

# *In silico* identification of transcriptional regulatory regions

Martti Tolvanen, IMT Bioinformatics, University of Tampere

Eija Korpelainen, CSC

Jarno Tuimala, CSC

## Program

- Introduction (Eija)
- Retrieval of promoter sequences (Martti)
- Detection of binding sites for known transcription factors (Eija)
  - binding site presentation
  - transcription factor databases
  - programs for matrix scans
- Phylogenetic footprinting and clusters (Eija)
- Pattern discovery from promoters of co-expressed genes (Martti and Jarno)

## Aims of the course

- introduce resources available for *in silico* analysis of transcriptional regulation
- ensure realistic expectations
- facilitate participation in the advanced course
- collect opinions on the different software and databases

## Terms

- Promoter
  - sufficient for transcription initiation, includes transcription start site (TSS)
- Proximal regulatory region
  - adjacent to promoter
- Distal regulatory region
  - further 5' or 3' from TSS
- Module = composite regulatory element, set of transcription factor binding sites (TFBS) that function together
- Framework = conserved set of transcription elements, in a same order and distance range

## in reality

- regulatory elements can be far
- binding sites have variability
- chromatin structure is important
- TFs work together

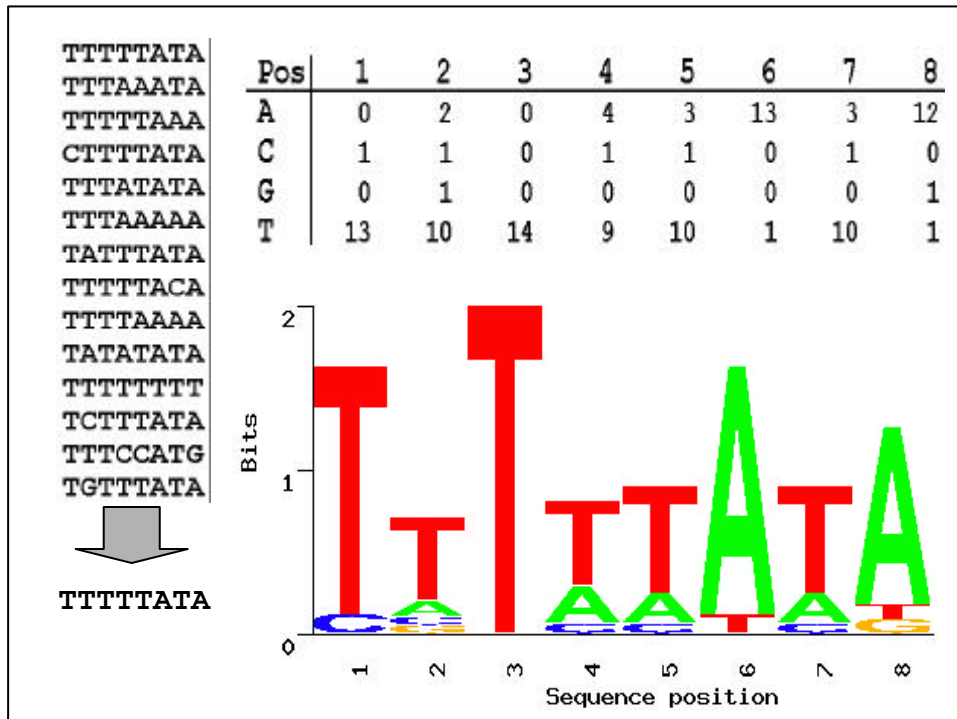
## in silico

- large search space ☹️
- lot of random hits ☹️
- chromatin structure is ignored ☹️
- module search improves specificity 😊

## 2. Retrieval of promoter sequences (Martti)

### 3. Detection of binding sites for known transcription factors

Presentation of TF binding sites



- consensus sequence
  - shows tolerance at each position
- matrix (profile)
  - shows preference at each position
- sequence logo
  - shows information content at each position

## From frequency to weight matrix

A	5	0	1	0	0
C	0	2	2	4	0
G	0	3	1	0	4
T	0	0	1	1	1

- add pseudocounts
- divide by background base frequencies
- convert to log-scale

A	1.6	-1.7	-0.2	-1.7	-1.7
C	-1.7	0.5	0.5	1.3	-1.7
G	-1.7	1.0	-0.2	-1.7	1.3
T	-1.7	-1.7	-0.2	-0.2	-0.2

TGCTG = 0.9

- assumes that positions are independent
- some programs weight positions based on info content

## Matrix performance

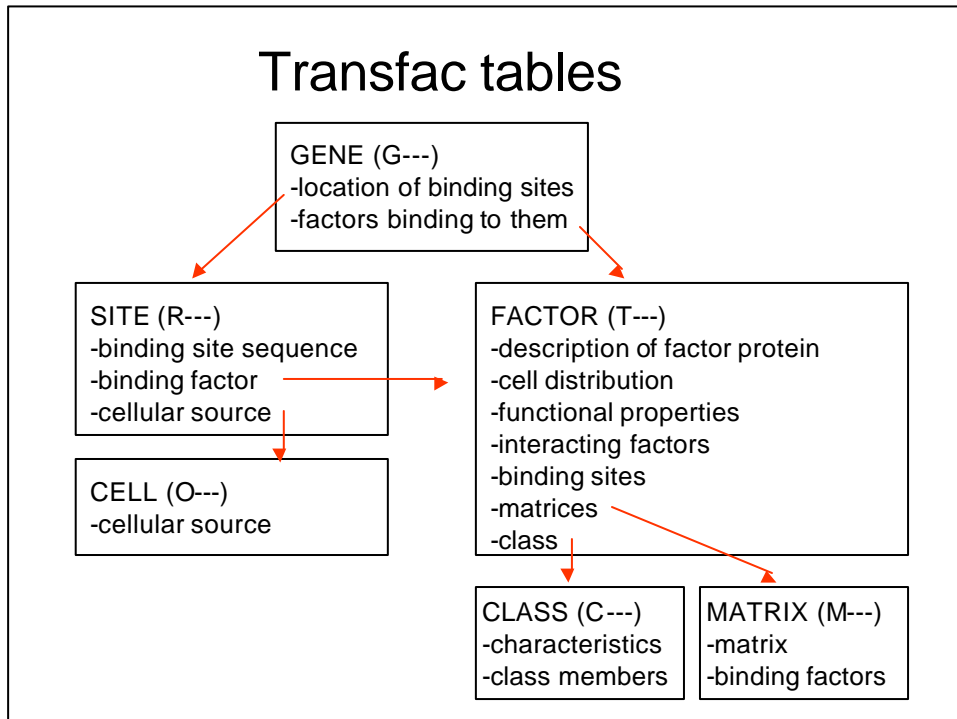
- because TFBSs are short and degenerate, predicted sites occur frequently at random (eg. MyoD sites 1 per 600 bp)
- sensitivity is good (real sites are not missed) but specificity is awful (a lot of false positives)
- nearly 100% of predicted sites have no function *in vivo*, because binding is regulated by chromatin structure and cooperativity
- best for one factor and a short piece of DNA.

