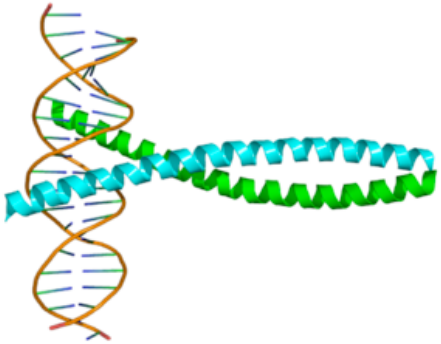
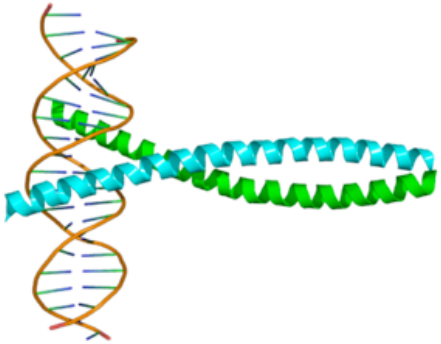


Transcription Factor Binding Sites

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

Daniel Rico, PhD.
drico@cnio.es



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

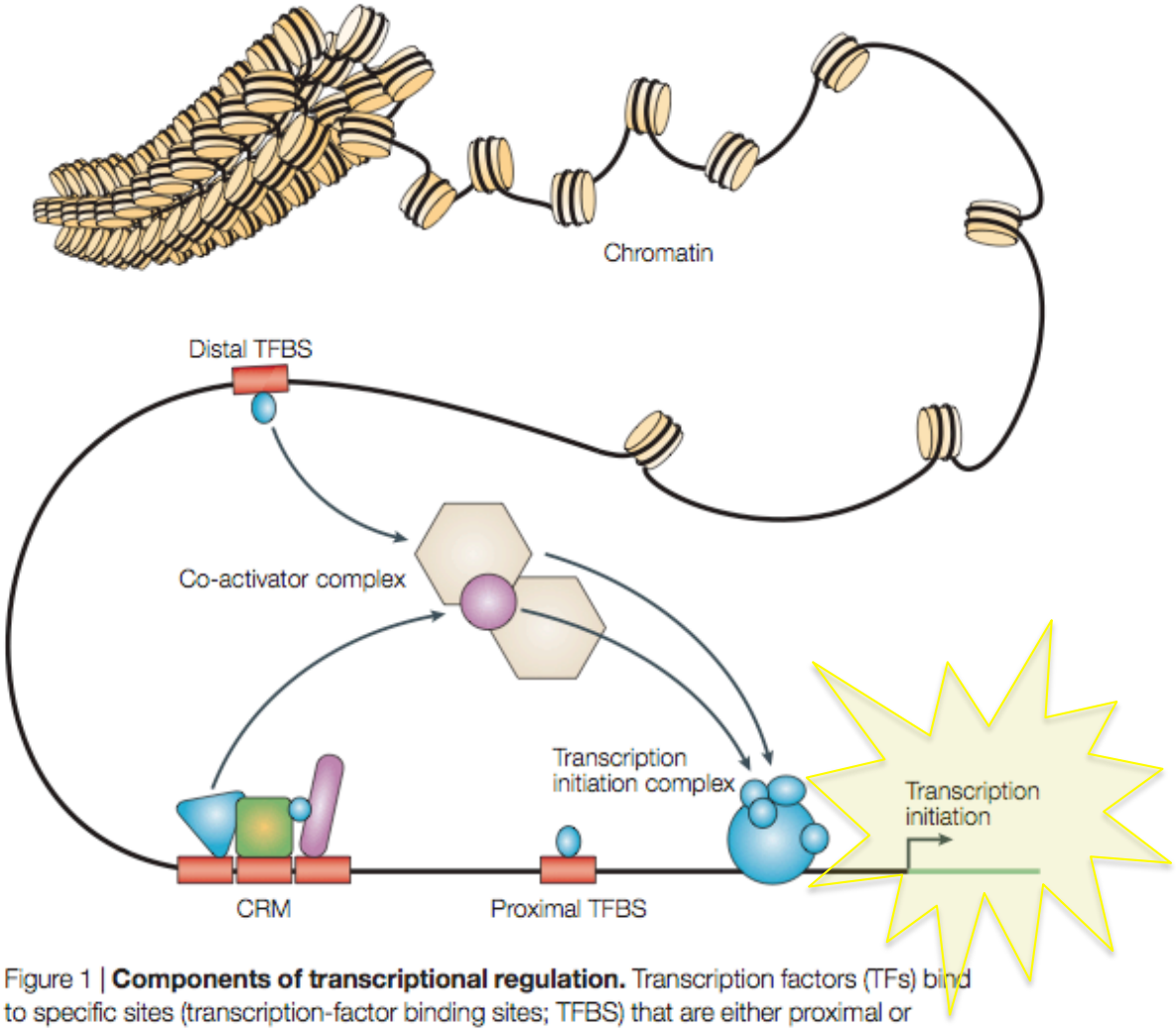


Figure 1 | **Components of transcriptional regulation.** Transcription factors (TFs) bind to specific sites (transcription-factor binding sites; TFBS) that are either proximal or distal to a transcription start site. Sets of TFs can operate in functional *cis*-regulatory modules (CRMs) to achieve specific 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

Quick Search

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/

Zhang Lab:
Computational Biology and Bioinformatics

Home
People
Databases and Software Tools
Collaborations
Courses
Seminars
Meetings and Workshops
Publications
Links
Positions Available

