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	in silico
• regulatory elements can be far	• large search space ☹
• binding sites have variability	• lot of random hits ☹
• chromatin structure is important	• chromatin structure is ignored ☹
• TFs work together	• module search improves specificity ☺

2. Retrieval of promoter sequences (Martti)

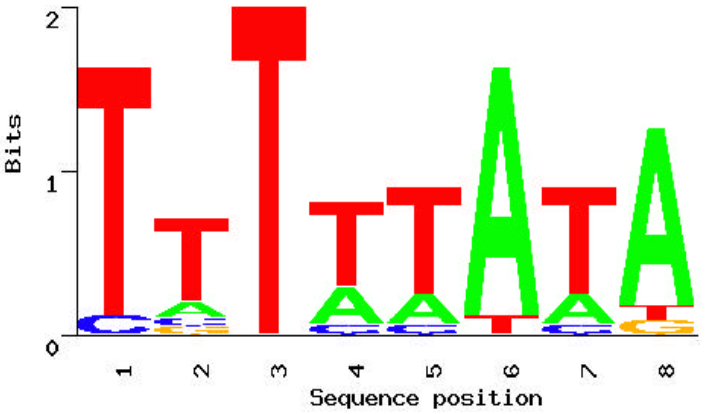
3. Detection of binding sites for known transcription factors

Presentation of TF binding sites

TTTTTATA
TTTAAATA
TTTTTAAA
CTTTTATA
TTTATATA
TTTAAAAA
TATTTATA
TTTTTACA
TTTTAAAA
TATATATA
TTTTTTTT
TCTTTATA
TTTCCATG
TGTTTATA

TTTTTATA

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



- 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

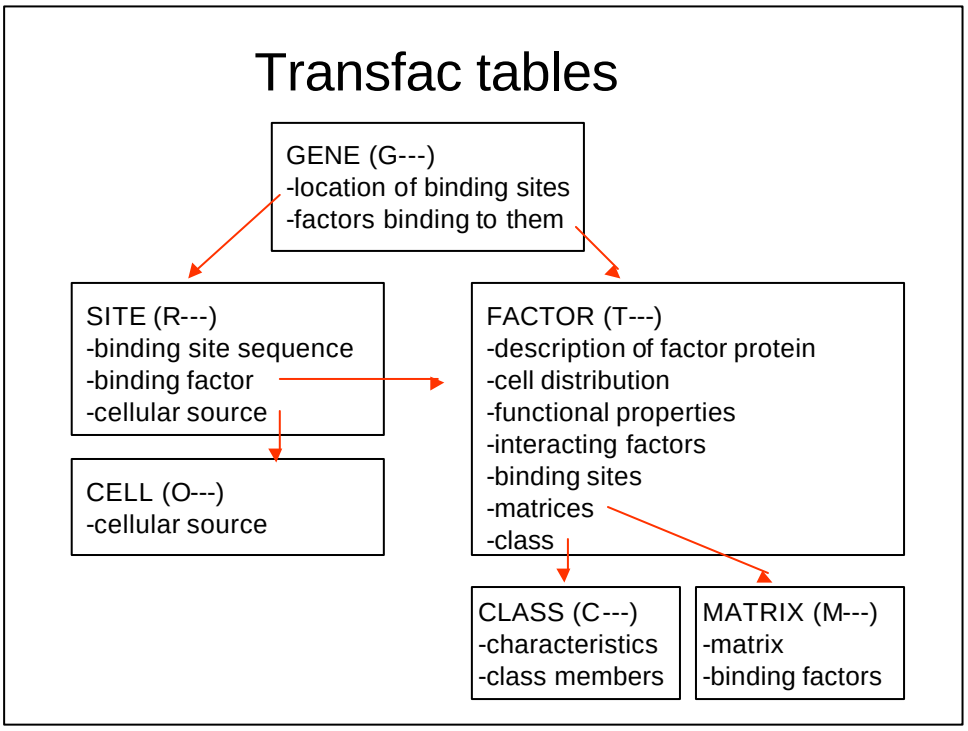
- because TFBSs are short and degenerate, predicted sites occur frequently at random (eg. MyoD sites 1 per 600 bp)
- sensitivity is good (real sites are not missed) but specificity is awful (a lot of false positives)
- nearly 100% of predicted sites have no function *in vivo,* because binding is regulated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

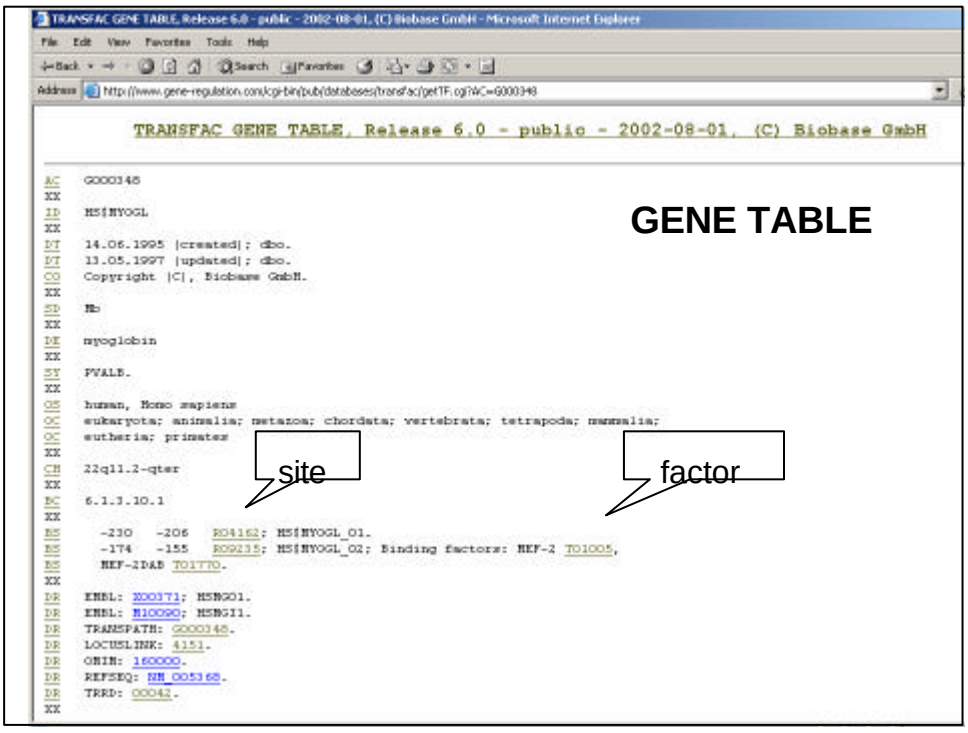
Transcription factor databases

The ”Transfac family”

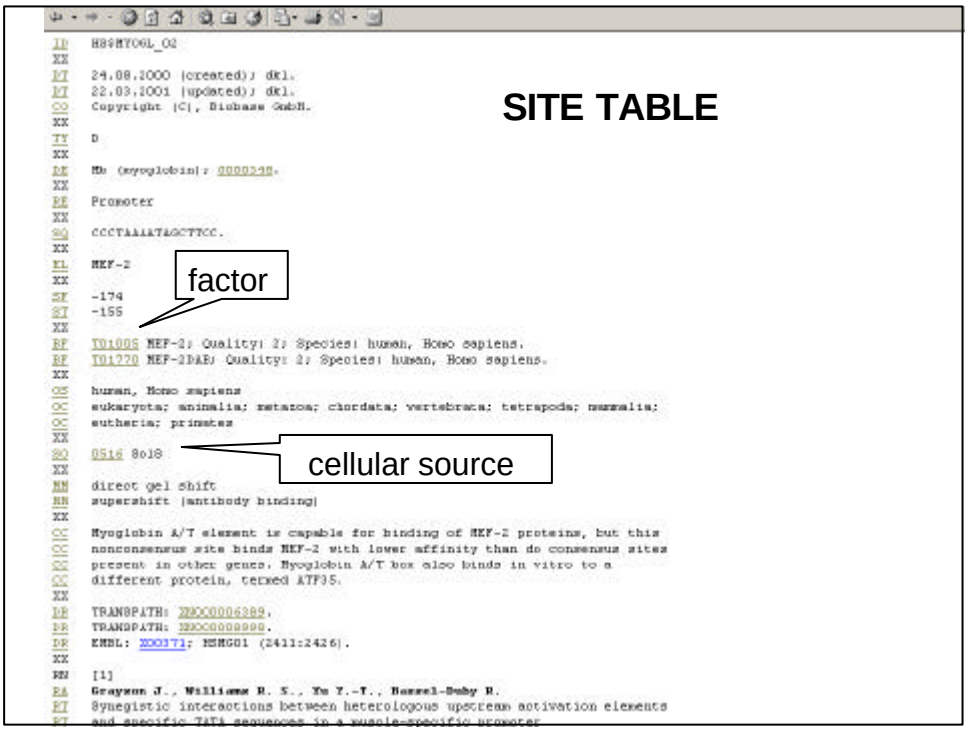
- consists of several databases and programs
- TRANSFAC
  - transcription factors, known binding sites in promoters, matrices
  - public v. 6.0 contains 336 matrices, professional v. 7.4.1 contains 695 matrices
- TRANSCOMPEL
  - composite regulatory elements (pairs of closely situated sites and transcription 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