## Transcription factor databases

### The "Transfac family"

- consists of several databases and programs
- TRANSFAC
  - transcription factors, known binding sites in promoters, matrices
  - public v. 6.0 contains 336 matrices, professional v. 7.4.1 contains 695 matrices
- TRANSCOMPEL
  - composite regulatory elements (pairs of closely situated sites and transcription factors binding to them)
- TRANSPATH
  - pathways involved in regulation of transcription factors
- Programs Match (for matrices), Patch (for sites) and Catch (for Transcompel searches)
- the latest public version is free for academics, registration required:  
<http://www.gene-regulation.com/pub/databases.html>
- commercial site: [www.biobase.de](http://www.biobase.de)

# Transfac tables



TRANSFAC GENE TABLE, Release 6.0 - public - 2002-08-01, (C) Biobase GmbH - Microsoft Internet Explorer

Address: <http://www.gene-regulation.com/cgi-bin/pub/databases/transfac/getTF.cgi?AC=6000348>

**TRANSFAC GENE TABLE, Release 6.0 - public - 2002-08-01, (C) Biobase GmbH**

```

AC 6000348
XX
ID HS{MYOGL
XX
DT 14.06.1995 [created]; cbo.
DT 13.05.1997 [updated]; cbo.
CO Copyright (C), Biobase GmbH.
XX
SD
XX
DE myoglobin
XX
SY PVALE.
XX
OS human, Homo sapiens
OC eukaryota; animalia; metazoa; chordata; vertebrata; tetrapoda; mammalia;
OC eutheria; primates
XX
CH 22q11.2-qter
XX
BC 6.1.3.10.1
XX
ES -230 -206 R04162; HS{MYOGL_01.
ES -174 -155 R09235; HS{MYOGL_02; Binding factors: MEF-2 TO1005,
ES MEF-2DAB TO1770.
XX
DE EHEL: 300371; HSN001.
DE EHEL: R10090; HSN011.
DE TRANSPATH: 6000348.
DE LOCUSLINK: 4151.
DE ONIN: 160000.
DE REFSEQ: NM\_005368.
DE TRRD: 00042.
XX
  
```

**GENE TABLE**

site

factor

**SITE TABLE**

```

ID   H89HY00L_O2
XX
DT   24.08.1000 (created); dkl.
DT   22.03.1001 (updated); dkl.
CO   Copyright (C), Biobase GmbH.
XX
TY   D
XX
DE   Hb (myoglobin); 0000248.
XX
EE   Promoter
XX
SQ   CCTAAATAGCTTCC.
XX
EL   HEF-2
XX
SE   -174
ST   -155
XX
RF   T01005 HEF-2; Quality: 2; Species: human, Homo sapiens.
RF   T01770 HEF-2&A; Quality: 2; Species: human, Homo sapiens.
XX
OS   human, Homo sapiens
OC   eukaryota; animalia; metazoa; chordata; vertebrata; tetrapoda; mammalia;
OC   eutheria; primates
XX
SO   0516 8a18
XX
NM   direct gel shift
NR   supershift (antibody binding)
XX
CC   Myoglobin A/T element is capable for binding of HEF-2 proteins, but this
CC   nonconsensus site binds HEF-2 with lower affinity than do consensus sites
CC   present in other genes. Myoglobin A/T box also binds in vitro to a
CC   different protein, termed ATF35.
XX
DE   TRANSPATH: XXXX0006388.
DE   TRANSPATH: XXXX0008999.
DE   EMBL: XXXX171; M8MG01 (2411:2426).
XX
EN   [1]
EA   Grayson J., Williams R. S., Yu Y.-Y., Hassell-Duby B.
ET   Synergistic interactions between heterologous upstream activation elements
ET   and specific TATA sequences in a muscle-specific promoter
  
```

factor

cellular source

**FACTOR TABLE**

```

ID   T01005
XX
DT   24.01.1994 (created); swi
DT   09.11.2000 (updated); dkl
CO   Copyright (C), Biobase GmbH
XX
FA   HEF-2
XX
ST   HEF-2; HEF-2A; HEF2.
XX
OS   human, Homo sapiens
OC   eukaryota; animalia; metazoa; chordata; vertebrata; tetrapoda; mammalia;
OC   eutheria; primates
XX
HO   D-HEF2 (Drosophila)
XX
CL   00014 BABS; 0.4.1.1.1.1.
XX
SQ   507 AA; 54.9 kDa (cDNA)
XX
SQ   MSKREIQTRIMEKSNQQTPTTRKPOLKREAVELDVLCDEKIALIIFKSNRKLQVAST
SQ   DMHVVLLVYTYMFMKESIMSIIVEALNKEKSPGCDSDPDTSYVLTNTEKFKYKINE
SQ   EFINNHPNHRKIAQGLPQGFSEKVTYVFTSFMALVTRPGRSLVSPGLAAGTLDKSHL
SQ   SFFQTLMEKSPGASQFPDSTMGAGHLSITLTLVRRGAGSSVWNRGPMNSRASPMLIG
SQ   ATCANGLGRKPTVRSPPFGGHLGMSKEDLKVIFPDRGMSPLKELKLELNTGR
SQ   ISSGATGCLATVPSVTFKSLFPGVYSARFATNFTYSLTSLALSKLQKPSKCHLS
SQ   LQKSRAPQHLGQALSLVAGQGLSQGSHLSINTWQNTIIEEPISPPQDNTDSSQ
SQ   QQQQQQQQQQVFFVCFQCFQCFQCFQCFQCFQCFQCFQCFQCFQCFQCFQCFQCFQ
SQ   IVLGRFPNTERESFVUKHRNDAVPT
XX
SC   conceptually translated from EMBL/GenBank/3001 @K09205
XX
FT   3 57 BABS box.
FT   58 86 HEF2 domain.
FT   87 192 replaced by 87-192 in aHEF-2.
FT   101 186 serine-/threonine-rich region (20/46).
FT   200 296 absent in BSRFC4/BSRFTCO.
FT   312 312 threonine phosphorylation site.
FT   319 319 threonine phosphorylation site.
FT   355 355 serin phosphorylation site.
FT   420 446 glutamine-/proline-rich region (27/27).
  
```

class

## FACTOR TABLE continued

```

XX
SF alternative splicing leads to aHEF-2, ERFC4, ERFC9;
SF direct interaction with other myogenic factors [bHLH proteins MyoD,
SF myogenin, MRF4], but not with E2 proteins [7];
SF GRE mutant exhibits MCM1-binding specificity [5];
SF heterodimerization with HEF-2D [1];
XX
CP cardiac, smooth, skeletal muscle, less in placenta, brain, lung, kidney
CP [8]; P19 cells differentiated into neurons or endodermal cells,
CP undifferentiated P19 cells [4];
CN liver [8];
XX
FF activator [8];
FF involved in myogenesis [6] [6];
FF cooperates with myogenic bHLH factors through E-box for gene activation in
FF skeletal muscle cells [7];
FF NADS domain of HEF2A is both necessary and sufficient for binding to MyoD
FF [7];
FF synergistic activation of transcription in neurogenic lineages with NASH1
FF [4];
FF threonines 312 and 319 are phosphorylation sites for p38 [1] [2];
FF serine 355 is a phosphorylation site for p38 [2];
FF phosphorylation of HEF-2A in a HEF-2A/HEF-2D heterodimer enhances
FF HEF-2-dependent gene expression [1];
XX
IN T01537 MRF4; mouse, Mus musculus.
IN T00525 MyoD; human, Homo sapiens.
IN T00526 MyoD; mouse, Mus musculus.
IN T00528 myogenin; mouse, Mus musculus.
XX
MX M00006 V4HEF2_01.
MX M00031 V4HEF2_02.
MX M00032 V4HEF2_03.
MX M00033 V4HEF2_04.
XX
ES E03583 AS(HEF2_01; Quality: 6.
ES E03586 CHICK(CTNT_05; Quality: 6; cTNT, G0000057; chick, Gallus gallus.
ES E00197 EDV(BELF1_01; Quality: 6; BELF1, G000118; EDV, Epstein-Barr virus.
ES E00198 EDV(BELF1_02; Quality: 6; BELF1, G000118; EDV, Epstein-Barr virus.
ES E00199 EDV(BELF1_03; Quality: 6; BELF1, G000118; EDV, Epstein-Barr virus.
ES E09149 HS(ENK3_01; Quality: 6; ENK3, G001831; human, Homo sapiens.
ES E09469 HS(HYF4_01; Quality: 2; Hyf4, G001943; human, Homo sapiens.
ES E09235 HS(HYOG1_02; Quality: 2; Hy, G000348; human, Homo sapiens.
ES E09244 MURK(BCT_05; Quality: 6; mck, G000367; mouse, Mus musculus.