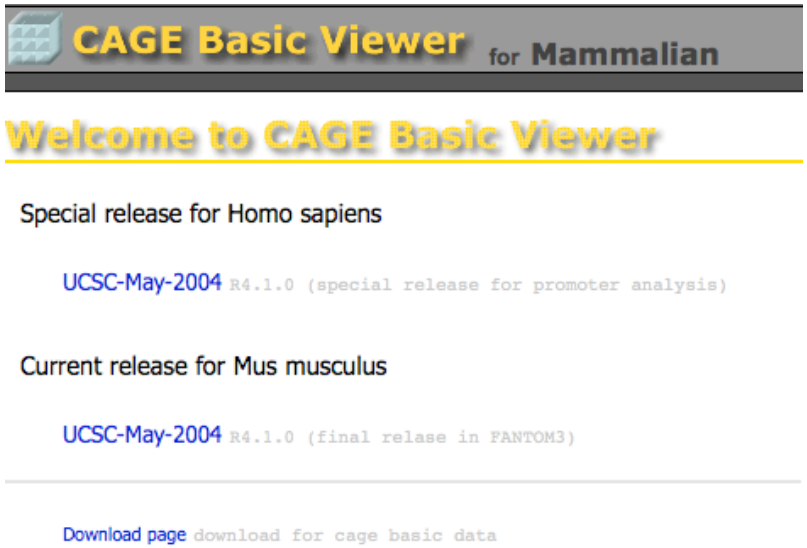
Databases

- 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)

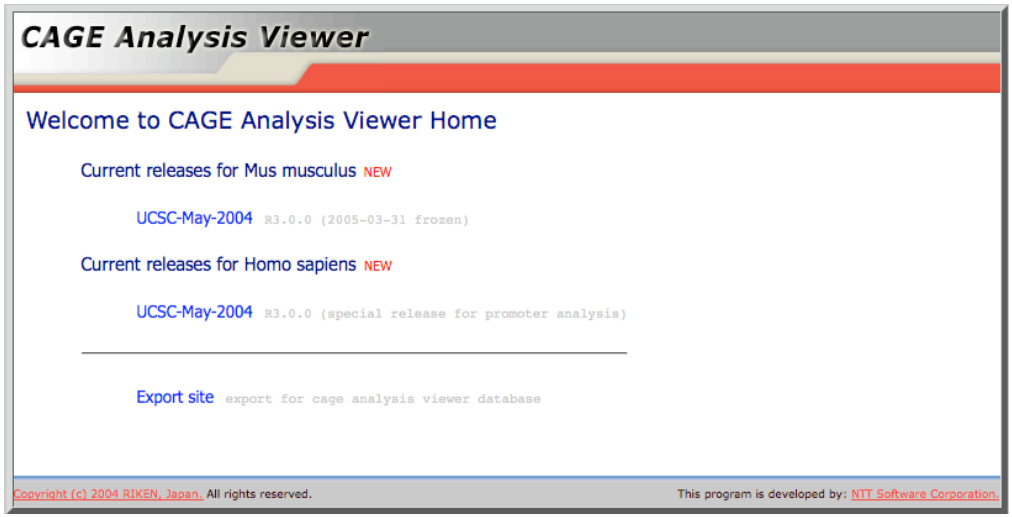
Software Tools

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

Experimental Transcription Start Sites (TSS)



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

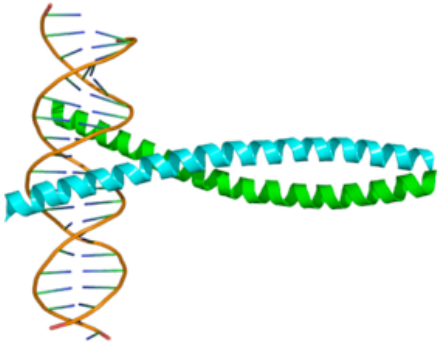


http://gerg01.gsc.riken.jp/cage_analysis/

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

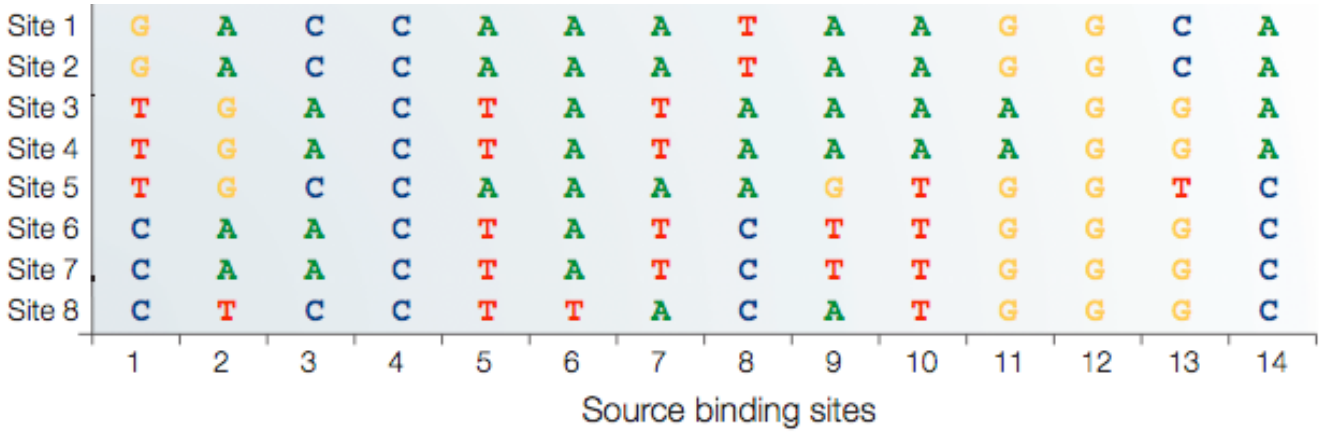
Hideya Kawaji[1], Takeya Kasukawa[1,2], Shiro Fukuda[2], Shintaro Katayama[2,*], Chikatoshi Kai[2], Jun Kawai[2,3], Piero Carninci[2,3] and Yoshihide Hayashizaki[2,3]