Transfac tables





SITE TABLE

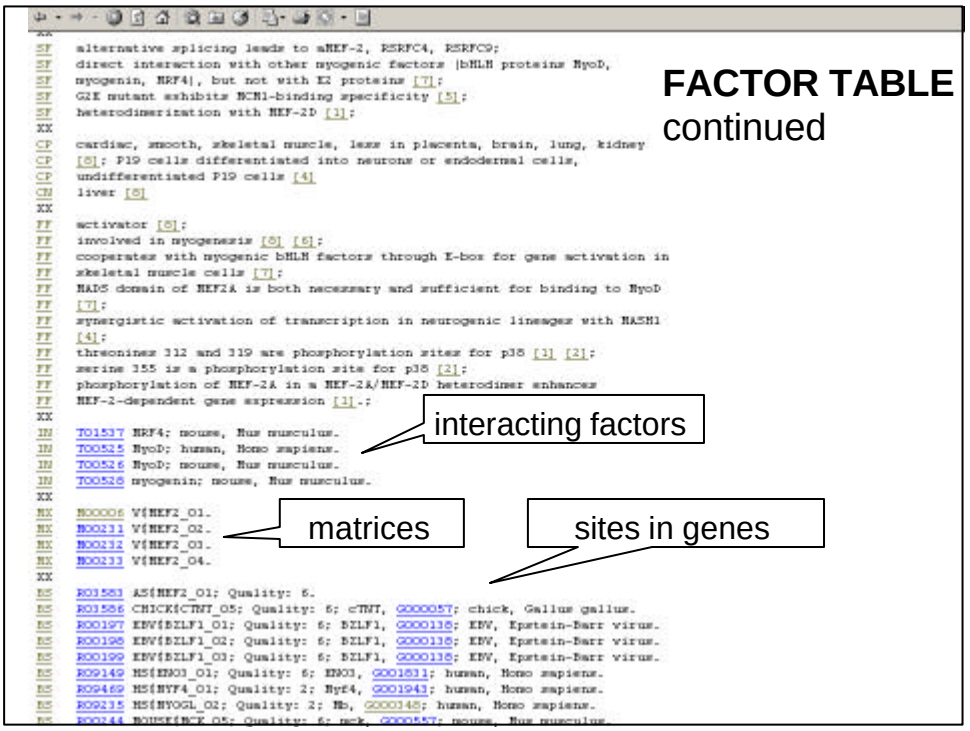


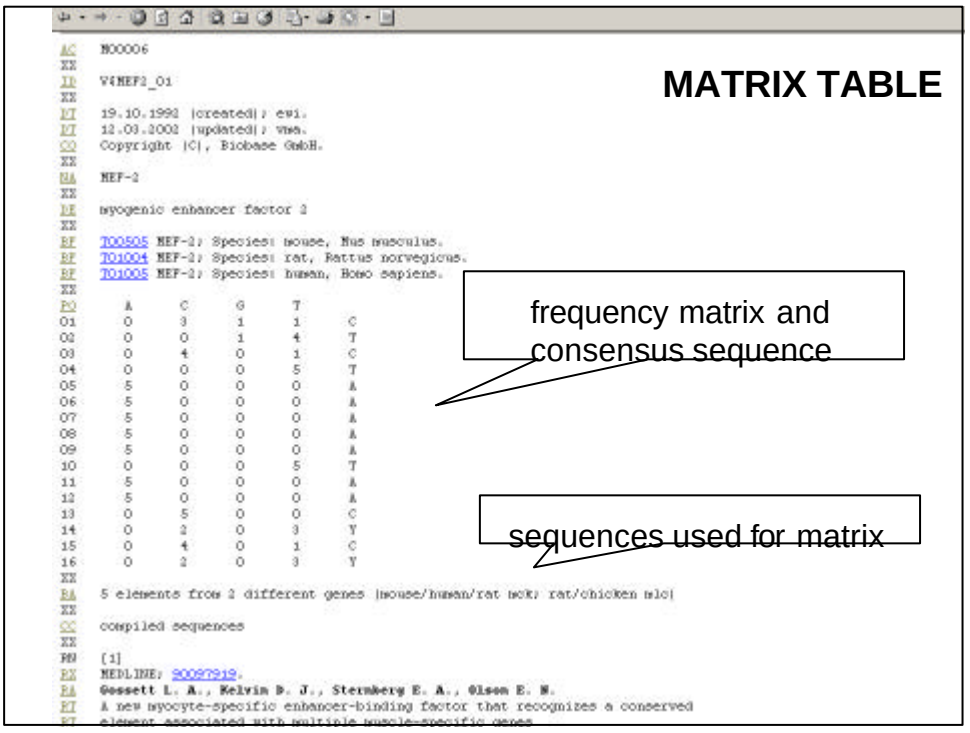
factor

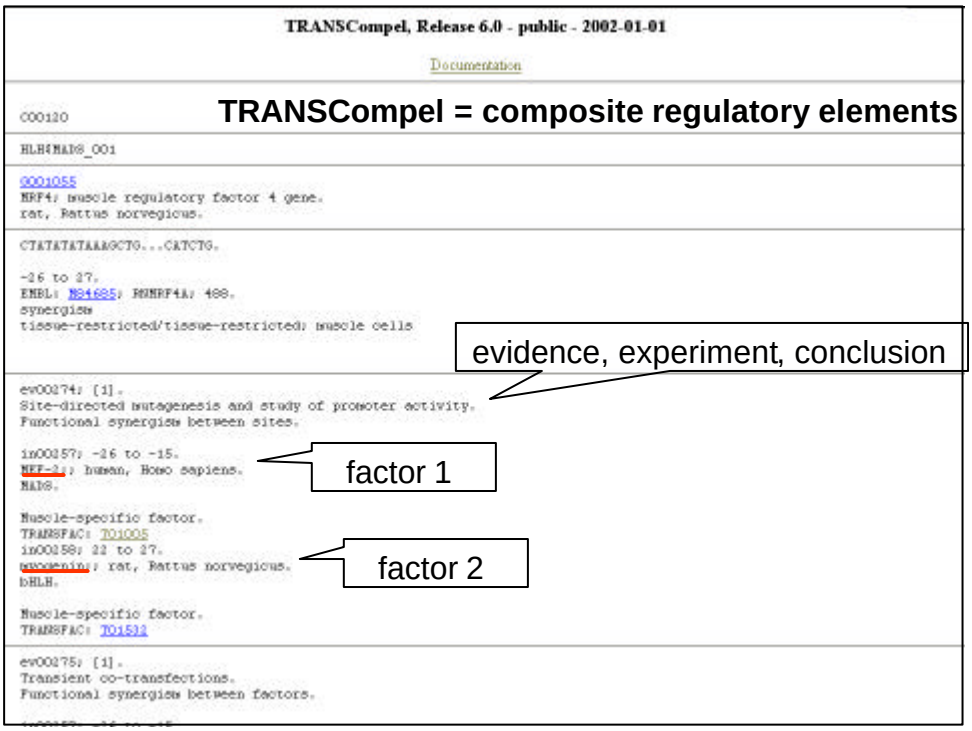
cellular source

FACTOR TABLE

class

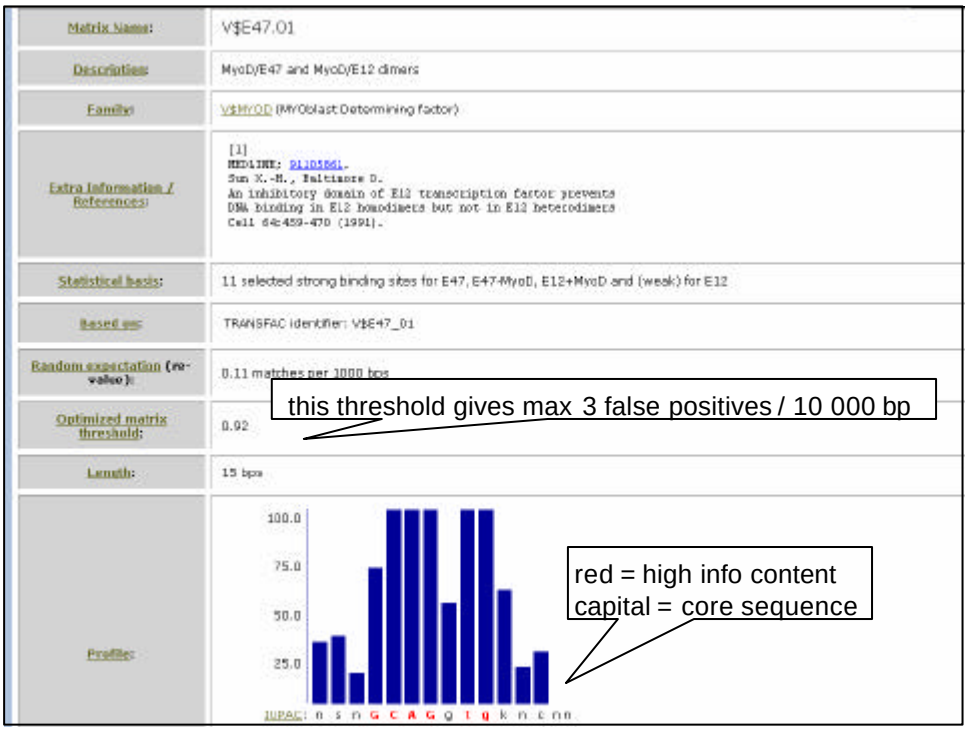

FACTOR TABLE
continued
interacting factors
matrices
sites in genes


MATRIX TABLE
frequency matrix and
consensus sequence
sequences used for matrix