```

interacting factors

matrices

sites in genes

## MATRIX TABLE

```

AC M00006
XX
IN V4HEF2_01
XX
LT 19.10.1990 [created]: ewi.
LT 12.09.2002 [updated]: vma.
CC Copyright (C), EMBASE GmbH.
XX
EA HEF-2
XX
EE myogenic enhancer factor 2
XX
EF T00505 HEF-2; Species: mouse, Mus musculus.
EF T01004 HEF-2; Species: rat, Rattus norvegicus.
EF T01005 HEF-2; Species: human, Homo sapiens.
XX
ES
ES2
ES1 0 3 0 1 1 C
ES2 0 0 0 1 4 T
ES3 0 0 0 0 1 C
ES4 0 0 0 0 5 T
ES5 5 0 0 0 0 A
ES6 5 0 0 0 0 A
ES7 5 0 0 0 0 A
ES8 5 0 0 0 0 A
ES9 5 0 0 0 0 A
ES10 0 0 0 0 5 T
ES11 5 0 0 0 0 A
ES12 5 0 0 0 0 A
ES13 0 5 0 0 0 C
ES14 0 2 0 0 3 Y
ES15 0 4 0 0 1 C
ES16 0 2 0 0 3 Y
XX
EA 5 elements from 2 different genes [mouse/human/rat boik; rat/chicken sio]
XX
CC compiled sequences
XX
FE [1]
EE MEDLINE: 90092919.
EA Gosssett L. A., Melvin D. J., Steinberg E. A., Olsen E. W.
ET A new myocyte-specific enhancer-binding factor that recognizes a conserved
ET element associated with multiple muscle-specific genes

```

frequency matrix and  
consensus sequence

sequences used for matrix

TRANSCompel, Release 6.0 - public - 2002-01-01

[Documentation](#)

---

000120      **TRANSCompel = composite regulatory elements**

---

HLH4MADS\_001

---

[0001055](#)  
MRF4; muscle regulatory factor 4 gene.  
rat, Rattus norvegicus.

---

CTATATATAAAGCTG...CATCTG.  
-26 to 27.  
ENBL: [M24685](#); RRMRF4A; 498.  
synergism  
tissue-restricted/tissue-restricted; muscle cells

---

ev00274: [1].  
Site-directed mutagenesis and study of promoter activity.  
Functional synergism between sites.

---

in00157: -26 to -15.  
[MRF-4](#); human, Homo sapiens.  
MADS.

---

Muscle-specific factor.  
TRANSFAC: [T01005](#)  
in00158: 22 to 27.  
[MADS](#); rat, Rattus norvegicus.  
bHLH.

---

Muscle-specific factor.  
TRANSFAC: [T01514](#)

---

ev00275: [1].  
Transient co-transfections.  
Functional synergism between factors.

---

000120: -26 to -15

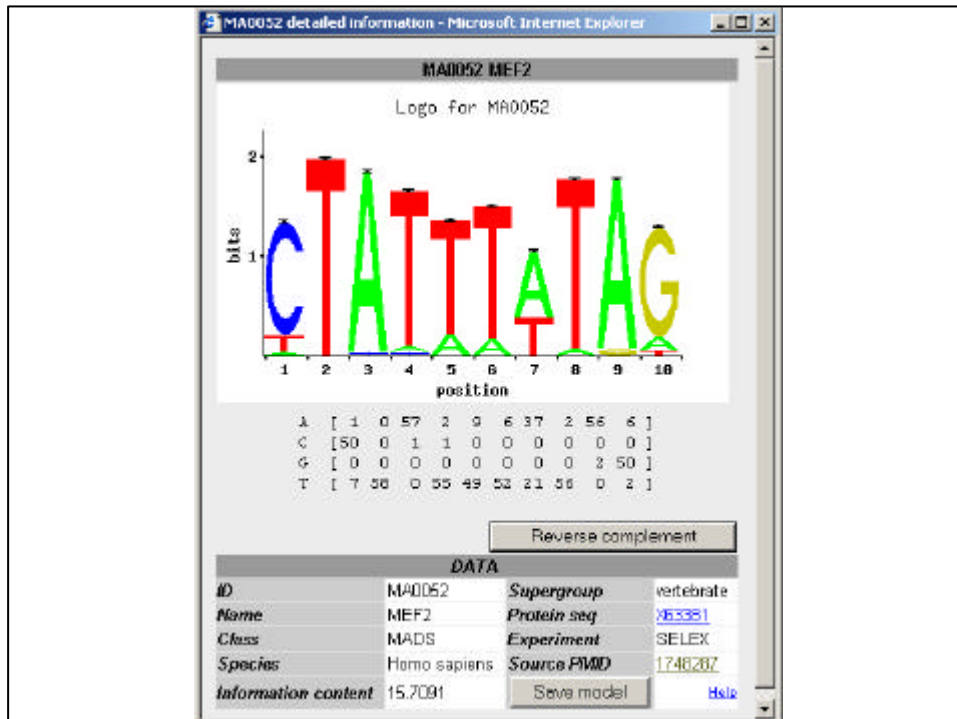
## The "Genomatix family"

- matrix database (535 matrices), promoter sequence database, and an extensive package of promoter analysis programs
- programs:
  - MatInspector: matrix search
  - ModelInspector: model search
  - fastM: model creation
  - FrameWorker: find a common framework
  - Gene2Promoter: promoter sequence retrieval
  - EIDorado: functional genomic elements, comparative genomics
  - SequenceShaper: TFBS design
  - BiblioSphere: literature mining combined with TFBS analysis
- limited free access for academics, requires registration:  
<http://www.genomatix.de/>

Matrix Name:	V\$E47_01
Description:	MyoD/E47 and MyoD/E12 dimers
Family:	V\$MYOD (MYOblast Determining factor)
Extra information / References:	[1] <b>HEADLINE: E12/E47-</b> Sun X.-H., Baltimore D. An inhibitory domain of E12 transcription factor prevents DNA binding in E12 homodimers but not in E12 heterodimers Cell 64:459-470 (1991).
Statistical basis:	11 selected strong binding sites for E47, E47-MyoD, E12+MyoD and (weak) for E12
Based on:	TRANSFAC identifier: V\$E47_01
Random expectation (p-value):	0.11 matches per 1000 bps
Optimized matrix threshold:	0.02
Length:	15 bps
Profile:	

## JASPAR

- open-access matrix database (111 matrices) and search tools
- programs:
  - matrix search with a sequence
  - compare own matrix to database matrices
- [http://jaspar.cgb.ki.se/cgi-bin/jaspar\\_db.pl](http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl)



## Transcription Regulatory Regions Database (TRRD)

- database containing information about regulatory regions using hierarchical approach
  - regulatory region
  - regulatory unit
  - transcription factor binding site
- experimental evidence, effects on transcription
- limited demo version available at <http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/>

**GENE TABLE**

[TRRDGENES4-A00042](#)