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



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

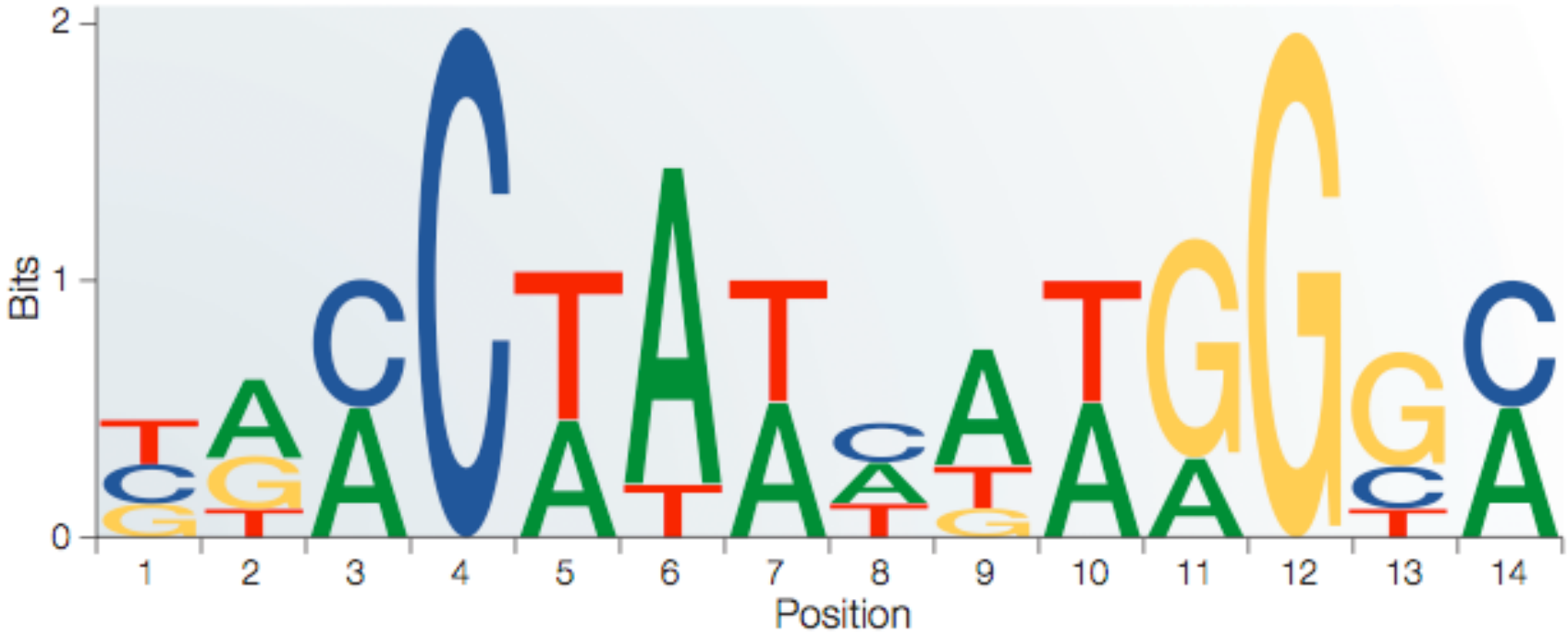
$f_{b,i}$ = counts of base b in position i; N = number of sites; $p(b,i)$ = corrected probability of base b in position i; $s(b)$ = pseudocount function

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



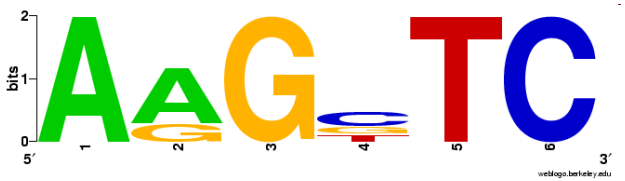
SEQUENCE LOGOS: The information content of a matrix column ranges from 0 (no base preference) and 2 (only 1 base used).

http://weblogo.berkeley.edu/ http://www.lecb.ncifcrf.gov/~toms/sequencelogo.html

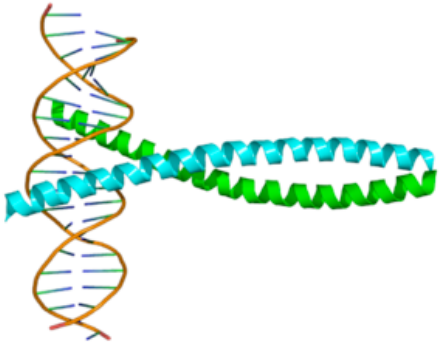
Summary

AAGTTC
AAGCTC
AGGCTC
AAGGTC

A	430000
C	000204
G	014100
T	000140



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®

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
Accession Number	M01118
Accession numbers, secondary	M00709
Identifier	V$WT1_Q6
Created Updated	02.06.2006 by dtc.
Copyright	Copyright (C), Biobase GmbH.
Name	WT1
Binding factors	T00899; WT1; Species: human, Homo sapiens. T02351; WT1; Species: mouse, Mus musculus. T02352; WT1; Species: rat, Rattus norvegicus. T01839; WT1 -KTS; Species: human, Homo sapiens. T01840; WT1 I; Species: human, Homo sapiens. T00900; WT1 I -KTS; Species: human, Homo sapiens. T01842; WT1 I-del2; Species: human, Homo sapiens. T01841; WT1-del2; Species: human, Homo sapiens. T09249; WT1-isoform1; Species: mouse, Mus musculus.

Binding Matrix

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

Basis

Binding sites

23 compiled sequences

Sequence	Derived from	Start	Length	Gaps	Orientation
gcctcacnn	R02307	-2	9		n.
nnCTCCCTC	R02308	-2	9		p.
cacacannn	R02309	-3	9		n.
cacacatac	R02310	2	9		n.
cacaccctc	R02311	3	9		n.
CACTCCAGG	R02312	7	9		p.
CGCCCCCGC	R02313	1	9		p.
gcccccgca	R02314	-1	9		n.
CGCCCCCGC	R02315	1	9		p.
cccacccgc	R04858	-4	9		n.
agcccacgc	R04859	1	9		n.
gcccccgcg	R04860	-1	9		n.
gcccccgcg	R04861	-1	9		n.
CACGCCCGC	R04862	-2	9		p.
ccctcctcc	R04863	-4	9		n.
ggctccggc	R04864	-5	9		n.