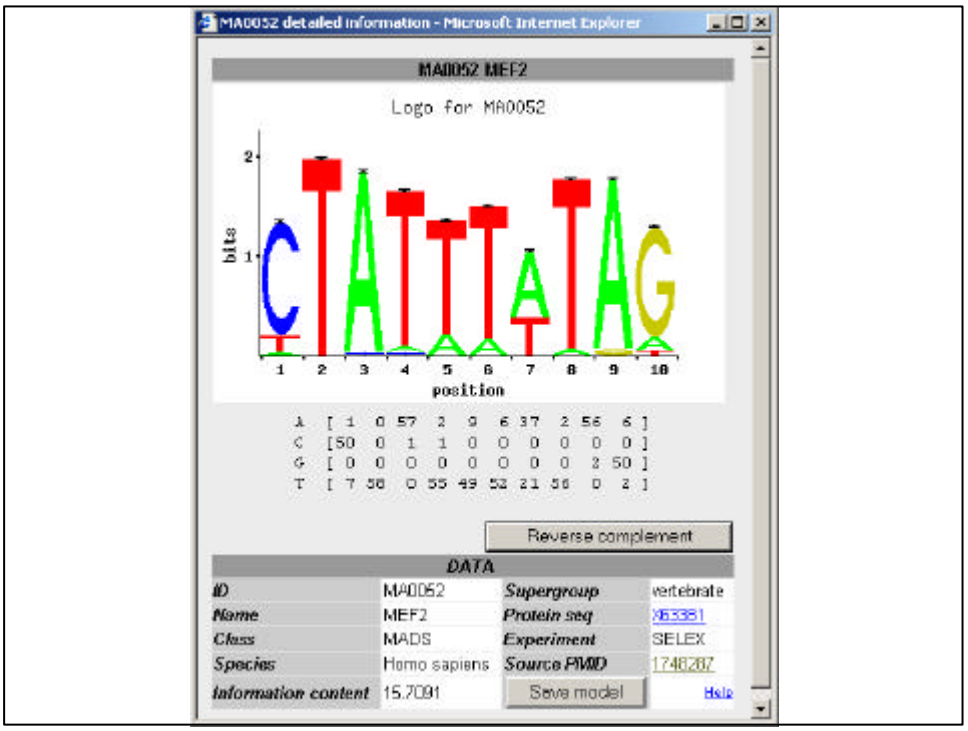
The ”Genomatix family”

- 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
  - ElDorado: 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/



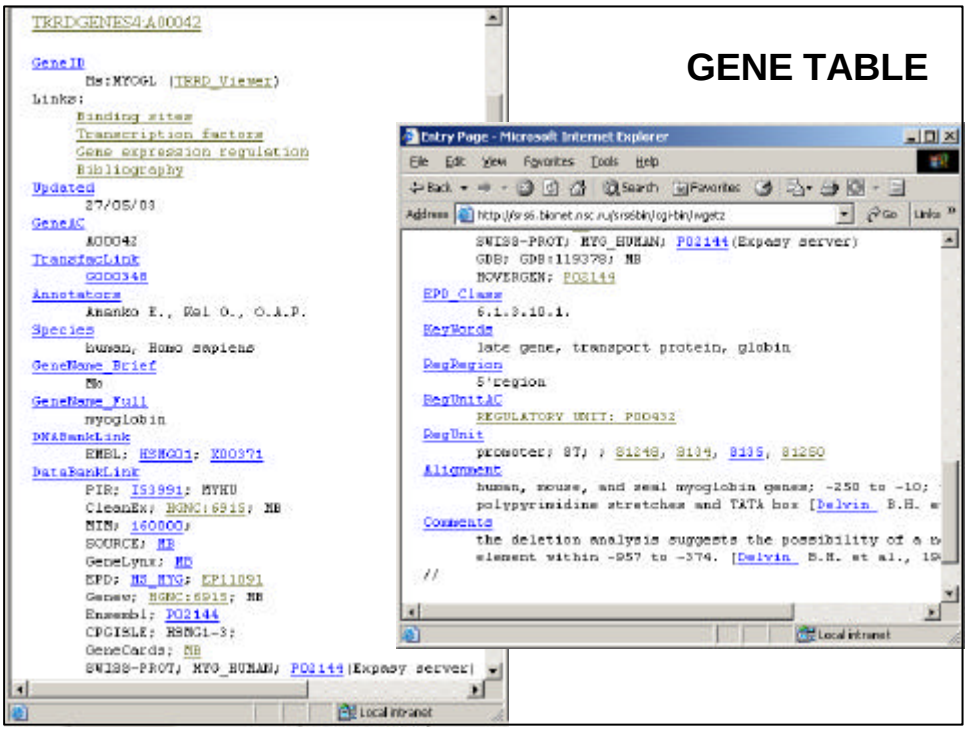
JASPAR

- 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

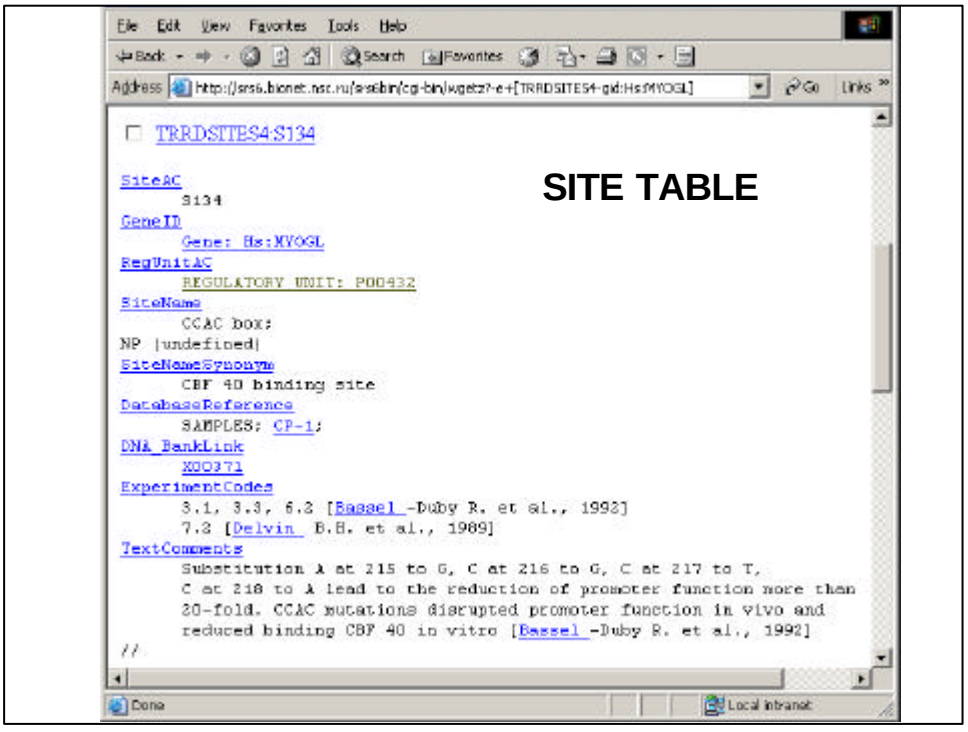


Transcription Regulatory Regions Database (TRRD)

