promoter	38	223k	0.1	Build 38	18 Apr. 2007
<a href="#">Rat 1.1</a>	RGSC v3.1	6.7k	promoter	28	116k	0.25	n/a	12 Feb. 2006
<a href="#">C.elegans 4</a>	WormBase WS170	3.8k	promoter	8	158k	1.0	Build 44-46	18 Jul. 2008
<a href="#">Human Stat1 ChIP-seq peaks 1</a>	NCBI v35	226	ChIP-seq	23	~6k	1.0	n/a	03 Apr. 2007

### Overview

The cisRED database holds conserved sequence motifs identified by genome scale motif discovery, similarity, clustering, co-occurrence and coexpression calculations. Sequence inputs include [low-coverage genome sequence](#) data and [ENCODE](#) data. A Nucleic Acids Research [article](#) describes the system architecture; please use this publication to cite cisRED. PubMed publications that cite cisRED are listed [here](#).

cisRED makes three levels of information available for regulatory elements:

- 'Atomic' motifs:** These are conserved, over-represented, sequence sets, typically 6 to 12 bp long, that have been discovered in a 'search region' sequence set.
- Groups of 'similar' motifs:** These are identified either by a) annotating motifs with site sequences from TRANSFAC, JASPAR and ORegAnno databases (annotation-based groups), or by b) 'de novo' hierarchical clustering with the OPTICS algorithm ('de novo' groups).
- Patterns of motif group labels that co-occur in many search regions:** These putative regulatory modules are ranked using genome-scale statistical and functional properties. Motifs in highly ranked patterns are likely the most reliable predictions.

In promoter-based cisRED databases, sequence search regions for motif discovery extend from 1.5 Kb upstream to 200b downstream of a transcription start site, net of most types of repeats and of coding exons. Many transcription factor binding sites are located in such regions. For each target gene's search region, we use a base set of probabilistic *ab initio* discovery tools, in parallel, to find over-represented atomic motifs. Discovery methods use comparative genomics with over 40 vertebrate input genomes.

### News

[C. elegans v4 database published](#)

January 16, 2009

The C. elegans cisRED database has been published in Nucleic Acids Research.

[C. elegans v4 tables are now public](#)

August 25, 2008

The new C. elegans database has been added to our public MySQL server.

[C. elegans v4 database](#)

July 18, 2008

This version of the C. elegans cisRED database features 8 nematode genomes and 3847 highly conserved transcripts.

[New mouse v4 database](#)

September 26, 2007

The v3.1 motif coordinates were 'lifted' to the NCBI m37 (mm9) assembly. The v4 motifs are compatible with (and will be available at) Ensembl 47.

[New human v9 database](#)

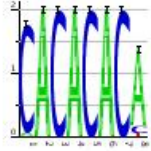
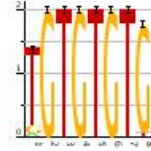
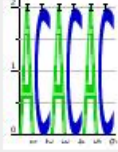
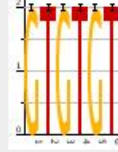
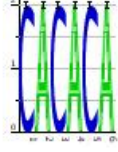
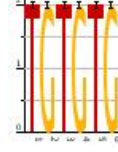
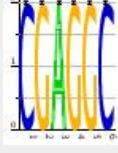
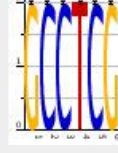
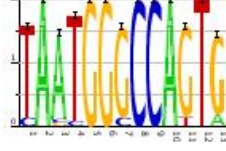
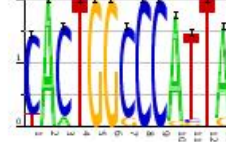
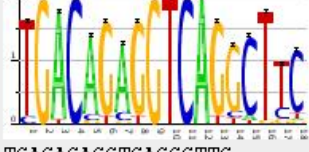
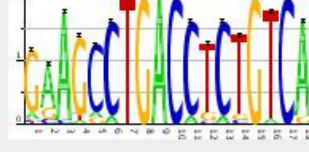
July 26, 2007

<http://www.cisred.org/>

- “For each target gene's search region, we use a base set of probabilistic ab initio discovery tools, in parallel, to find over-represented atomic motifs. Discovery methods use comparative genomics with over 40 vertebrate input genomes.”