Transfac example: WT1

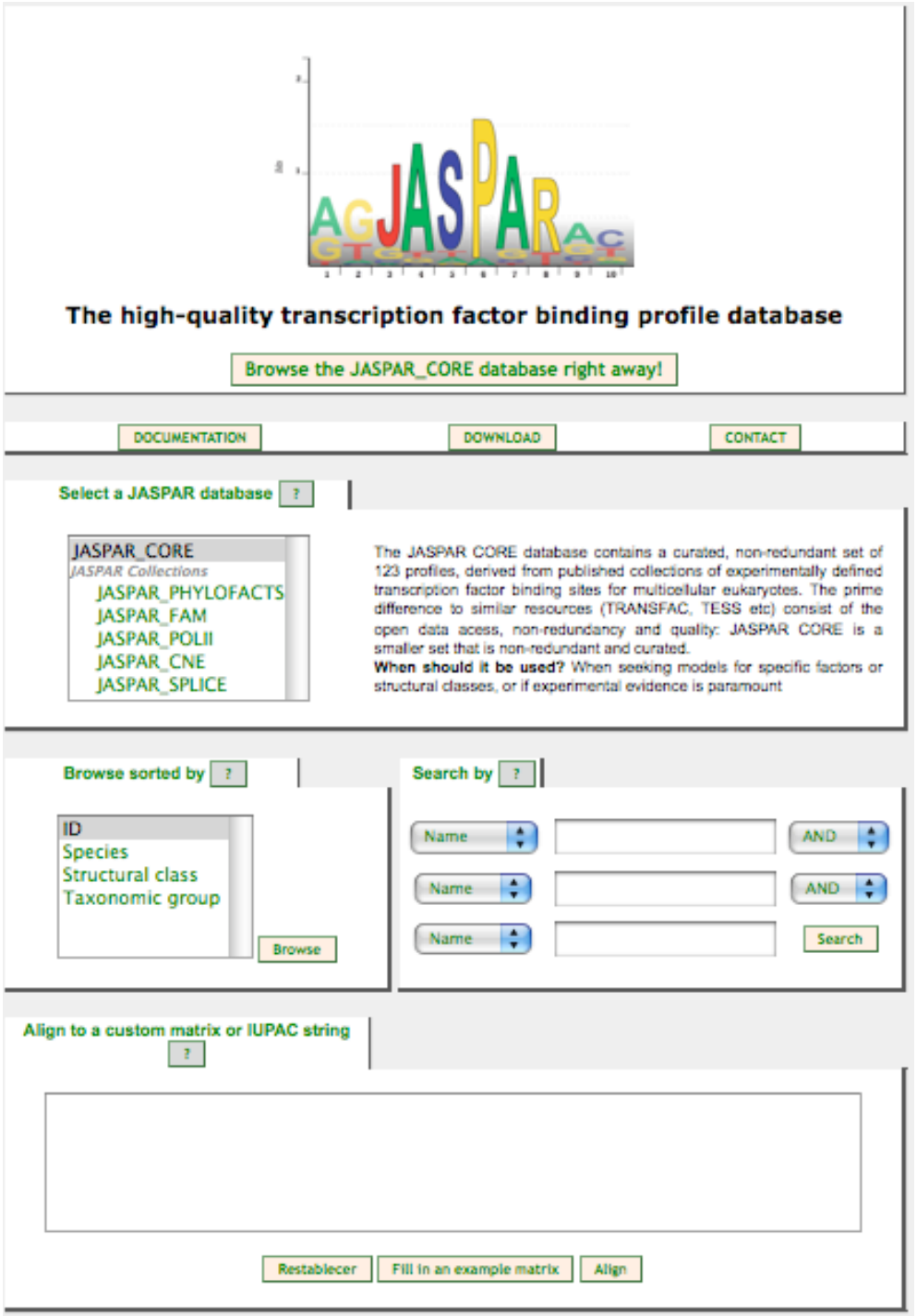
Accession Number	R04859
Identifier	RAT$IGF1R_02
Created Updated	05.03.1998 by ili. 08.11.2000 by vma.
Copyright	Copyright (C), Biobase GmbH.
Sequence type	DNA
Description	IGF-1 receptor (insulin-like growth factor I receptor); Gene: G001118.
Species	rat, Rattus norvegicus
axonomic classification	eukaryota; animalia; metazoa; chordata; vertebrata; tetrapoda; mammalia; eutheria; rodentia; myomorpha; muridae; murinae
Sequence	GCGTGGGCT
rst position of element	-250
te terminating position	-242
Binding factors	T00899; WT1; Quality: 2; Species: human, Homo sapiens. T01839; WT1 -KTS; Quality: 2; Species: human, Homo sapiens. T01840; WT1 I; Quality: 2; Species: human, Homo sapiens. T00900; WT1 I -KTS; Quality: 2; Species: human, Homo sapiens.
Matrices	M01118; V$WT1_Q6
Cellular factor source	0306; rec(human-E.coli).
Method(s)	DNase I footprinting
Comments	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 [1
xternal database links	EMBL: M37807; RNIGFIRC (168:176). TRANSPRO: RNO_10607.
Reference number uthor(s), Title, Journal	[1]; RE0006666. PUBMED: 8175666. 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

Accession Number	R16085
Identifier	RAT$EGFR_01
Created Updated	02.12.2004 by abd2. 17.03.2005 by oke.
Copyright	Copyright (C), Biobase GmbH.
Sequence type	DNA
Description	EGFR (epidermal growth factor receptor); Gene: G009653.
Species	rat, Rattus norvegicus
Taxonomic classification	eukaryota; animalia; metazoa; chordata; vertebrata; tetrapoda; mammalia; eutheria; rodentia; myomorpha; muridae; murinae
Gene region	promoter
Sequence	cTCCTCCTCCacttagtcccTCCTCCTCCcgcccaacctcccccacgtcccgaccaggg
Element	NGF-responsive
First position of element	-318
Site terminating position	-260
Binding factors	T02352; WT1; Quality: 3; Species: rat, Rattus norvegicus.
Matrices	M01118; V$WT1_Q6
Cellular factor source	0925; PC12+NGF. 1729; PC-12.
Method(s)	direct gel shift functional analysis gel shift competition supershift (antibody binding)
Comments	TCC TCCTCC repeat sequences are required for down-regulation of rat EGFR by NGF in PC12 cells and is mediated through WT1 [1] [2]
External database links	EMBL: AF142153; AF142153 (2407:2465). TRANSPRO: RNO_3204.
Reference number	[1]; RE0025181. PUBMED: 11071895.
Author(s), Title, Journal	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



http://jaspar.cgb.ki.se/

Jaspar example: Pax6

Summary page for ID: MA0069 NAME: Pax6 from the JASPAR_CORE database ?