- 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



SITE TABLE

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

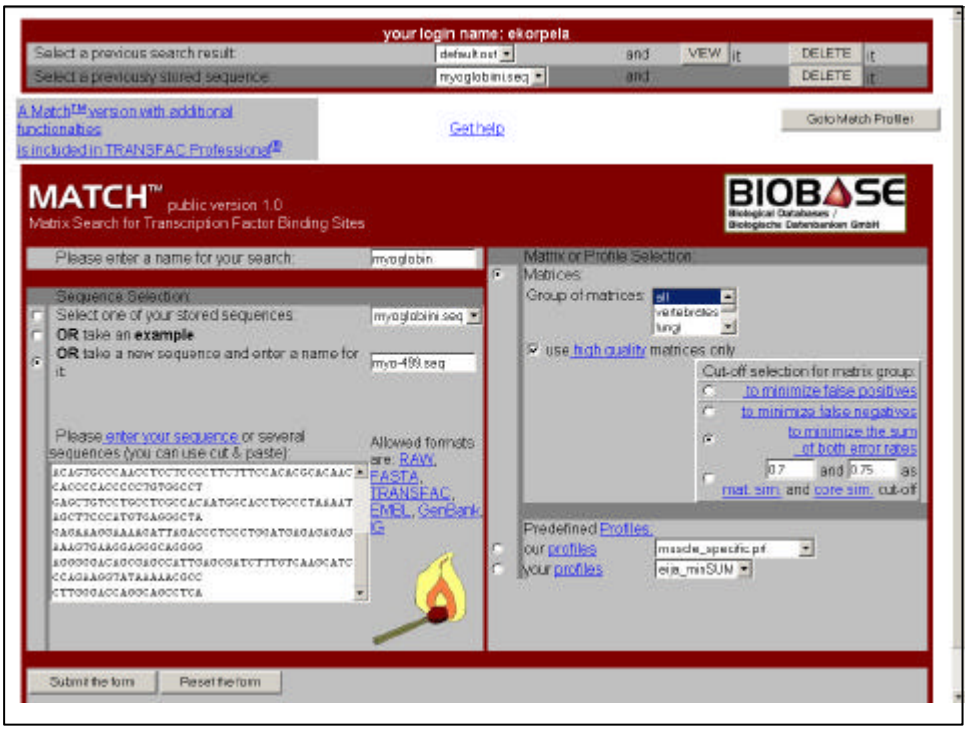
Programs for matrix scans

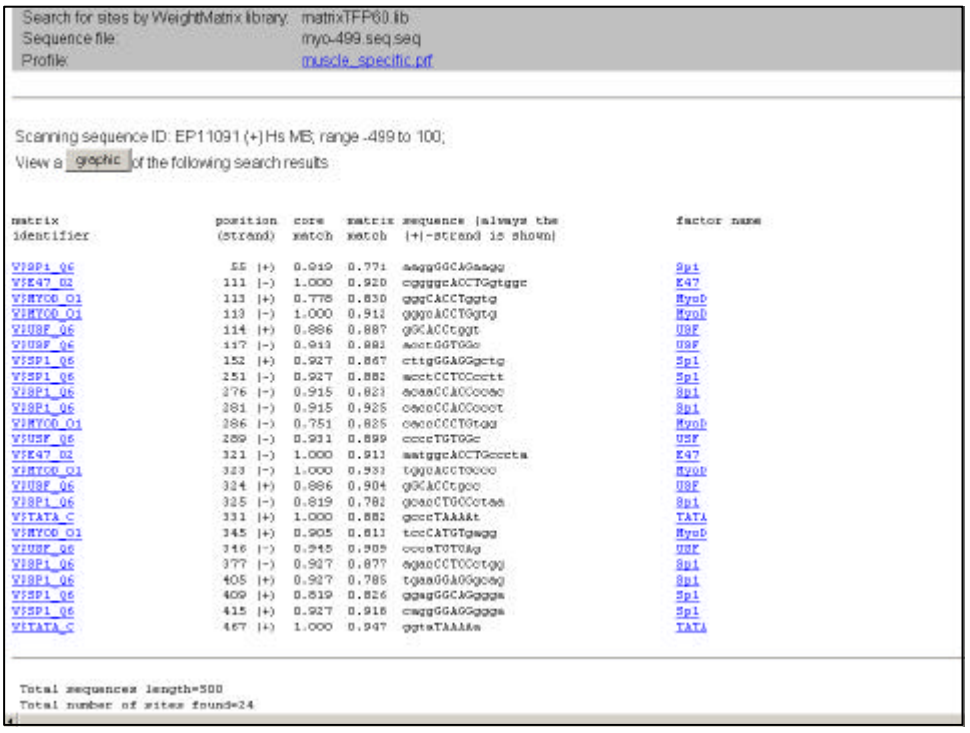
Match

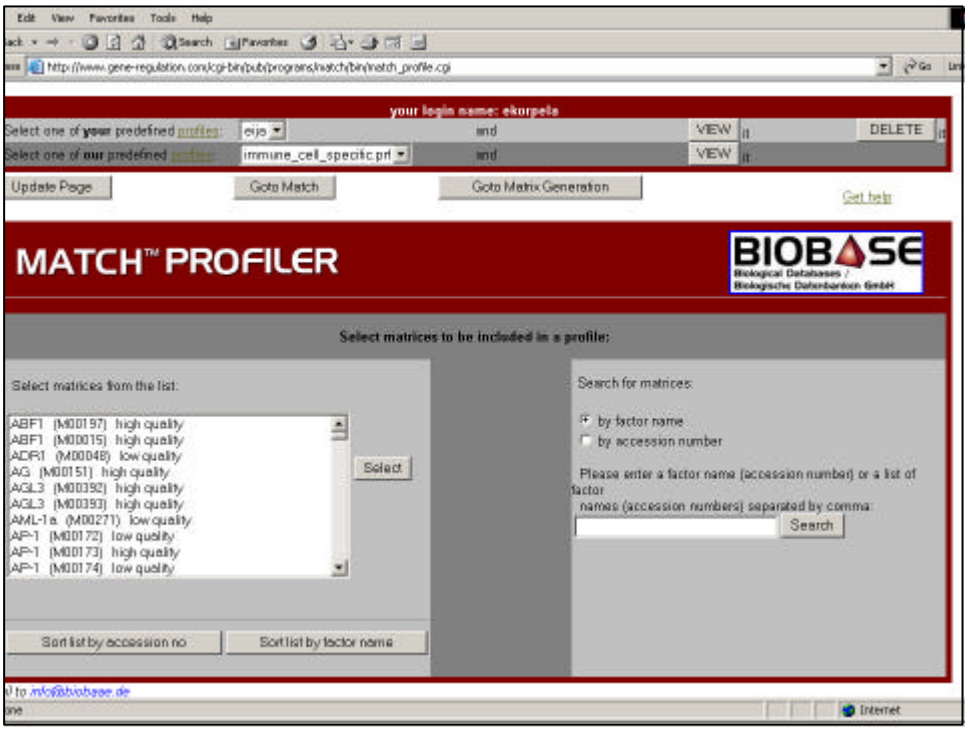
- uses Transfac matrices
- thresholds
  - minimize false negatives (minFN)
  - minimize false positives (minFP)
  - minimize the sum of both error rates
  - matrix similarity (takes information content into account)
  - core similarity (5 most conserved consecutive positions)
- profiles (matrix subset 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

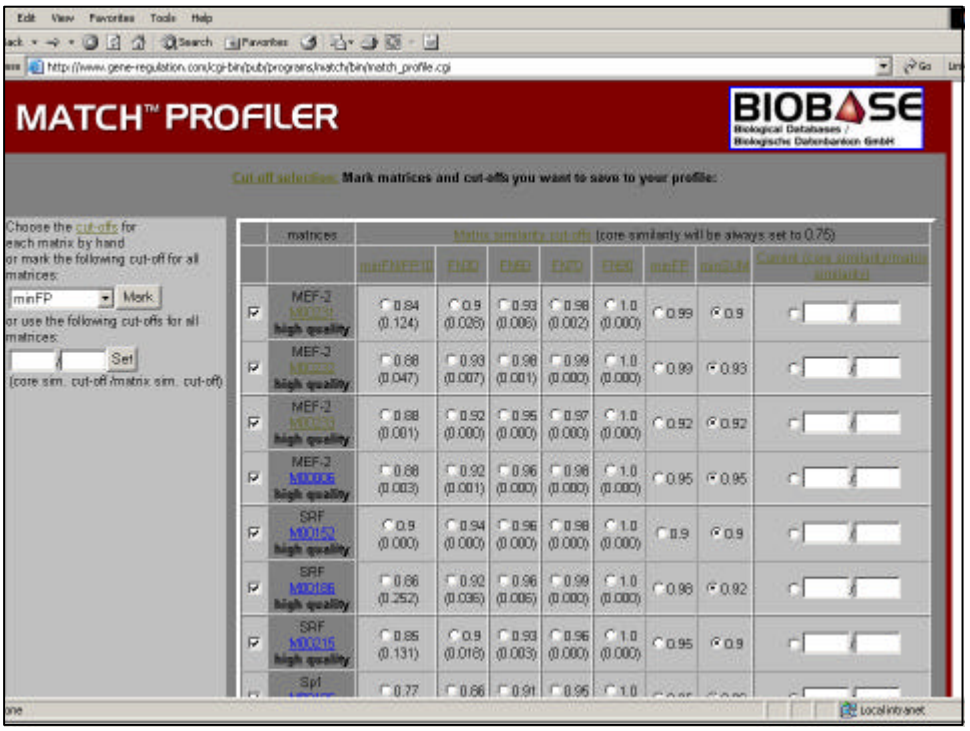
Match thresholds

- 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
it
DELETE
it
Select a previously stored sequence:
myoglobin.seq
and
DELETE
it
A Match™ version with additional functionalities is included in TRANSFAC Professional®
Get help
Goto Match Profiler
MATCH™ public version 1.0
Matrix Search for Transcription Factor Binding Sites
BIOBASE
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):
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
muscle_specific.prf
your profiles
eija_miSUN
Submit the form
Reset the form