**Gene ID**  
Hs:MYOGL (TRRD\_Viewer)

**Links:**  
[Binding sites](#)  
[Transcription factors](#)  
[Gene expression regulation](#)  
[Bibliography](#)

**Updated**  
27/05/03

**GeneAC**  
A00042

**TransfacLink**  
A00042

**Annotators**  
Akanki E., Kral O., O.L.P.

**Species**  
HUMAN, Homo sapiens

**GeneName\_Brief**  
Myo

**GeneName\_Full**  
myoglobin

**DNABankLink**  
EMBL: [HS0001](#); [X00371](#)

**DataBankLink**  
 PIR: [I51991](#); MYGU  
 CleanEx: [HGNC:6915](#); NB  
 NIN: [160000](#)  
 SOURCE: [MR](#)  
 GeneLynx: [MR](#)  
 EPD: [MS\\_MYG](#); [EPI1091](#)  
 Genes: [HGNC:6915](#); NB  
 Ensembl: [P02144](#)  
 CPGISLE: [R5NG1-3](#)  
 GeneCards: [NB](#)  
 SWISS-PROT; MYO\_HUMAN; [P02144](#) (Expaty server)

**Entry Page - Microsoft Internet Explorer**  
 Address: <http://rsb.bionet.nsc.ru/rsbbr/cgi-bin/wgetz>  
 SWISS-PROT; MYO\_HUMAN; [P02144](#) (Expaty server)  
 GDB: [GDB:119378](#); NB  
 NOVERGEN: [P02144](#)  
[EPD Class](#)  
6.1.3.10.1.  
[KeyWords](#)  
late gene, transport protein, globin  
[RegRegion](#)  
5' region  
[RegUnitAC](#)  
REGULATORY UNIT: [P00432](#)  
[RegUnit](#)  
promoter; 87; [81248](#), [8119](#), [8135](#), [81250](#)  
[Alignment](#)  
human, mouse, and seal myoglobin genes: -250 to -10;  
 polypyrimidine stretches and TATA box [[Delvin](#), B.H. et al., 1989]  
[Comments](#)  
the deletion analysis suggests the possibility of a b  
 alignment within -957 to -374. [[Delvin](#), B.H. et al., 1989]  
 //

**SITE TABLE**

[TRRDSITES4-S134](#)

**SiteAC**  
S134

**GeneID**  
Gene: Hs:MYOGL

**RegUnitAC**  
REGULATORY UNIT: [P00432](#)

**SiteName**  
CCAC box;  
NP [undefined]

**SiteNameEPBox**  
CBF 40 binding site

**DatabaseReference**  
SAMPLES: [CP-1](#);

**DNA\_BankLink**  
[X00371](#)

**ExperimentCodes**  
3.1, 3.3, 6.3 [[Bassel](#) -Duby R. et al., 1992]  
7.3 [[Delvin](#), B.H. et al., 1989]

**TextComments**  
Substitution A at 215 to G, C at 216 to G, C at 217 to T,  
C at 218 to A lead to the reduction of promoter function more than  
20-fold. CCAC mutations disrupted promoter function in vivo and  
reduced binding CBF 40 in vitro [[Bassel](#) -Duby R. et al., 1992]  
 //

## Other TF databases...

- Arabidopsis
  - <http://arabidopsis.med.ohio-state.edu/AtTFDB/index.jsp>
- Plants
  - <http://www.dna.affrc.go.jp/htdocs/PLACE/>
- *S. cerevisiae*
  - <http://cgsigma.cshl.org/jian/>
- *E. coli*
  - [http://bayesweb.wadsworth.org/binding\\_sites/index.html](http://bayesweb.wadsworth.org/binding_sites/index.html)

## Programs for matrix scans

## Match

- uses Transfac matrices
- thresholds
  - minimize false negatives (minFN)
  - minimize false positives (minFP)
  - minimize the sum of both error rates
  - matrix similarity (takes information content into account)
  - core similarity (5 most conserved consecutive positions)
- profiles (matrix subset with defined cut-offs) for muscle, liver and immune cells, and cell cycle. Also "Best selection".
- possibility to create own matrices and profiles with thresholds
- the version using Transfac 6.0 public matrices is free for academics, requires registration:  
<http://www.gene-regulation.com/pub/programs.html#match>

## Match thresholds

- minimize false negatives (minFN)
  - FN10 = recognizes at least 90% of oligonucleotides generated from the matrix (= misses 10% of hits)
- minimize false positives (minFP)
  - no matches to non-regulatory sequences (exon2 and exon3)
  - strict, useful for looking for most promising sites in long genomic sequences
- minimize the sum of both errors (minSum)
  - set 100% false positives = number of matches from exon sequences using minFN. Raise the threshold and sum the new percentage of false negatives and false positives each time. Finally, pick the threshold that gives the minimum sum.

your login name: ekorpela

Select a previous search result: default out and VIEW |t DELETE |t

Select a previously stored sequence: myoglobin seq and DELETE |t

[A Match™ version with additional functionalities is included in TRANSFAC Professional®](#) [Get help](#) [Go to Match Profile](#)

## MATCH™ public version 1.0

Matrix Search for Transcription Factor Binding Sites

**BIOBASE**  
Biological Databases / Biologische Datenbanken GmbH

Please enter a name for your search: myoglobin

Sequence Selection

Select one of your stored sequences: myoglobin seq

OR take an **example**

OR take a new sequence and enter a name for it: myo-499 seq

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

```

ACAGTGGCCAACTCTCTCCCTCTCTTCACAGGACAAAC
CAAGCCAGCCCGTGTGGCT
CAAGCTGTCTCTCTCTCTCCAGCAATGGCACTTCCCTAAAAT
AGCTTCCCTATGTGAAGGCTA
GAGAAAGGAAAGATTAGACCTCTCTCTCTATGAGAGAGAG
AAAGTGAAGGAGGGGAGGGG
AGGGTACAGCCAGCCAGCTTGAAGGCTCTTCTTAAAGATC
CCAGAAAGGTATAAAGAGCC
CTTGGGACCAAGCAAGCTCA

```

Allowed formats are **RAW**, **FASTA**, **TRANSFAC**, **EMBL**, **GenBank**, **Cl**

Matrix or Profile Selection

Matrices

Group of matrices: all  
no falsePos  
lung

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: muscle\_specific\_pf

your profiles: e1a\_misSUM

Submit the form    Reset the form

Search for sites by WeightMatrix library: matrixTFP60.lib

Sequence file: myo-499 seq seq

Profile: muscle\_specific\_pf

Scanning sequence ID: EP11091 (+) Hs MB, range -499 to 100;  
View a [graphic](#) of the following search results