DATA	
name	Pax6
class	PAIRED
species	Homo sapiens
total_ic	13.7960
sysgroup	vertebrate
acc	P26367
type	SELEX
medline	8132558
comment	

SEQUENCE LOGO ?

Make a SVG logo ?

FREQUENCY MATRIX ?

```
A  [  2   2   4  39   3   1   1  21   1   2  36  11   1   1 ]
C  [  4   2  26   2  34   0  37   2   4  14   0  11   5   0 ]
G  [  4   0   1   1   1  41   4   2   1  25   6  13   3  17 ]
T  [ 33  39  12   1   5   1   1  18  37   2   1   8  34  25 ]
```

Reverse complement ?

EXPECTED PREDICTIONS/BP ?

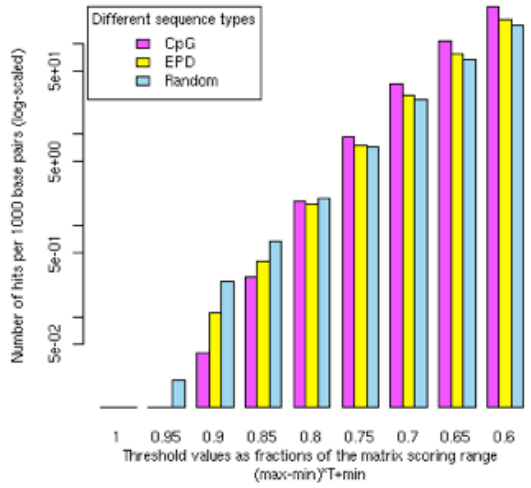
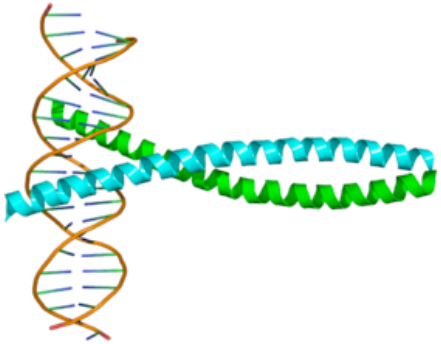


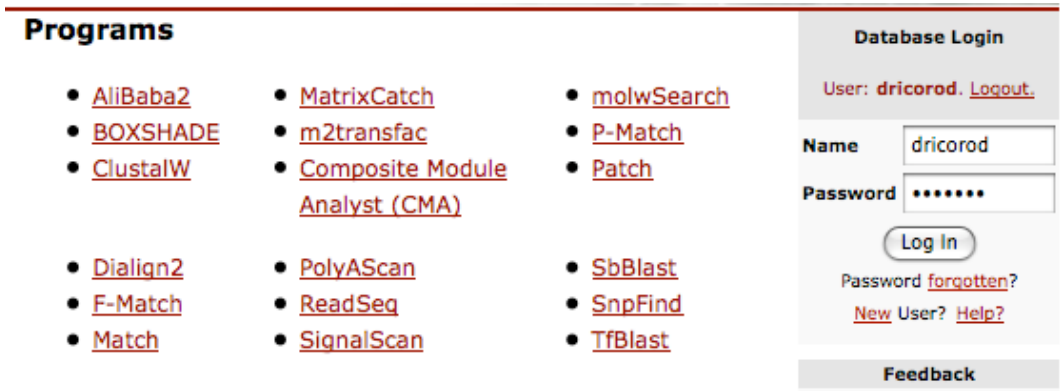
Table of number of hits per 1000 base pairs for each sequence type

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



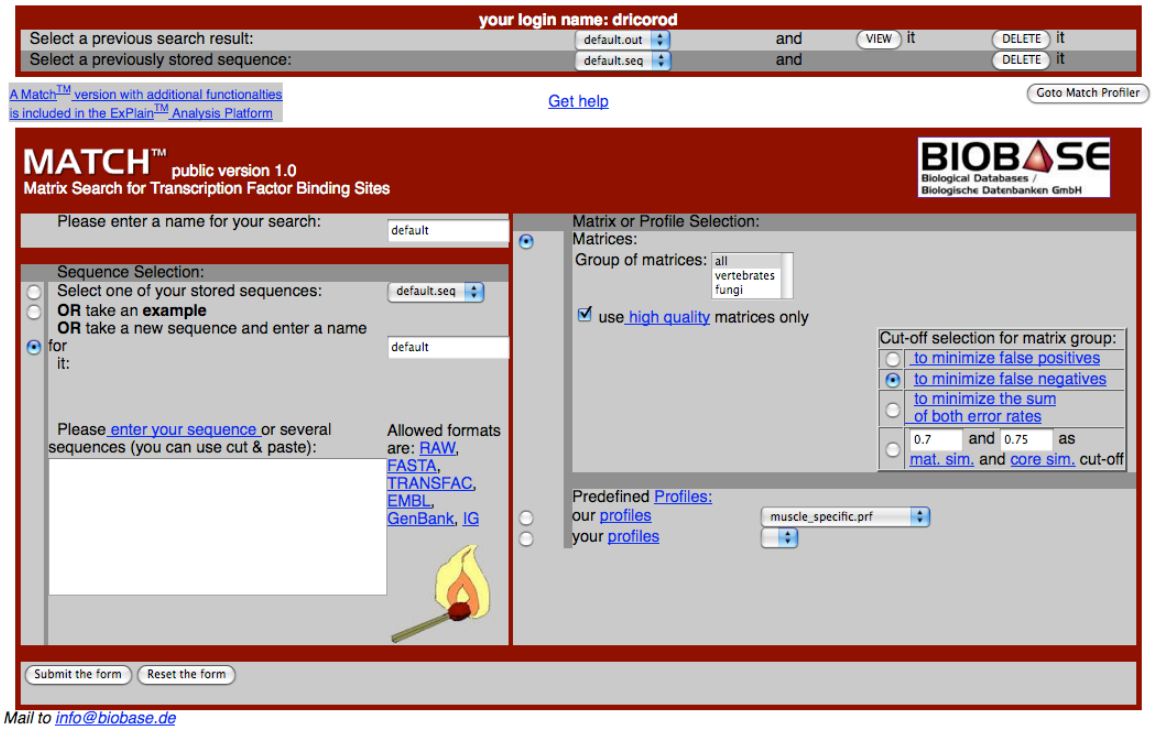
http://www.gene-regulation.com/cgi-bin/pub/programs/match/bin/match.cgi