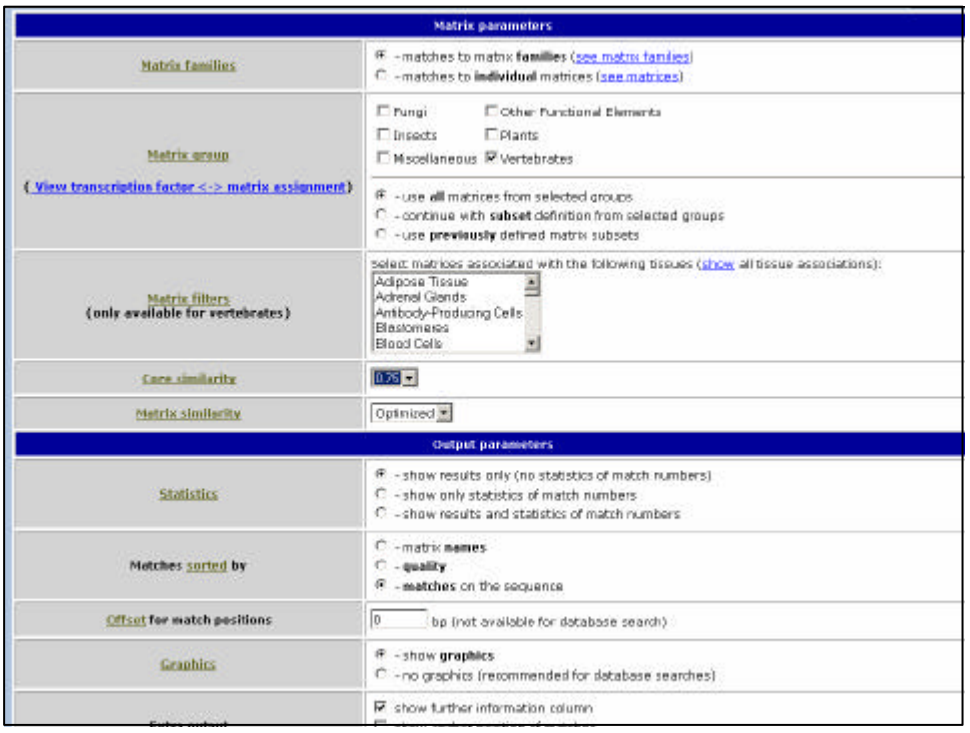

Search for sites by WeightMatrix library: matrixTFP60.lib
Sequence file: myo-499.seq.seq
Profile: muscle_specific.prf
Scanning sequence ID: EP11091 (+) Hs MB; range -499 to 100;
View a graphic of the following search results
matrix identifier | position (strand) | core match | matrix match | sequence (always the (+)-strand is shown) | factor name
V$SP1_Q6 | 55 | (+) | 0.919 | 0.771 | aaggGGCAGaagg | Sp1
V$E47_02 | 111 | (-) | 1.000 | 0.920 | cggggcACCTGgtggc | E47
V$MYOD_01 | 113 | (+) | 0.778 | 0.830 | gggCACCTggtg | MyoD
V$MYOD_01 | 113 | (-) | 1.000 | 0.912 | ggcaACCTGgtg | MyoD
V$USF_Q6 | 114 | (+) | 0.886 | 0.887 | gGCACCtggt | USF
V$USF_Q6 | 117 | (-) | 0.913 | 0.881 | acctGGTGCc | USF
V$SP1_Q6 | 152 | (+) | 0.927 | 0.867 | cttgGGAGGgctg | Sp1
V$SP1_Q6 | 251 | (-) | 0.927 | 0.881 | acctCCTCCcctt | Sp1
V$SP1_Q6 | 276 | (-) | 0.915 | 0.821 | acaaCCACCccac | Sp1
V$SP1_Q6 | 281 | (-) | 0.915 | 0.925 | cacaCCACCcact | Sp1
V$MYOD_01 | 286 | (-) | 0.751 | 0.825 | cacaCCCTGtgg | MyoD
V$USF_Q6 | 289 | (-) | 0.931 | 0.899 | cccaTGTGGc | USF
V$E47_02 | 321 | (-) | 1.000 | 0.911 | aatggcACCTGcccta | E47
V$MYOD_01 | 323 | (-) | 1.000 | 0.931 | tggcACCTGccc | MyoD
V$USF_Q6 | 324 | (+) | 0.886 | 0.904 | gGCACCtgcc | USF
V$SP1_Q6 | 325 | (-) | 0.819 | 0.781 | gcaaCTGCCctaa | Sp1
V$TATA_C | 331 | (+) | 1.000 | 0.881 | gcccTAAAAt | TATA
V$MYOD_01 | 345 | (+) | 0.905 | 0.811 | tccCATGTgagg | MyoD
V$USF_Q6 | 346 | (-) | 0.915 | 0.905 | cccaTGTCAg | USF
V$SP1_Q6 | 377 | (-) | 0.927 | 0.877 | agaaCCTCCctgg | Sp1
V$SP1_Q6 | 405 | (+) | 0.927 | 0.785 | tgaaGGAGGcctg | Sp1
V$SP1_Q6 | 409 | (+) | 0.819 | 0.826 | ggagGGCAGgggg | Sp1
V$SP1_Q6 | 415 | (+) | 0.927 | 0.916 | cagggGGAGGgggg | Sp1
V$TATA_C | 467 | (+) | 1.000 | 0.947 | ggtaTAAAAa | TATA
Total sequences length=500
Total number of sites found=24

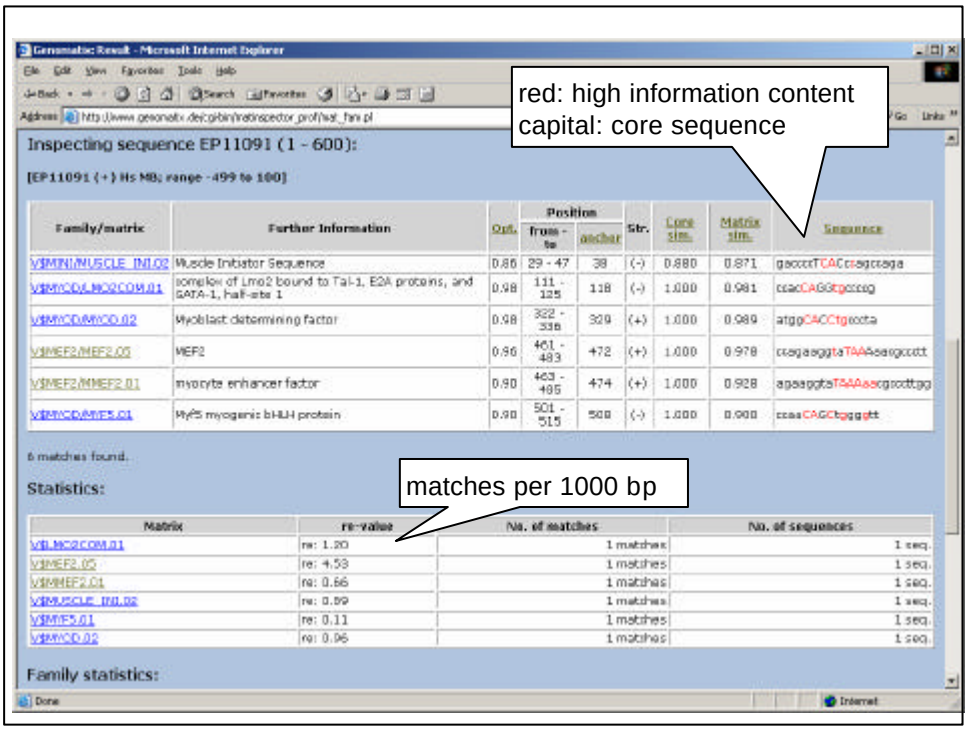

your login name: ekorpela
Update Page
Goto Match
Goto Matrix Generation
MATCH™ PROFILER
BIOBASE
Select matrices to be included in a profile:
Select matrices from the list:
Select
Search for matrices:
by factor name
by accession number
Search
Sort list by accession no
Sort list by factor name


MATCH™ PROFILER
BIOBASE
Mark matrices and cut-offs you want to save to your profile:
matrices
Mark
Set