# CisRED

Group(s) crtHsap#(name) [p-value]	Motif craHsap#	Discovery p-value	Location	Width	(+)motif	(-)motif
<b>1 annotated group(s):</b> 40083 (FOXP3) [2.94E-04]	4029	8.39E-05	chr11:31,789,274-31,789,281	8	 CACACACA	 TGTGTGTG
<b>1 annotated group(s):</b> 40083 (FOXP3) [2.94E-04]	4049	8.39E-05	chr11:31,789,275-31,789,280	6	 ACACAC	 GTGTGT
<b>1 annotated group(s):</b> 40083 (FOXP3) [2.94E-04]	4006	8.39E-05	chr11:31,789,276-31,789,281	6	 CACACA	 TGTGTG
<b>0 annotated groups</b>	3987	8.39E-05	chr11:31,789,292-31,789,297	6	 CGAGGC	 GCCTCG
<b>1 annotated group(s):</b> 40102 (HOXA5) [8.26E-04]	3967	4.07E-03	chr11:31,789,312-31,789,324	13	 TAATGGCCAGTG	 CACTGCCCCATTA
<b>6 annotated group(s):</b> 40175 (RORalpha1) [1.61E-04] 40019 (ATF3) [1.70E-04] 50071 (RORalpha-1) [2.00E-04] ...	3941	4.07E-03	chr11:31,789,327-31,789,344	18	 TGACAGAGGTCAGGCTTC	 GAAAGCTGACGCTGCTGAA

## Programs

- [AliBaba2](#)
- [BOXSHADE](#)
- [ClustalW](#)
- [Dialign2](#)
- [F-Match](#)
- [Match](#)
- [MatrixCatch](#)
- [m2transfac](#)
- [Composite Module Analyst \(CMA\)](#)
- [PolyAScan](#)
- [ReadSeq](#)
- [SignalScan](#)
- [molwSearch](#)
- [P-Match](#)
- [Patch](#)
- [SbBlast](#)
- [SnpFind](#)
- [TfBlast](#)

### Database Login

User: [dricorod](#). [Logout](#).

Name

Password

Password [forgotten?](#)

[New User?](#) [Help?](#)

[Feedback](#)

### Match - 1.0 Public

Match<sup>TM</sup> is designed for searching potential binding sites for transcription factors (TF binding sites) nucleotide sequences. Match<sup>TM</sup> uses a library of mononucleotide weight matrices from TRANSFAC<sup>®</sup> 6.0

**Authors:** Alexander Kel - BIOBASE GmbH, Ellen Goessling - BIOBASE GmbH

**License:** free for non-commercial use only

[Use Match on this site](#)

### Patch 1.0

Search for potential transcription factor binding sites in your own sequences with the pattern search program using TRANSFAC 6.0 public sites.

**Authors:** Jochen Striepe, Ellen Goessling

**License:** free only for non-commercial use

[Use Patch on this site](#)

### P-Match - Public 1.0 Public


P-Match is a new tool for identifying transcription factor binding sites (TF binding sites) in DNA sequences. It combines pattern matching and weight matrix approaches thus providing higher accuracy of recognition than each of the methods alone. P-Match uses a library of mononucleotide weight matrices from TRANSFAC<sup>®</sup> 6.0 along with the site alignments associated with these matrices.