Match - 1.0 Public

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

Authors: Alexander Kel - BIOBASE GmbH, Ellen Goessling - BIOBASE GmbH
License: free for non-commercial use only

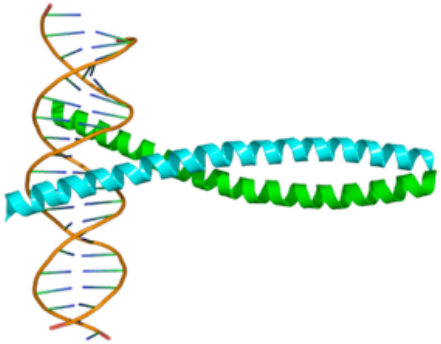
Use Match on this site



The high-quality transcription factor binding profile database

Browse the JASPAR_CORE database right away!

Click here to select all TFBS

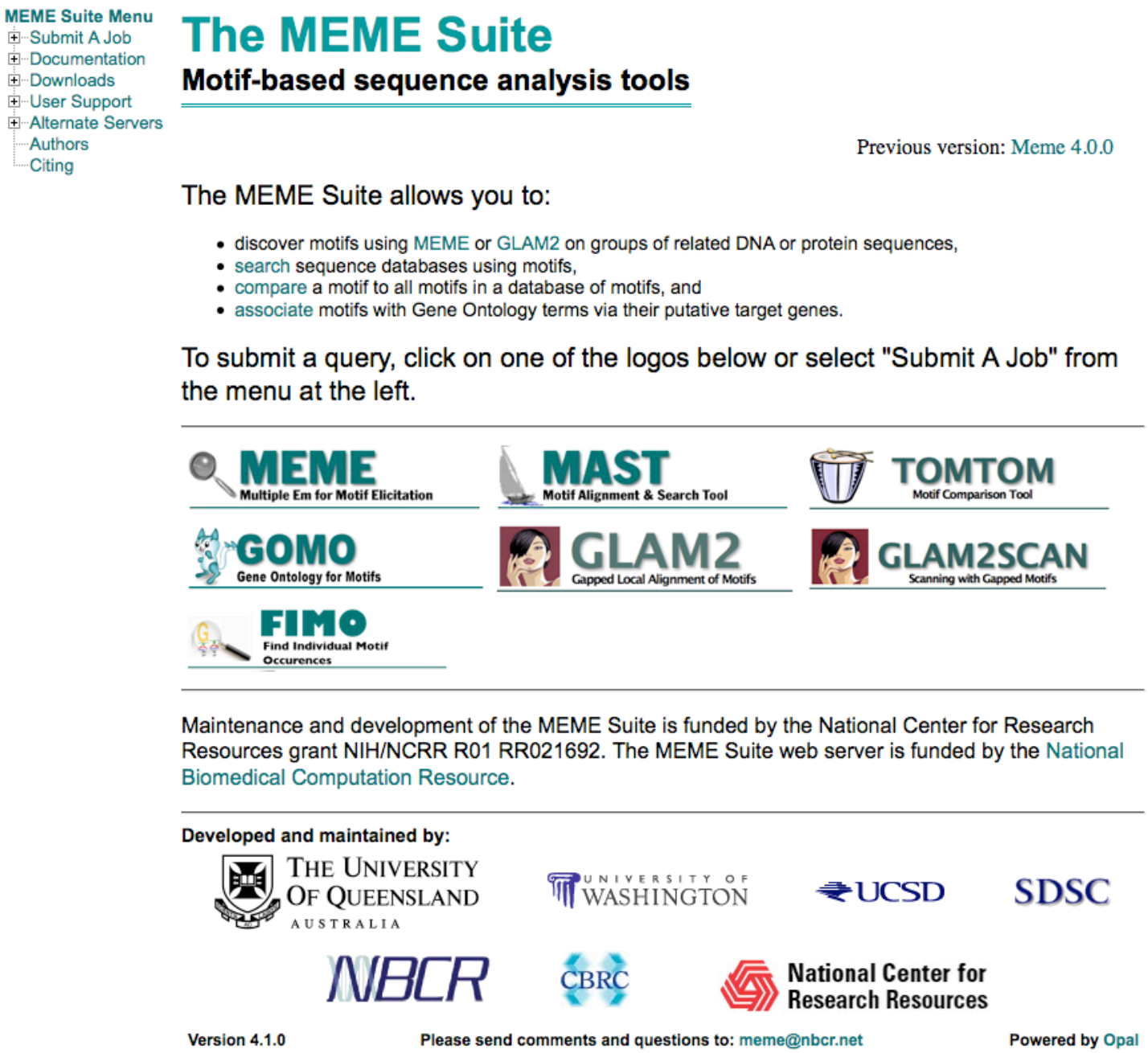


Transcription Factor Binding Sites

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	
…		…		



http://meme.sdsc.edu/meme/

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

cisRED

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
Human 9	NCBI v36b	18.7k	promoter	41	236k	0.01	Build 38-49	26 Jul. 2007
Mouse 4	NCBI m37	17.5k	promoter	38	223k	0.1	Build 47-49	26 Sep. 2007
Mouse 3.1	NCBI m35	17.5k	promoter	38	223k	0.1	Build 38	18 Apr. 2007
Rat 1.1	RGSC v3.1	6.7k	promoter	28	116k	0.25	n/a	12 Feb. 2006
C.elegans 4	WormBase WS170	3.8k	promoter	8	158k	1.0	Build 44-46	18 Jul. 2008
Human Stat1 ChIP-seq peaks 1	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:

1. **'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.
2. **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).
3. **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

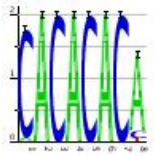
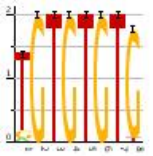
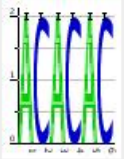
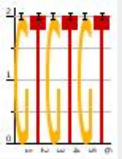
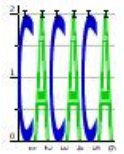
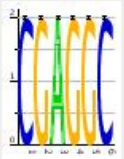
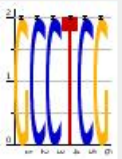
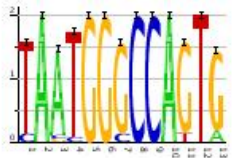
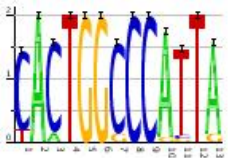
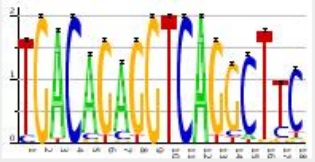
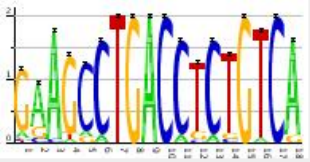
The v3.1 motif coordinates were 'lifted' to the NCBI m37 (mm9) assembly. The v4 motifs are compatible with (and will be available at) Ensembl 47.

New human v9 database — July 26, 2007

http://www.cisred.org/

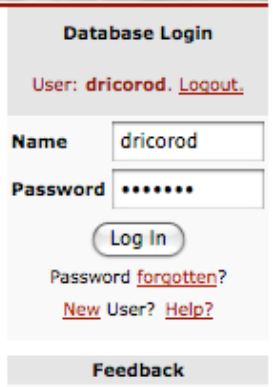
- “For each target gene's search region, we use a base set of probabilistic ab initio discovery tools, in parallel, to find over-represented atomic motifs. Discovery methods use comparative genomics with over 40 vertebrate input genomes.”

CisRED

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

Programs

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



Match - 1.0 Public

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

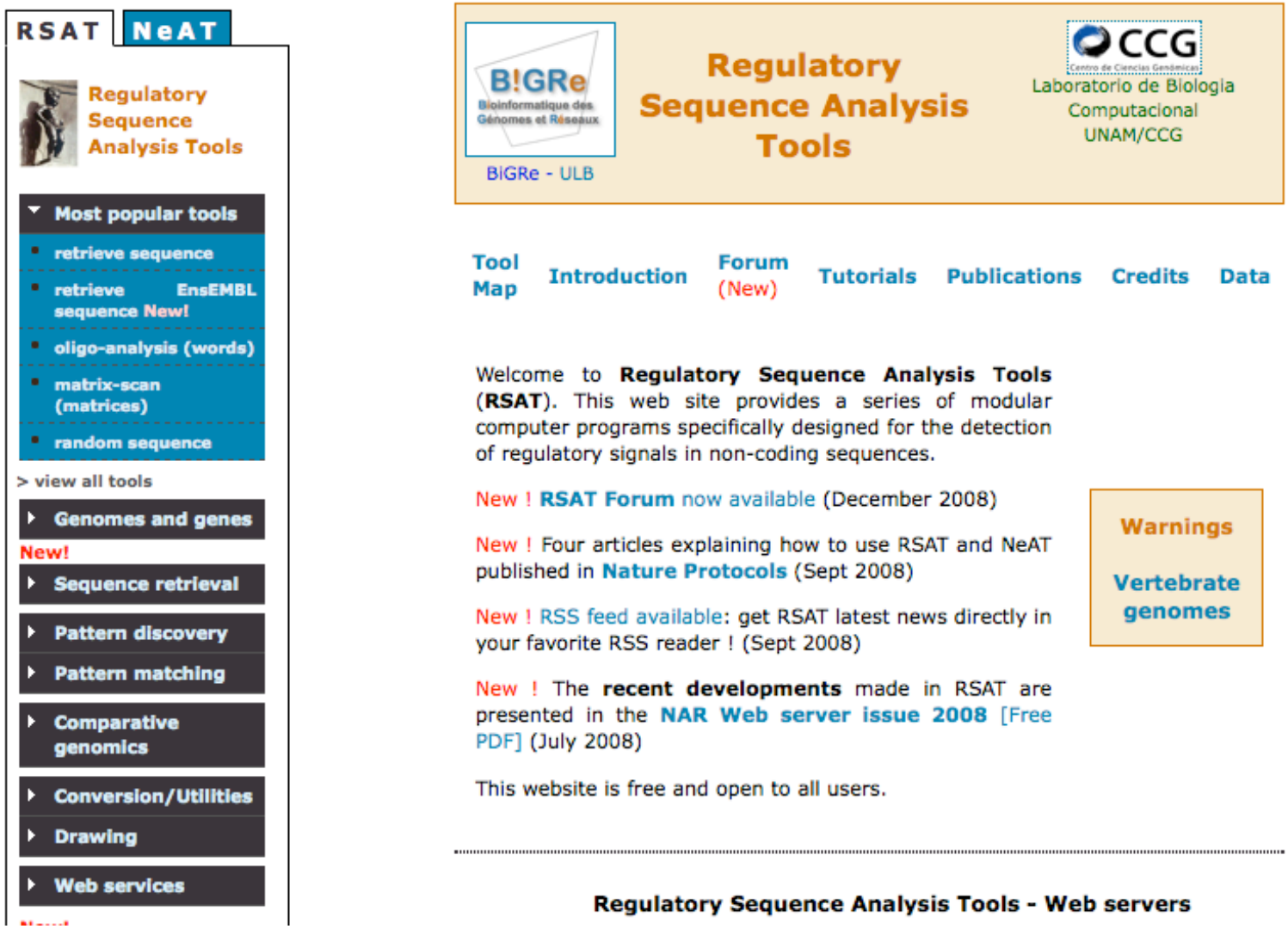
Use Patch on this site

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® 6.0 along with the site alignments associated with these matrices.