MatInspector professional

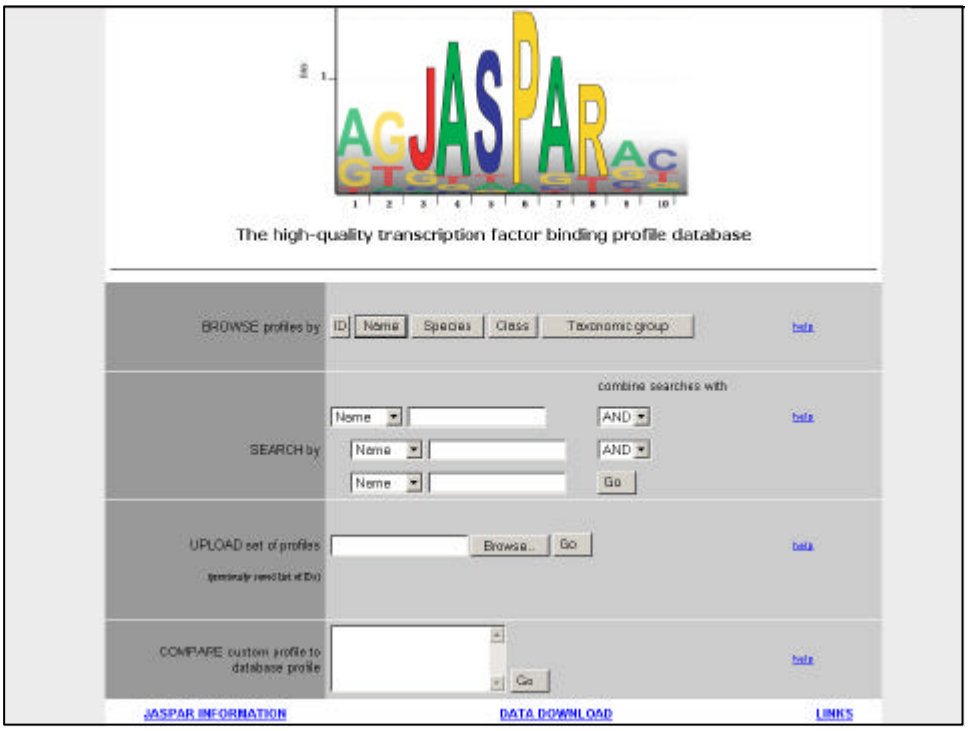
- Genomatix tool
- thresholds
  - matrix similarity (takes information content into account)
  - core similarity (4 most conserved consecutive positions)
  - optimized matrix threshold (max 3 hits per 10 000 bp of non-regulatory test sequence)
- matrices grouped into families
  - similar TFBSs with a 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





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


JASPAR
The high-quality transcription factor binding profile database
BROWSE profiles by
ID
Name
Species
Class
Taxonomic group
help
SEARCH by
combine searches with
Name
AND
Go
UPLOAD set of profiles
Browse...
COMPARE custom profile to database profile
JASPAR INFORMATION
DATA DOWNLOAD
LINKS

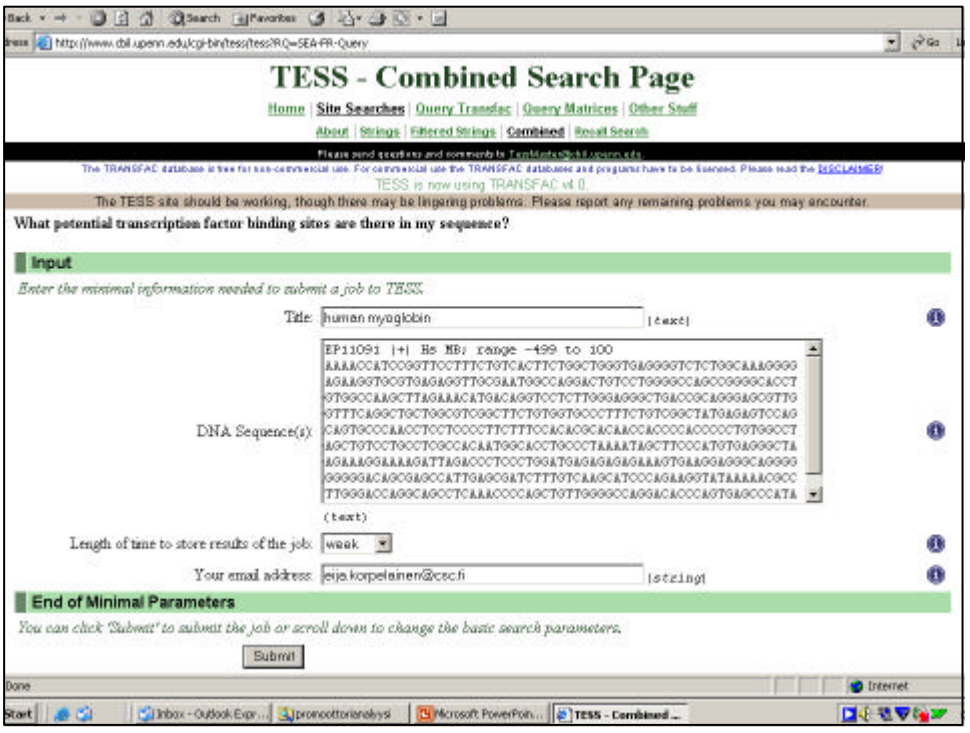

SEARCH
Name
AND
Species
Class
Submit
ID
NAME
SPECIES
CLASS
LOGO
MA0001
AGL3
Arabidopsis thaliana
MADS
MA0002
AML-1
Homo sapiens
RUNT
MA0003
AP2alpha
AP2
MA0004
ARNT
Mus musculus
bHLH
MA0110
ATHB5
Arabidopsis thaliana
HOMEO-ZIP
MA0005
Agamous
Arabidopsis thaliana
MADS
MA0006
Ahr-ARNT
Mus musculus
bHLH
view
Analyze this (fasta-formatted) sequence with selected
profile score threshold
10
%
Analyze