**Authors:** Dmitry Chekmenev, Carla Haid and Alexander Kel - BIOBASE GmbH

[Use P-Match on this site](#)

RSAT

NeAT



Regulatory Sequence Analysis Tools

Most popular tools

retrieve sequence

retrieve sequence Ensembl **New!**

oligo-analysis (words)

matrix-scan (matrices)

random sequence

> view all tools

Genomes and genes

**New!**

Sequence retrieval

Pattern discovery

Pattern matching

Comparative genomics

Conversion/Utilities

Drawing

Web services



Regulatory Sequence Analysis Tools

BIGRe - ULB



CCG

Centro de Ciencias Genómicas

Laboratorio de Biología Computacional

UNAM/CCG

[Tool Map](#)
[Introduction](#)
[Forum \(New\)](#)
[Tutorials](#)
[Publications](#)
[Credits](#)
[Data](#)
[Links](#)

Welcome to **Regulatory Sequence Analysis Tools (RSAT)**. This web site provides a series of modular computer programs specifically designed for the detection of regulatory signals in non-coding sequences.

**New ! RSAT Forum** now available (December 2008)

**New !** Four articles explaining how to use RSAT and NeAT published in **Nature Protocols** (Sept 2008)

**New ! RSS feed available:** get RSAT latest news directly in your favorite RSS reader ! (Sept 2008)

**New !** The **recent developments** made in RSAT are presented in the **NAR Web server issue 2008** [Free PDF] (July 2008)

This website is free and open to all users.