matrix identifier	position (strand)	core match	matrix sequence (always the (+)-strand is shown)	factor name
<a href="#">V18P1_Q6</a>	55 (+)	0.919	0.771 aaggGGCAaagg	<a href="#">Sp1</a>
<a href="#">V1E47_Q2</a>	111 (-)	1.000	0.920 cggggcACCTGgtgc	<a href="#">E47</a>
<a href="#">V1HY00_Q1</a>	113 (+)	0.778	0.830 gggCACCTggtg	<a href="#">MyoB</a>
<a href="#">V1RT00_Q1</a>	113 (-)	1.000	0.912 gggcACCTGgtg	<a href="#">MyoB</a>
<a href="#">V1URF_Q6</a>	114 (+)	0.886	0.887 gGCACctggt	<a href="#">USF</a>
<a href="#">V1URF_Q6</a>	117 (-)	0.913	0.881 aactGGTggc	<a href="#">USF</a>
<a href="#">V1SP1_Q6</a>	152 (+)	0.927	0.867 ctTgGGAGggtg	<a href="#">Sp1</a>
<a href="#">V1SP1_Q6</a>	251 (-)	0.927	0.882 aactCCTCcoctt	<a href="#">Sp1</a>
<a href="#">V1SP1_Q6</a>	276 (-)	0.915	0.821 ccaacCAGCCooc	<a href="#">Sp1</a>
<a href="#">V1SP1_Q6</a>	281 (-)	0.915	0.925 ccaacCAGCCooc	<a href="#">Sp1</a>
<a href="#">V1HY00_Q1</a>	286 (-)	0.751	0.825 ccaacCCTGtgc	<a href="#">MyoB</a>
<a href="#">V1URF_Q6</a>	289 (-)	0.931	0.899 ccccTGTggc	<a href="#">USF</a>
<a href="#">V1E47_Q2</a>	321 (-)	1.000	0.913 aatTggcACCTGcctta	<a href="#">E47</a>
<a href="#">V1HY00_Q1</a>	323 (-)	1.000	0.931 TggcACCTCooc	<a href="#">MyoB</a>
<a href="#">V1URF_Q6</a>	324 (+)	0.886	0.904 gGCACctgoc	<a href="#">USF</a>
<a href="#">V1SP1_Q6</a>	325 (-)	0.819	0.782 gcaacCTGccctaa	<a href="#">Sp1</a>
<a href="#">V1TATA_C</a>	331 (+)	1.000	0.881 gcccTAAAAt	<a href="#">TATA</a>
<a href="#">V1HY00_Q1</a>	345 (+)	0.905	0.813 tccCATGTgagg	<a href="#">MyoB</a>
<a href="#">V1URF_Q6</a>	346 (-)	0.945	0.905 ccccTGTggg	<a href="#">USF</a>
<a href="#">V1SP1_Q6</a>	377 (-)	0.927	0.877 agaacCCTCctgg	<a href="#">Sp1</a>
<a href="#">V1SP1_Q6</a>	405 (+)	0.927	0.785 TgaaGGAGgctg	<a href="#">Sp1</a>
<a href="#">V1SP1_Q6</a>	409 (+)	0.819	0.826 gaggGGCAgagg	<a href="#">Sp1</a>
<a href="#">V1SP1_Q6</a>	415 (+)	0.927	0.918 caggGGAGgagg	<a href="#">Sp1</a>
<a href="#">V1TATA_C</a>	467 (+)	1.000	0.947 ggttaTAAAAt	<a href="#">TATA</a>

Total sequences length=500  
Total number of sites found=24

your login name: ekorpets

Select one of your predefined profiles: **ejjo** and VIEW DELETE

Select one of our predefined profiles: **immune\_cell\_specific.pr** and VIEW

Update Page Goto Match Goto Matrix Generation Get help

# MATCH™ PROFILER

Biological Databases / Biologische Datenbanken GmbH

Select matrices to be included in a profile:

Select matrices from the list:

- ABF1 (M00197) high quality
- ABF1 (M00015) high quality
- ACR1 (M00046) low quality
- AG (M00151) high quality
- AGL3 (M00392) high quality
- AGL3 (M00393) high quality
- AML-1a (M00271) low quality
- AP-1 (M00172) low quality
- AP-1 (M00173) high quality
- AP-1 (M00174) low quality

Search for matrices:

by factor name  
by accession number

Please enter a factor name (accession number) or a list of factor names (accession numbers) separated by comma:

Sort list by accession no. Sort list by factor name

mailto:info@biobase.de

Cutoff selection. Mark matrices and cut-offs you want to save to your profile:

Choose the **cut-off** for each matrix by hand or mark the following cut-off for all matrices:

minFP  Mark

or use the following cut-offs for all matrices:

Set  
(core sim. cut-off / matrix sim. cut-off)

matrices	Matrix similarity cut-offs (core similarity will be always set to 0.75)								Current core similarity matrix similarity
	minFP	FP	FP	FP	FP	minFP	minFP		
<input checked="" type="checkbox"/> MEF-2 (M0022) high quality	<input type="checkbox"/> 0.84 (0.124)	<input type="checkbox"/> 0.9 (0.025)	<input type="checkbox"/> 0.93 (0.005)	<input type="checkbox"/> 0.98 (0.002)	<input type="checkbox"/> 1.0 (0.000)	<input type="checkbox"/> 0.99	<input type="checkbox"/> 0.9	<input type="text"/>	
<input checked="" type="checkbox"/> MEF-2 (M0022) high quality	<input type="checkbox"/> 0.88 (0.047)	<input type="checkbox"/> 0.93 (0.007)	<input type="checkbox"/> 0.98 (0.001)	<input type="checkbox"/> 0.98 (0.000)	<input type="checkbox"/> 1.0 (0.000)	<input type="checkbox"/> 0.99	<input type="checkbox"/> 0.93	<input type="text"/>	
<input checked="" type="checkbox"/> MEF-2 (M0022) high quality	<input type="checkbox"/> 0.88 (0.001)	<input type="checkbox"/> 0.92 (0.000)	<input type="checkbox"/> 0.95 (0.000)	<input type="checkbox"/> 0.97 (0.000)	<input type="checkbox"/> 1.0 (0.000)	<input type="checkbox"/> 0.92	<input type="checkbox"/> 0.92	<input type="text"/>	
<input checked="" type="checkbox"/> MEF-2 (M0022) high quality	<input type="checkbox"/> 0.88 (0.003)	<input type="checkbox"/> 0.92 (0.001)	<input type="checkbox"/> 0.96 (0.000)	<input type="checkbox"/> 0.98 (0.000)	<input type="checkbox"/> 1.0 (0.000)	<input type="checkbox"/> 0.95	<input type="checkbox"/> 0.95	<input type="text"/>	
<input checked="" type="checkbox"/> SRF (M00152) high quality	<input type="checkbox"/> 0.9 (0.000)	<input type="checkbox"/> 0.94 (0.000)	<input type="checkbox"/> 0.96 (0.000)	<input type="checkbox"/> 0.98 (0.000)	<input type="checkbox"/> 1.0 (0.000)	<input type="checkbox"/> 0.9	<input type="checkbox"/> 0.9	<input type="text"/>	
<input checked="" type="checkbox"/> SRF (M00156) high quality	<input type="checkbox"/> 0.86 (0.252)	<input type="checkbox"/> 0.92 (0.035)	<input type="checkbox"/> 0.96 (0.005)	<input type="checkbox"/> 0.99 (0.000)	<input type="checkbox"/> 1.0 (0.000)	<input type="checkbox"/> 0.95	<input type="checkbox"/> 0.92	<input type="text"/>	
<input checked="" type="checkbox"/> SRF (M00215) high quality	<input type="checkbox"/> 0.85 (0.131)	<input type="checkbox"/> 0.9 (0.015)	<input type="checkbox"/> 0.93 (0.003)	<input type="checkbox"/> 0.96 (0.000)	<input type="checkbox"/> 1.0 (0.000)	<input type="checkbox"/> 0.95	<input type="checkbox"/> 0.9	<input type="text"/>	
<input checked="" type="checkbox"/> Sp1 (M00174) low quality	<input type="checkbox"/> 0.77	<input type="checkbox"/> 0.86	<input type="checkbox"/> 0.91	<input type="checkbox"/> 0.95	<input type="checkbox"/> 1.0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="text"/>	

# MatInspector professional

- Genomatix tool
- thresholds
  - matrix similarity (takes information content into account)
  - core similarity (4 most conserved consecutive positions)
  - optimized matrix threshold (max 3 hits per 10 000 bp of non-regulatory test sequence)
- matrices grouped into families
  - similar TFBSs with a similar biological function
  - only the best match within a family is listed
- tissue filter (in commercial version only)
- matches relative to TSS, interactive graphical output