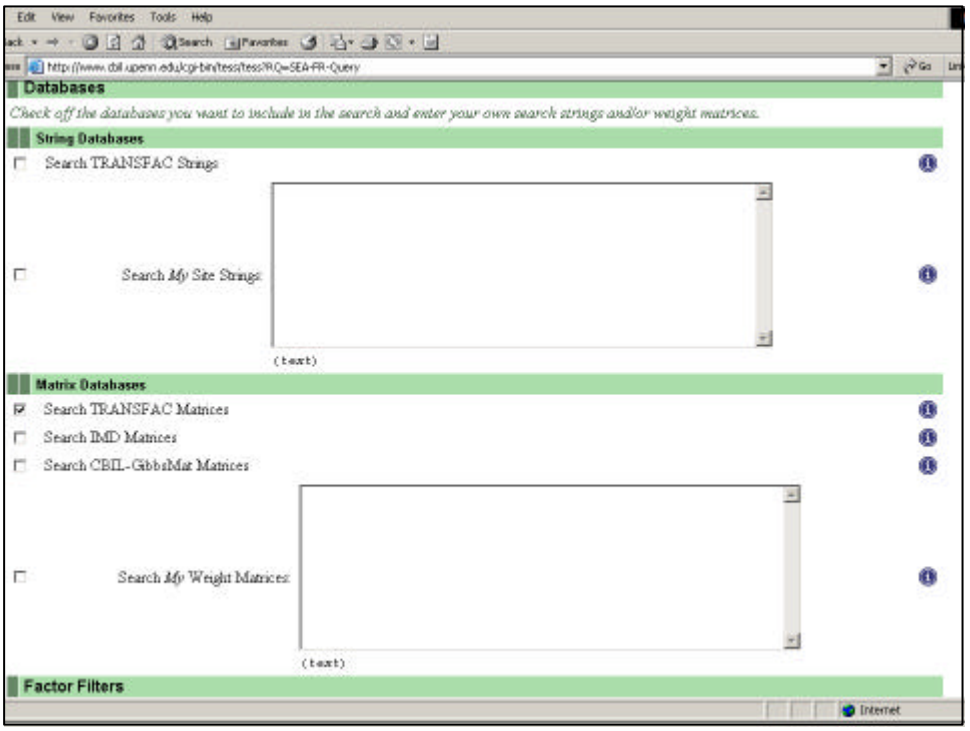
TESS

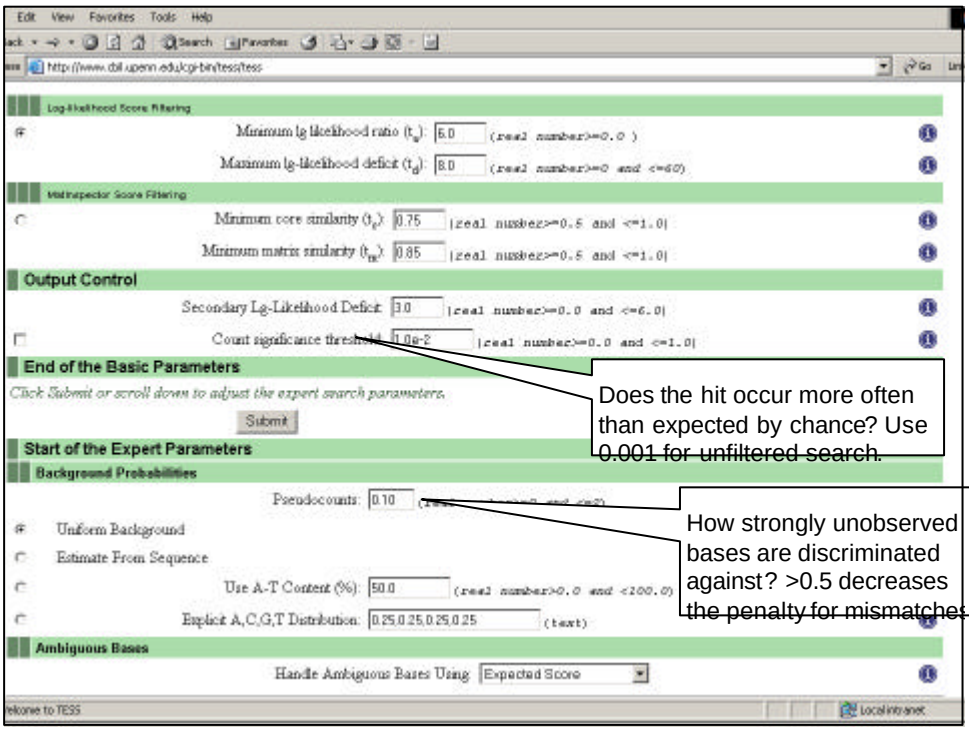
- uses Transfac 4.0 public
- thresholds
  - matrix similarity and core similarity
  - minimum log likelyhood ratio and deficit
  - significance threshold
- several filters to select matrices
- possibility to alter pseudocounts and background
- string search included
- http://www.cbil.upenn.edu/tess/

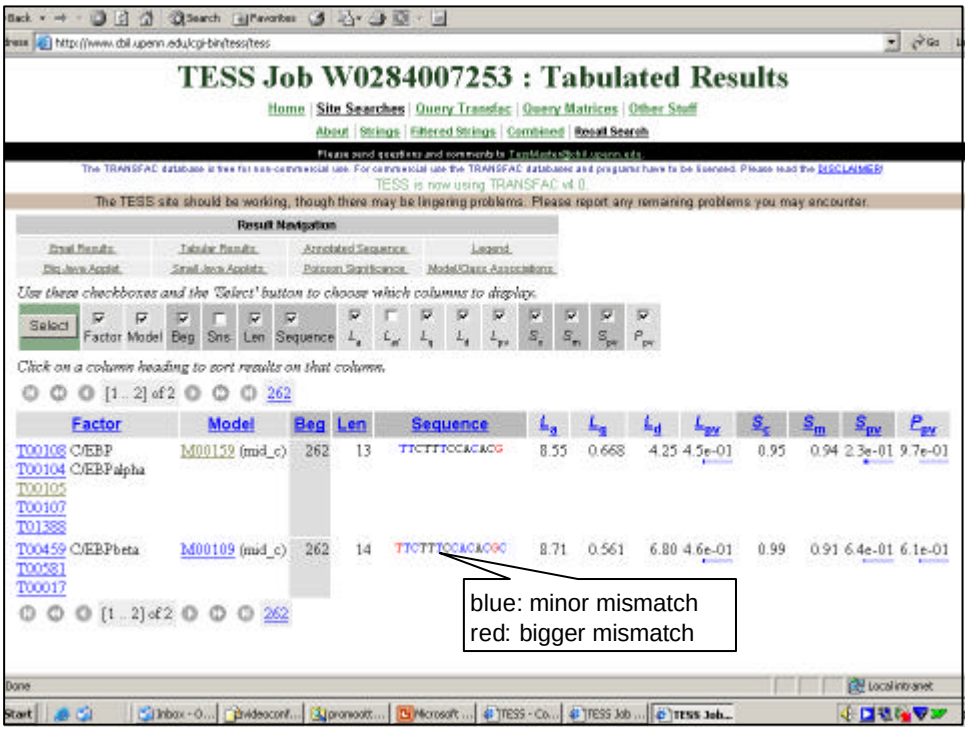
TESS filters

- 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)




Edit View Favorites Tools Help
Databases
Check off the databases you want to include in the search and enter your own search strings and/or weight matrices.
String Databases
Search TRANSFAC Strings
Search My Site Strings:
(text)
Matrix Databases
Search TRANSFAC Matrices
Search IMD Matrices
Search CBIL-GibbsMat Matrices
Search My Weight Matrices:
(text)
Factor Filters
Internet


Edit View Favorites Tools Help
Minimum lg likelihood ratio (t_s): 6.0 (real number>=0.0)
Maximum lg-likelihood deficit (t_d): 8.0 (real number>=0 and <=60)
Minimum core similarity (t_c): 0.75 (real number>=0.5 and <=1.0)
Minimum matrix similarity (t_m): 0.85 (real number>=0.5 and <=1.0)
Output Control
Secondary Lg-Likelihood Deficit: 3.0 (real number>=0.0 and <=6.0)
Count significance threshold: 1.0e-2 (real number>=0.0 and <=1.0)
End of the Basic Parameters
Click Submit or scroll down to adjust the expert search parameters.
Submit
Start of the Expert Parameters
Background Probabilities
Pseudocounts: 0.10
Uniform Background
Estimate From Sequence
Use A-T Content (%): 50.0 (real number>0.0 and <100.0)
Explicit A,C,G,T Distribution: 0.25,0.25,0.25,0.25 (text)
Ambiguous Bases
Handle Ambiguous Bases Using: Expected Score
Does the hit occur more often than expected by chance? Use 0.001 for unfiltered search.
How strongly unobserved bases are discriminated against? >0.5 decreases the penalty for mismatches
Local intranet



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 backgound 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

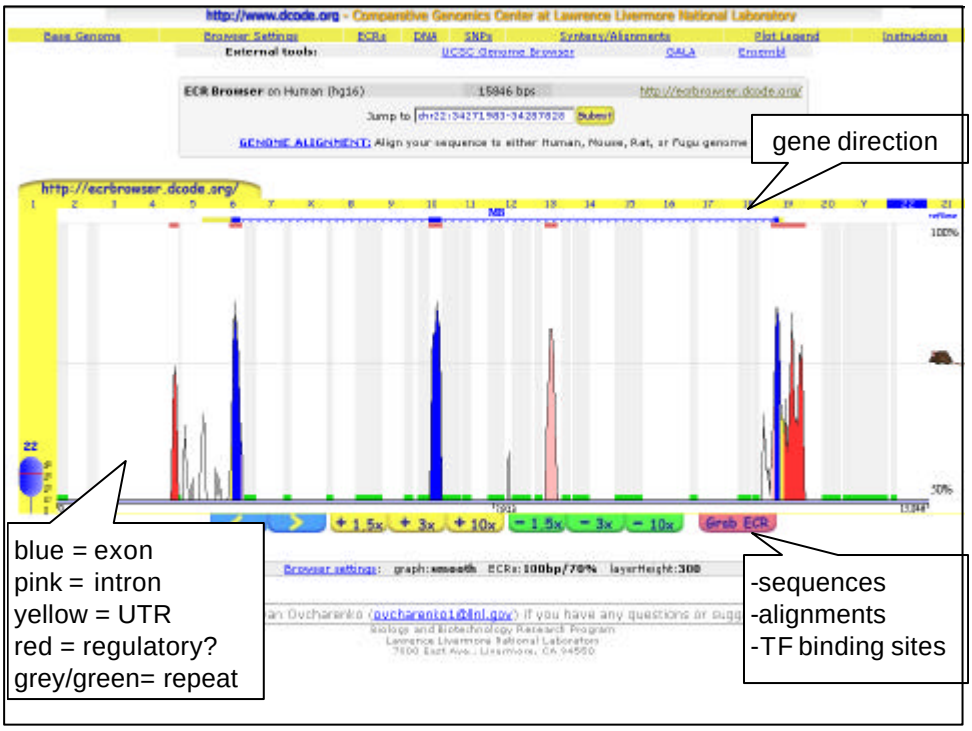
- regulatory regions tend to be evolutionarily conserved
- align orthologous genes
  - arose from a common ancestral gene through speciation
  - likely to be involved in similar biological functions
- suitable evolutionary distance
  - 70 million years reveals most regulatory regions (eg. human/mouse)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