**Warnings**

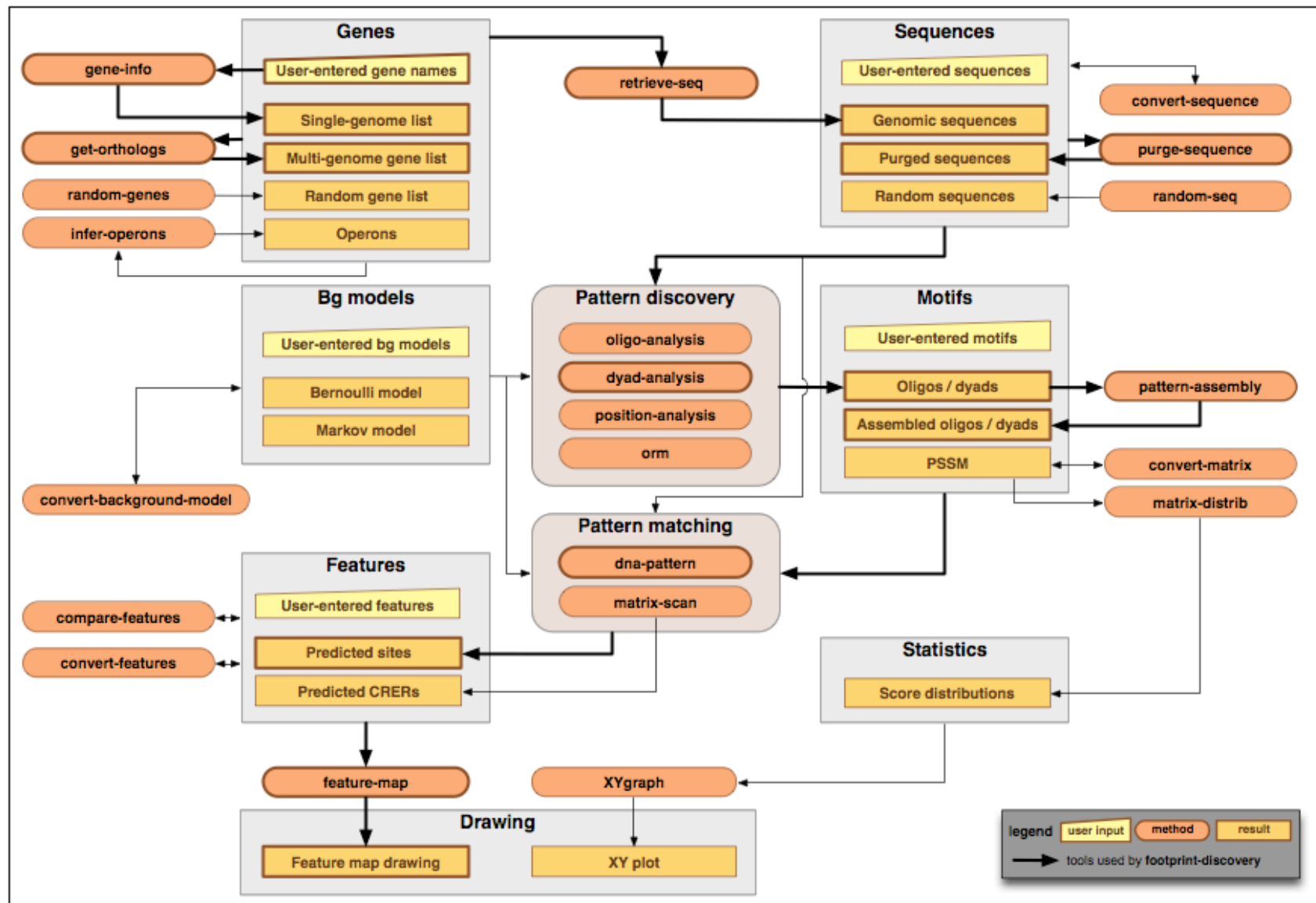
**Vertebrate genomes**

---

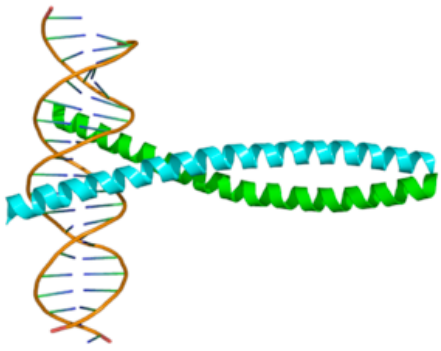
**Regulatory Sequence Analysis Tools - Web servers**

<http://rsat.ulb.ac.be/rsat/>

## RSA-tools - Map of the tools



<http://rsat.ulb.ac.be/rsat/>

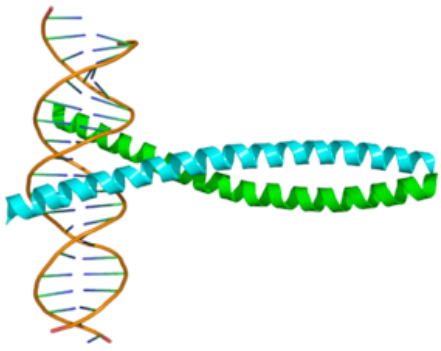


Thank you!

And thanks a lot for some of the slides to:

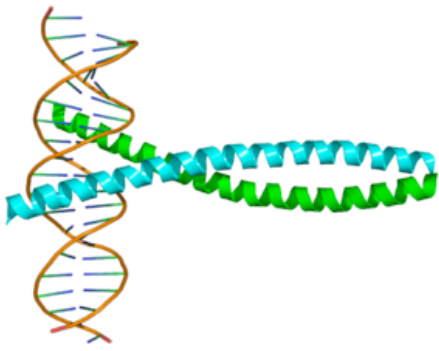
Philippe Gautier, Bioinformatics  
MRC-Human Genetics Unit  
Edinburgh UK

[http://www.hgu.mrc.ac.uk/Users/Philippe.Gautier/tfbs\\_seminar/noncoding.html](http://www.hgu.mrc.ac.uk/Users/Philippe.Gautier/tfbs_seminar/noncoding.html)



## Transcription Factor Binding Sites

1. Promoters and gene regulation in Eukaryotes
2. Position Weight Matrices (PWM)
3. PWM Databases
4. Pattern Matching: TFBS prediction using PWMs
5. Pattern Discovery: Finding unknown motifs
6. Exercise: Obtain mouse and human fosB promoters  
and predict TFBS with Match and JASPAR



## EXERCISE

### Step by step

- a. Go to <http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=searchPromForm> and retrieve the mouse and human fosB promoters (-1500..299)
- b. Save your promoters as .fasta files.
- c. Go to Match and search for TFBS in the mouse promoter with the defaults. Change the options to minimize false negatives.
- d. Go to JASPAR and search for TFBS in the mouse promoter with the defaults.
- e. Compare the results.