Matrix parameters	
Matrix families	<input checked="" type="radio"/> - matches to matrix families ( <a href="#">see matrix families</a> ) <input type="radio"/> - matches to individual matrices ( <a href="#">see matrices</a> )
Matrix group ( <a href="#">View transcription factor &lt;-&gt; matrix assignment</a> )	<input type="checkbox"/> Fungi <input type="checkbox"/> Other Functional Elements <input type="checkbox"/> Insects <input type="checkbox"/> Plants <input type="checkbox"/> Miscellaneous <input checked="" type="checkbox"/> Vertebrates <input checked="" type="radio"/> - use all matrices from selected groups <input type="radio"/> - continue with subset definition from selected groups <input type="radio"/> - use previously defined matrix subsets
Matrix filters (only available for vertebrates)	select matrices associated with the following tissues ( <a href="#">show all tissue associations</a> ): Adipose Tissue Adrenal Glands Antibody-Producing Cells Blastomeres Blood Cells
Core similarity	0.75
Matrix similarity	Optimized
Output parameters	
Statistics	<input checked="" type="radio"/> - show results only (no statistics of match numbers) <input type="radio"/> - show only statistics of match numbers <input type="radio"/> - show results and statistics of match numbers
Matches sorted by	<input type="radio"/> - matrix names <input type="radio"/> - quality <input checked="" type="radio"/> - matches on the sequence
Offset for match positions	0 bp (not available for database search)
Graphics	<input checked="" type="radio"/> - show graphics <input type="radio"/> - no graphics (recommended for database searches)
Further output	<input checked="" type="checkbox"/> show further information column

Genomatix: Result - Microsoft Internet Explorer

Address: [http://www.genomatix.de/cgi-bin/retrosector\\_prof/ret\\_1m.pl](http://www.genomatix.de/cgi-bin/retrosector_prof/ret_1m.pl)

Inspecting sequence EP11091 (1 - 600):

[EP11091 (+) Hs MB; range -499 to 100]

Family/matrix	Further Information	Out.	Position		Str.	Core sim.	Matrix sim.	Sequence
			From - to	anchor				
<a href="#">VIMINI/MUSCLE_INTL_02</a>	Muscle Initiator Sequence	0.86	29 - 47	38	(-)	0.880	0.871	gacccT <b>CAC</b> ccagctaga
<a href="#">VIMICD/MC2COM_01</a>	complex of Lmo2 bound to Tal-1, E2A proteins, and GATA-1, half-site 1	0.98	111 - 125	118	(-)	1.000	0.981	ccac <b>CAGG</b> gcccag
<a href="#">VIMICD/MYOD_02</a>	Myoblast determining factor	0.98	322 - 336	329	(+)	1.000	0.989	atgg <b>CAC</b> ctgcocta
<a href="#">VIMEF2/MEF2_02</a>	MEF2	0.96	461 - 483	472	(+)	1.000	0.978	ccagaaggt <b>TAA</b> acgcccctt
<a href="#">VIMEF2/MEF2_01</a>	myocyte enhancer factor	0.90	463 - 495	474	(+)	1.000	0.928	apaaaggt <b>TAA</b> acgccccttg
<a href="#">VIMICD/MYES_01</a>	Myf5 myogenic bHLH protein	0.90	501 - 515	508	(-)	1.000	0.908	ccaa <b>CAGC</b> tgggctt

6 matches found.

Statistics:

Matrix	E-value	No. of matches	No. of sequences
<a href="#">VIMICD/MC2COM_01</a>	re: 1.30	1 matches	1 seq.
<a href="#">VIMEF2_02</a>	re: 4.53	1 matches	1 seq.
<a href="#">VIMEF2_01</a>	re: 0.66	1 matches	1 seq.
<a href="#">VIMMUSCLE_INTL_02</a>	re: 0.89	1 matches	1 seq.
<a href="#">VIMES_01</a>	re: 0.11	1 matches	1 seq.
<a href="#">VIMICD_02</a>	re: 0.06	1 matches	1 seq.

Family statistics:

Done

Internet

red: high information content  
capital: core sequence

matches per 1000 bp

## JASPAR

- open-access matrix database (111 matrices) and search tools
- programs:
  - matrix search with a sequence
  - compare self-made matrix to database matrices
- [http://jaspar.cgb.ki.se/cgi-bin/jaspar\\_db.pl](http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl)

The high-quality transcription factor binding profile database

BROWSE profiles by:      [help](#)

SEARCH by:       [help](#)

combine searches with:    [help](#)

UPLOAD set of profiles:    [help](#)  
generally restrict to Do

COMPARE custom profile to database profile:   [help](#)

[JASPAR INFORMATION](#) [DATA DOWNLOAD](#) [LINKS](#)

SEARCH


ID	NAME	SPECIES	CLASS	LOGO	
MA0001	AGL3	Arabidopsis thaliana	MADS		<input type="button" value="VIEW"/>
MA0002	AML-1	Homo sapiens	RUNT		<input type="button" value="VIEW"/>
MA0003	AP2alpha	Homo sapiens	AP2		<input type="button" value="VIEW"/>
MA0004	ARNT	Mus musculus	bHLH		<input type="button" value="VIEW"/>
MA0110	ATH5	Arabidopsis thaliana	HOMEOD		<input type="button" value="VIEW"/>
MA0005	Agamous	Arabidopsis thaliana	MADS		<input type="button" value="VIEW"/>
MA0006	Atc-ARNT	Mus musculus	bHLH		<input type="button" value="VIEW"/>

Analyze this (fasta-formatted) sequence with selected profile score threshold  %

```
>EP11091 (+) Bc NB1 range -499 to -1
CAAAAACATCCGGTTCCTTCTCTCACTTCGGCTGGGTGAGGGGT
CAGAAAGTGCCTGAGAGGTTGCGAATGGCCAGGACTGTCTGGGG
GGTGGCCAGCTTAGAAACATGACAGGTCCTCTGGGAGGGCTGAG
GGTTTCAGGCTGCTGGGCTGGCTTCCTGGTGGCTTCTGTGG
ACAGTGGCCAACTGCTCCCTCTCTTCCACAGGCACACACCC
GAGCTGTCTGCTGGCCACAAAGGCCTGCTGCTAAATAGCTT
GAGAAAGGAAAAGATTAGACCCTCCCTGGATGAGAGAGAGAAAT
AGGGGGACAGCGACCATGGAGATCTTTGTCAAGCATCCCAGAA
CTTGGGACAGGCAGGCTCA
```

**Site search mode**

Putative transcription factor binding sites found along *EP11091*



**Table view**

Transcription factor		EP11091			
Sequence	From	To	Score	Strand	
MFE2	333	342	8.521	-	

**Sequence view**

```

EP11091  CAAAACCATCCGGTTCCTTTCTGTCACTTC79GCTG66TGAGGGGTCCTCTG9CAAAG000
          |      |      |      |      |      |      |      |      |      |
          10     20     30     40     50     60

EP11091  CAGAGGTGCGCTGAGAGGTTTGGAAATGGCCAGGACTGTCTCTGGGGCCAGCCGGGGCACCT
          |      |      |      |      |      |      |      |      |      |
          70     80     90     100    110    120

EP11091  GGTGGCCAAAGCTTAGAAAATGACAGGTCTCTTTGGGAGGGGCTGACCCGCAAGGAGCGCTTG
          |      |      |      |      |      |      |      |      |      |
          130    140    150    160    170    180
  