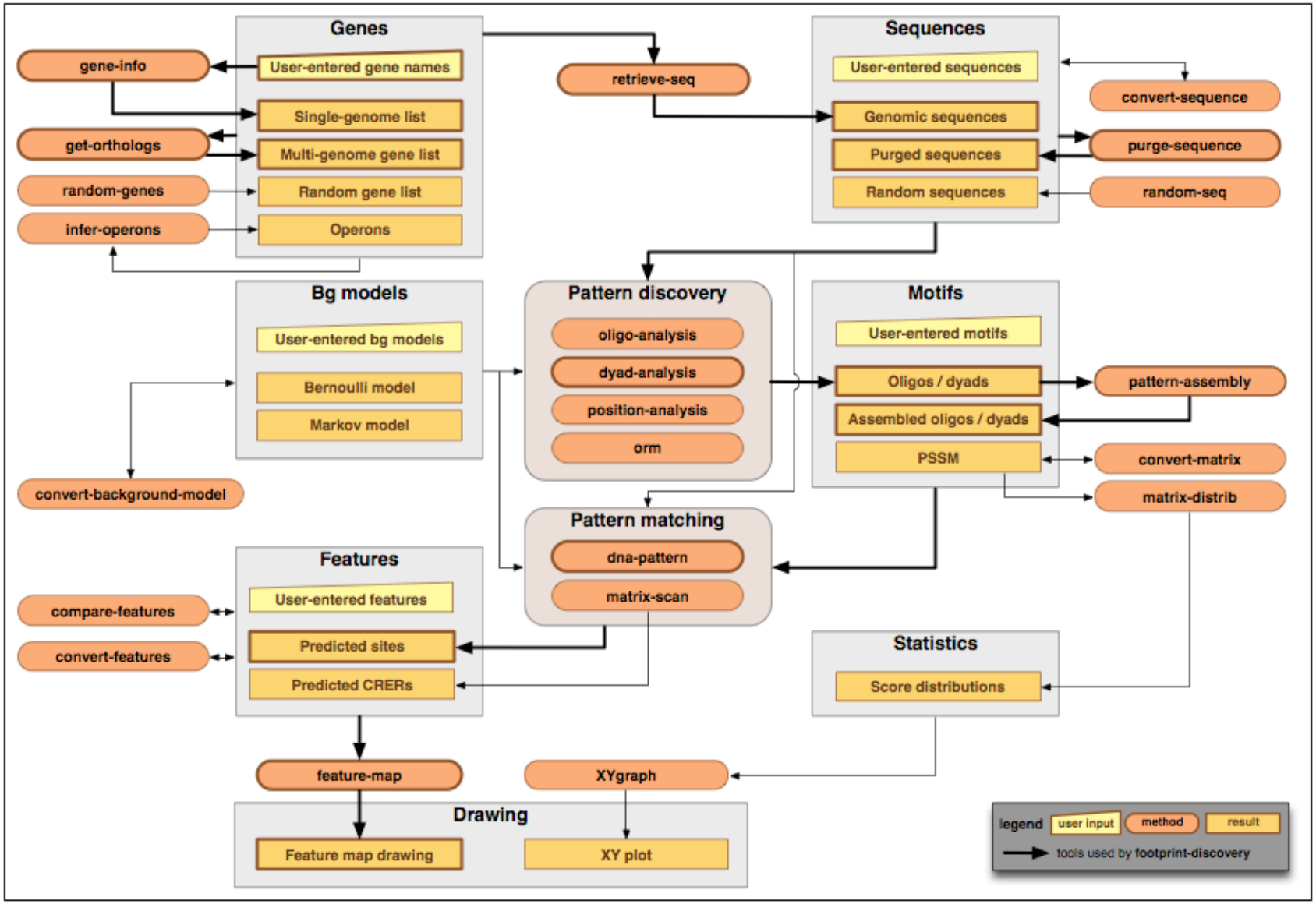
Authors: Dmitry Chekmenev, Carla Haid and Alexander Kel - BIOBASE GmbH

Use P-Match on this site

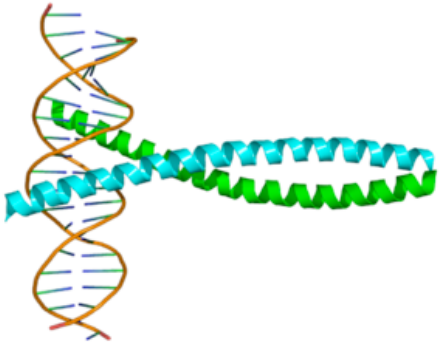


http://rsat.ulb.ac.be/rsat/

RSA-tools - Map of the tools



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

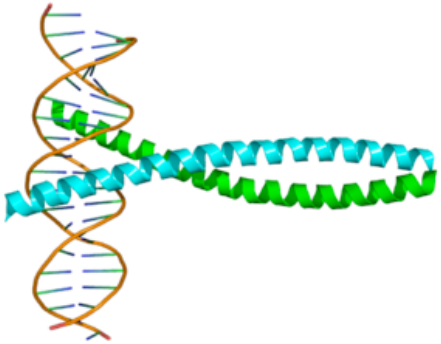


Thank you!

And thanks a lot for some of the slides to:

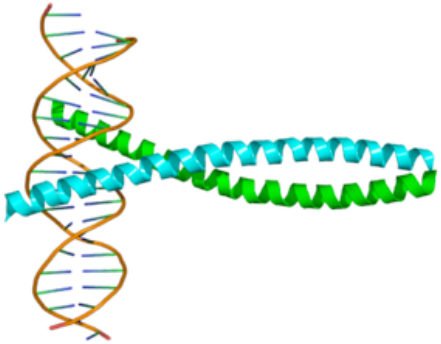
Philippe Gautier, Bioinformatics
MRC-Human Genetics Unit
Edinburgh UK

http://www.hgu.mrc.ac.uk/Users/Philippe.Gautier/tfbs_seminar/noncoding.html



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.