```

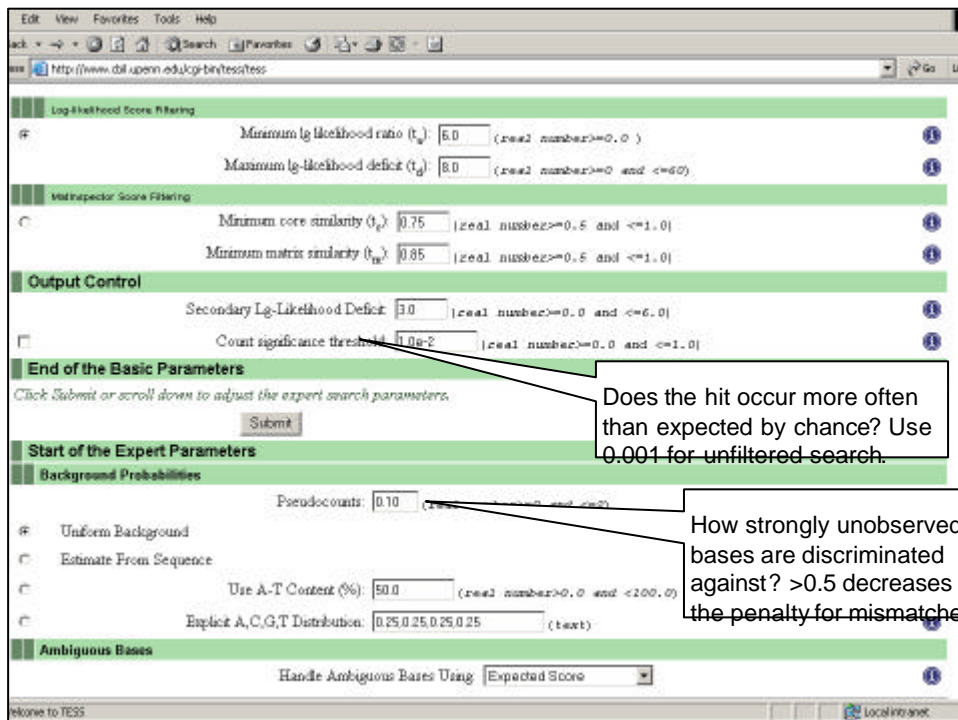
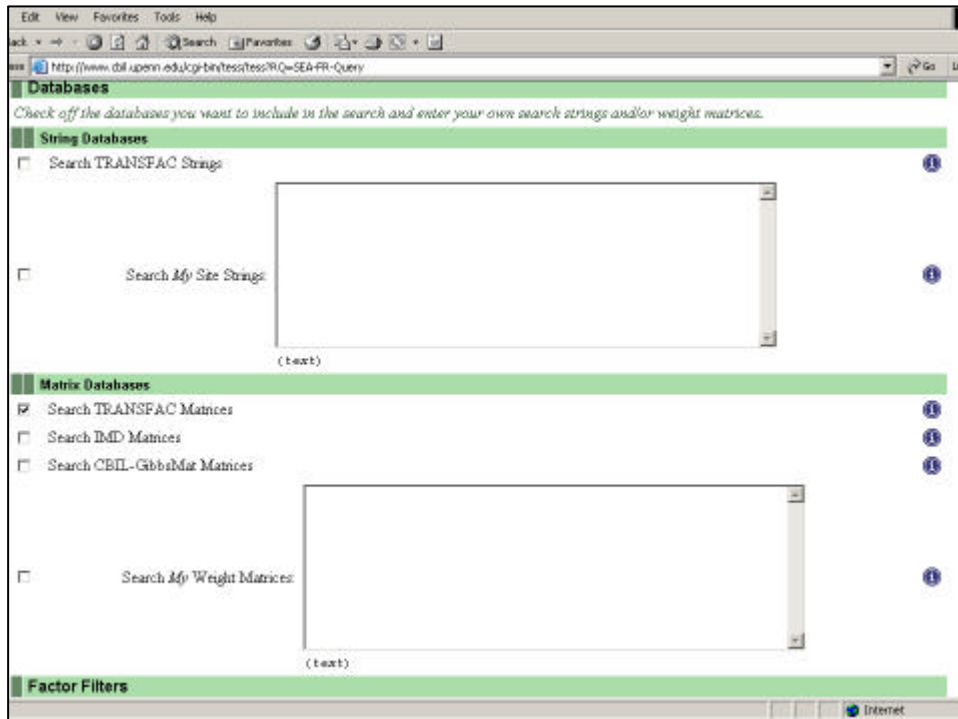
## TESS

- uses Transfac 4.0 public
- thresholds
  - matrix similarity and core similarity
  - minimum log likelihood ratio and deficit
  - significance threshold
- several filters to select matrices
- possibility to alter pseudocounts and background
- string search included
- <http://www.cbil.upenn.edu/tess/>

# TESS filters

- cell positive specificity (e.g. liver)
- organism classification (e.g. mammalia)
- organism species (e.g. Mus musculus)
- factor class (e.g. bHLH)
- factor name and synonyms
- interacting factors
- words in reference titles (e.g. homeobox)

The screenshot shows the TESS - Combined Search Page in a web browser. The page title is "TESS - Combined Search Page". The browser address bar shows "http://www.dbl.upenn.edu/cgi-bin/tes/tes?RC=SEA-FR-Query". The page has a navigation menu with links: Home, Site Searches, Query Transfac, Query Matrices, Other Stuff, About, Strings, Filtered Strings, Combined, and Royal Search. Below the navigation menu, there is a disclaimer: "Please send questions and comments to TomMack@dbf.upenn.edu". The main heading is "What potential transcription factor binding sites are there in my sequence?". Under the heading, there is an "Input" section with the instruction "Enter the minimal information needed to submit a job to TESS". The "Title" field contains "human myoglobin". The "DNA Sequence(s)" field contains a long DNA sequence: "EP11091 (+) Hs MB: range -499 to 100 AAAACCATCCGGTTCTTCTGTCACTTCTGGCTGGGTTGAGGGGTCCTGGCAGAGGGG AGAAGTCCGTGAGAGGTTTCCGAAATGGCCAGGACTGTCTTGGGGCCAGCCGGGGCACTT GTGGCCAACTTAGAAACATGACAGGTCCTTGGGAGGGGCTGACCCGCAAGGAGGCTTGT GTTTCAGGCTCTGGGCTCCGGCTTCTGTGGTGGCCCTTCTGTGTGGCTAGAGAGTCCAG CAGTGGCCAACTCCCTCCCTTCTTTCACAGGACAGACCCAGCCCACTCCCTGTGGCTT AGCTGTCTGGCTCCGCAATGGCACTGGCCCTAAAGTACTTCCCATGTGAGGGGCTA AGAAAGGAAAGATTAGACCTCCCTGGATGAGAGAGAGAAAGTTGAAAGGAGGGCAGGGG GGGGACAGCCAGCCATTGAGGATCTTGTCAAGCATCCAGAGAGGTTATAAAGAGCC TTGGGACAGGACGCTCAAGCCCAAGCTGTGGGGCCAGGACCCAGTGAAGCCATA". The "Length of time to store results of the job" is set to "week". The "Your email address" field contains "erja.korpelainen@cccfi". Below the input fields, there is an "End of Minimal Parameters" section with the instruction "You can click 'Submit' to submit the job or scroll down to change the basic search parameters." and a "Submit" button. The browser status bar shows "Done" and "Internet". The taskbar at the bottom shows "Start", "Inbox - Outlook Expr...", "promocitorianalys", "Microsoft PowerPoi...", and "TESS - Combined ...".



Back → → → Search Favorites → → →

http://www.dbl.upenn.edu/cgi-bin/teess

## TESS Job W0284007253 : Tabulated Results

[Home](#) | [Site Searches](#) | [Query Transfers](#) | [Query Matrices](#) | [Other Stuff](#)  
[About](#) | [Strings](#) | [Filtered Strings](#) | [Combined](#) | [Reset Search](#)

Please send queries and comments to [Teess@dbi.upenn.edu](mailto:Teess@dbi.upenn.edu)  
 The TRANSFAC database is for non-commercial use. For commercial use the TRANSFAC database and programs have to be licensed. Please read the [DISCLAIMER](#)  
 TESS is now using TRANSFAC v4.0.  
 The TESS site should be working, though there may be lingering problems. Please report any remaining problems you may encounter.

**Result Navigation**

Small Results	Tabular Results	Associated Sequence	Legend
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Use these checkboxes and the 'Select' button to choose which columns to display:

Select  Factor  Model  Beg  Sns  Len  Sequence  L<sub>g</sub>  L<sub>g'</sub>  L<sub>d</sub>  L<sub>gv</sub>  S<sub>c</sub>  S<sub>m</sub>  S<sub>gv</sub>  P<sub>gv</sub>

Click on a column heading to sort results on that column.

[ 1 .. 2 ] of 2    262

Factor	Model	Beg	Len	Sequence	L <sub>g</sub>	L <sub>g'</sub>	L <sub>d</sub>	L <sub>gv</sub>	S <sub>c</sub>	S <sub>m</sub>	S <sub>gv</sub>	P <sub>gv</sub>
<a href="#">T00108</a> C/EBP	<a href="#">M00159</a> (mod_c)	262	13	TTCTTTCCACACG	8.55	0.668	4.25	4.5e-01	0.95	0.94	2.3e-01	9.7e-01
<a href="#">T00104</a> C/EBPalpha												
<a href="#">T00105</a>												
<a href="#">T00107</a>												
<a href="#">T01388</a>												
<a href="#">T00459</a> C/EBPbeta	<a href="#">M00109</a> (mod_c)	262	14	TTCTTTCCACACGG	8.71	0.561	6.80	4.6e-01	0.99	0.91	6.4e-01	6.1e-01
<a href="#">T00581</a>												
<a href="#">T00017</a>												

[ 1 .. 2 ] of 2    262

blue: minor mismatch  
red: bigger mismatch

Done

Start    Inbo...    Adboconf...    promoot...    Microsoft...    TTESS - Co...    TTESS Job...    TTESS Job...

## Which one should I use?

- MatInspector
  - algorithm uses information vector
  - large, up-to-date matrix library
  - search with matrix families
  - tissue filter (in professional version)
- Match
  - algorithm uses information vector
  - large, up-to-date matrix library (public vs professional!)
  - ready made profiles (but with fixed thresholds)
- JASPAR
  - smaller but good quality matrix library
- TESS
  - allows filtering
  - allows fiddling with pseudocounts and background model

## How to weed out false positives?

- Phylogenetic footprinting
- Clusters/ modules
- Expression data

## 4. Clusters and phylogenetic footprinting

## Clusters/ modules/ frameworks

- Transcription factors don't act alone
- Biological knowledge to group TFs
- Spacing consideration (need data to establish rules)
- Genomatix: ModelInspector, fastM, FrameWorker
- MSCAN (<http://tfscan.cgb.ki.se/cgi-bin/MSCAN>)
- ClusterBuster (<http://zlab.bu.edu/cluster-buster/cbust.html>)

## Phylogenetic footprinting

- regulatory regions tend to be evolutionarily conserved
- align orthologous genes
  - arose from a common ancestral gene through speciation
  - likely to be involved in similar biological functions
- suitable evolutionary distance
  - 70 million years reveals most regulatory regions (eg. human/mouse)
  - coding sequences: 450 million years (human/fugu)
- results depend on the alignment
- phylogenetic footprinting eliminates ~90% of predictions (and ~10% of real hits)

## What kind of alignment?

- global alignment
  - assumes that similar regions in the sequence appear in the same order and orientation
  - on average, human and mouse have order and orientation preserved up to 8 Mb regions
  - Needleman-Wunsch, AVID
- local alignment
  - evidence that TFBSs are prone to reordering
  - less power in finding weakly conserved regions
  - Smith-Waterman, BLASTZ
- glocal alignment?

## ECR browser

- ECR = evolutionary conserved region
- visualizes all pairwise genomic alignments between human, mouse, rat, Fugu, Tetraodon, and zebrafish genomes
- extract sequences that correspond to any ECR, view sequence alignments, send results to further analysis (rVista)
- <http://www.dcode.org/>



# rVista

- alignment generated by BLASTZ
- identification of TFBS matches in individual sequences (using Match and Transfac 7.3 matrices)
- identification of aligned TFBS
- calculation of local conservation extending upstream and downstream from each orthologous TFBS
- calculates distance between all neighbouring TFBSs, possible to cluster individual or multiple TFs
- <http://www.dcode.org/>

The screenshot shows the rVista web interface. At the top, there is a navigation bar with "Back", "Search", and "Favorites" buttons. The address bar shows "http://rvista.dcode.org/". The main content area features a sequence alignment of two DNA sequences with the word "RVISTA" overlaid in large blue letters. Below the alignment, there is a paragraph explaining the tool's purpose: "Finding potential regulatory elements in noncoding regions of the human genome is a challenging problem. Analyzing novel sequences for the presence of known transcription factor binding sites or their weight matrices produces a huge number of false positive predictions that are randomly and uniformly distributed. rVista combines database searches with comparative sequence analysis, reducing the number of false positive predictions by 99.9% while maintaining a high zero false rate of the search."

There are 3 different ways to run rVista:

1. If you have 2 or more sequence files (of any size!) in FASTA format, then you can easily align them using fast and interactive alignment tool, **rPicture**, located at <http://rpicture.dcode.org/>. rPicture alignments could be automatically submitted for rVista processing.
2. If you know the location of your sequence in either human, mouse, rat, or fugu genome then you can fetch precalculated alignments for the rVista processing from the **ECR Browser** located at <http://ecrbrowser.dcode.org/>
3. If you precalculated alignments using **Advanced PipMaker** located at <http://bio.cse.msu.edu/pip-bin/pipmaker/advance/>, then you can submit them using the form below:  
1. Please note that "ENHANCING" PipMaker option is required to be selected in order for rVista to be able to process blastz alignments correctly.

2. How to run Advanced PipMaker for rVista

Alignment file (called "text")

Annotates file, base seq (optional):

Annotates file, second seq (optional):

Return to previously submitted request. ID:

Instructions: [dcode.version.2.0](#)  
Questions or comments: Gaby Loots ([gloots1@nl.gov](mailto:gloots1@nl.gov)) and Ivan Cycharenko ([cycharenko1@nl.gov](mailto:cycharenko1@nl.gov))  
Contact: [Lutz AG Cycharenko I](#), Pachter L, Dubchak I, Rubin EM. rVista for comparative sequence-based discovery of functional transcription factor binding sites, Genome Res. 2002 May;12(5):822-9 [PDF]



**RVISTA** | **VISUALIZATION & CLUSTERING**

**Picture:**  
Bases per layer: 0.5kb  
Picture width (in pixels): 800

**Clustering:**  
(Clustering: X sites over Y bases)  
(1 site per N bases = NO clustering)

Individual clustering  
1 100 SP4\_G6

Combinatorial clustering  
1 100

**Show:**  
 conserved  
 aligned  
 all

**SUBMIT**

**SELECT FACTORS**      **CLUSTERED SITES**

**REGULATORY VISTA**