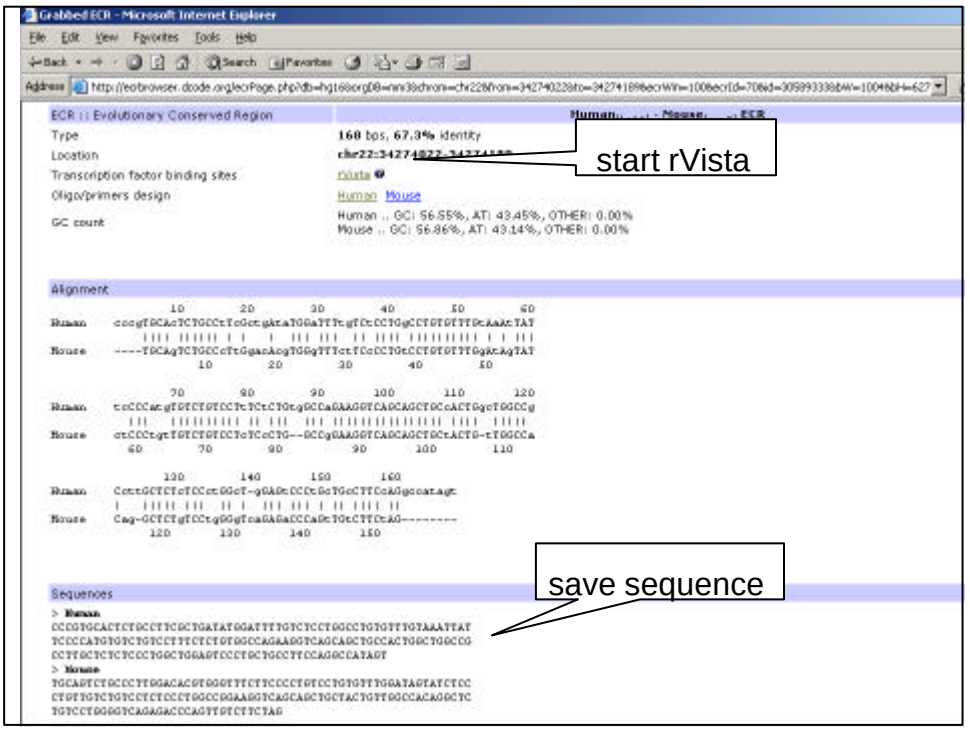
What kind of alignment?

- 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

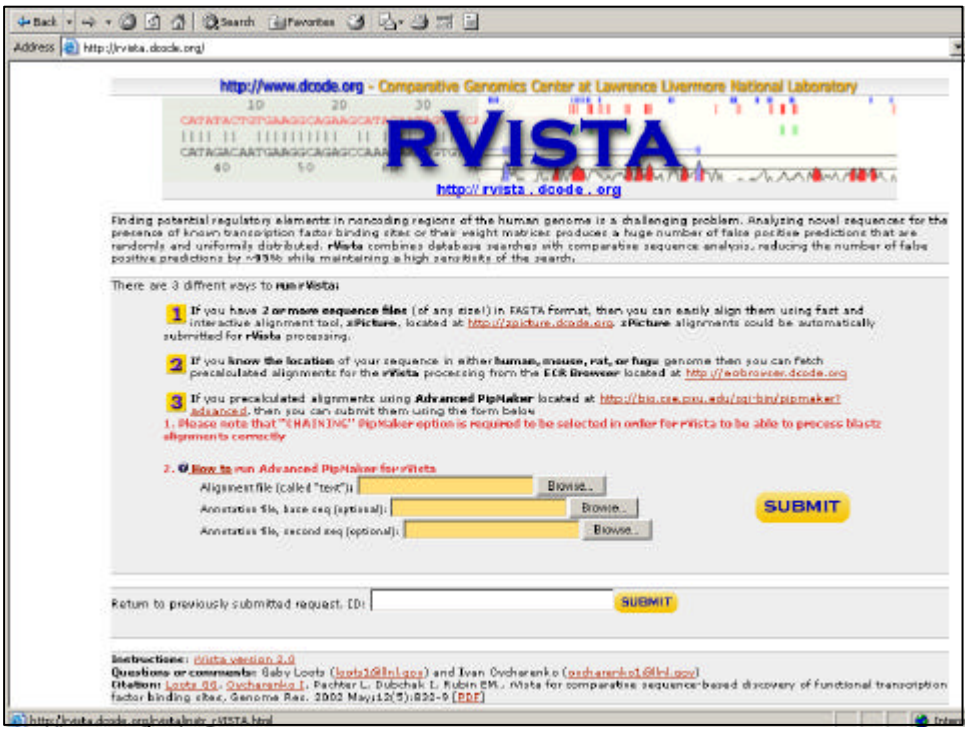
- ECR = evolutionary conserved region
- visualizes all pairwise genomic alignments between human, mouse, rat, Fugu, Tetraodon, and zebrafish genomes
- extract sequences that correspond to any ECR, view sequence alignments, send results to further analysis (rVista)
- http://www.dcode.org/

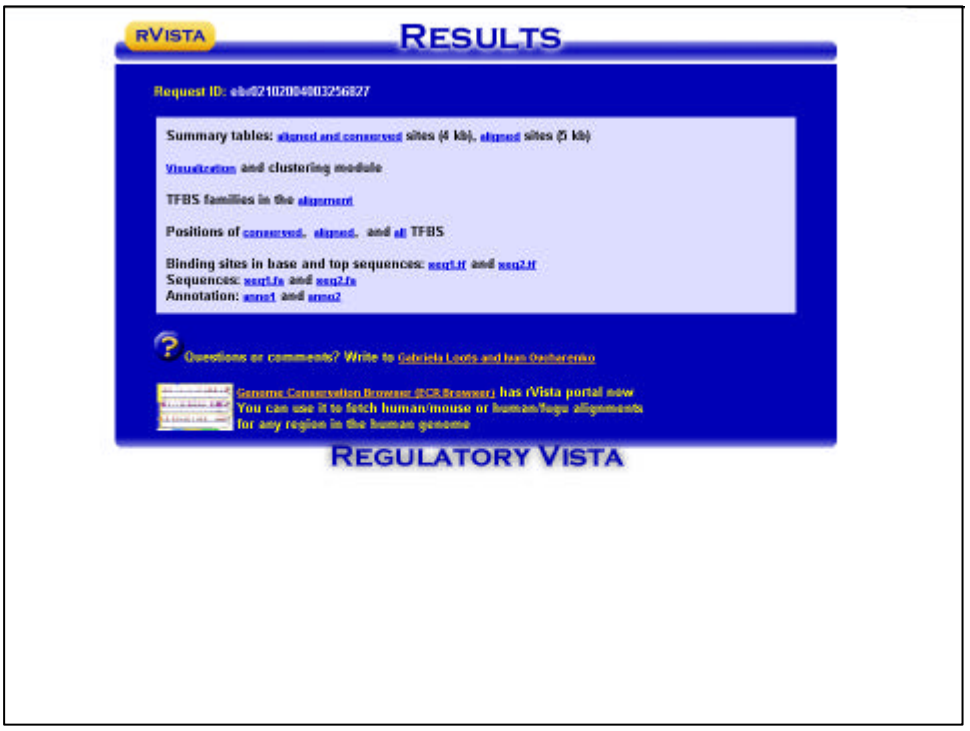

gene direction
blue = exon
pink = intron
yellow = UTR
red = regulatory?
grey/green= repeat
-sequences
-alignments
-TF binding sites

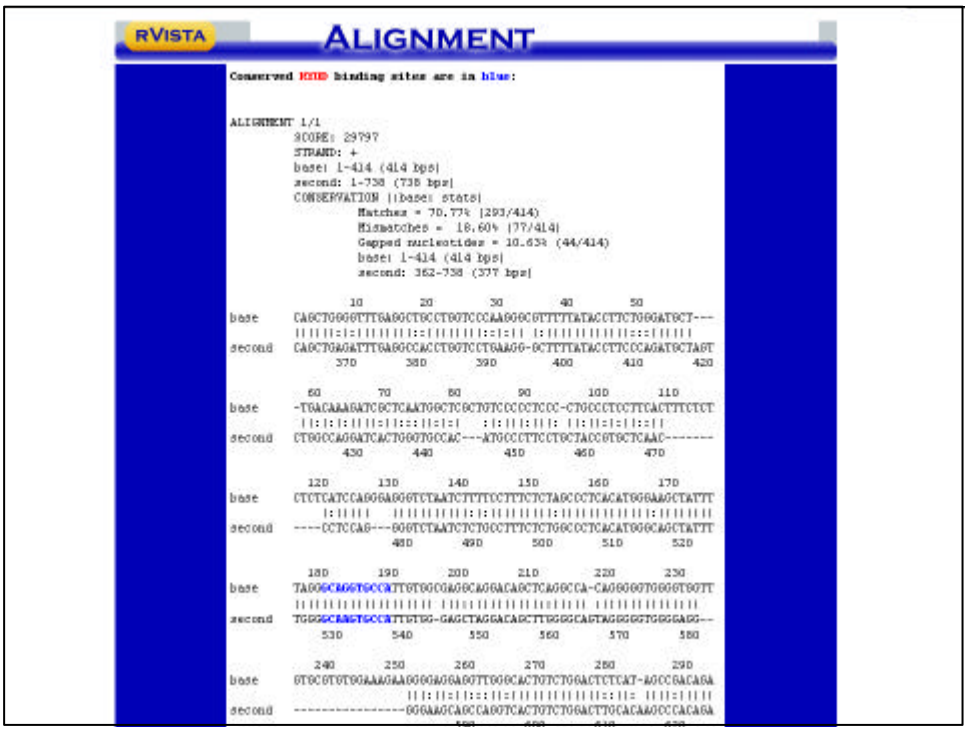

start rVista
save sequence

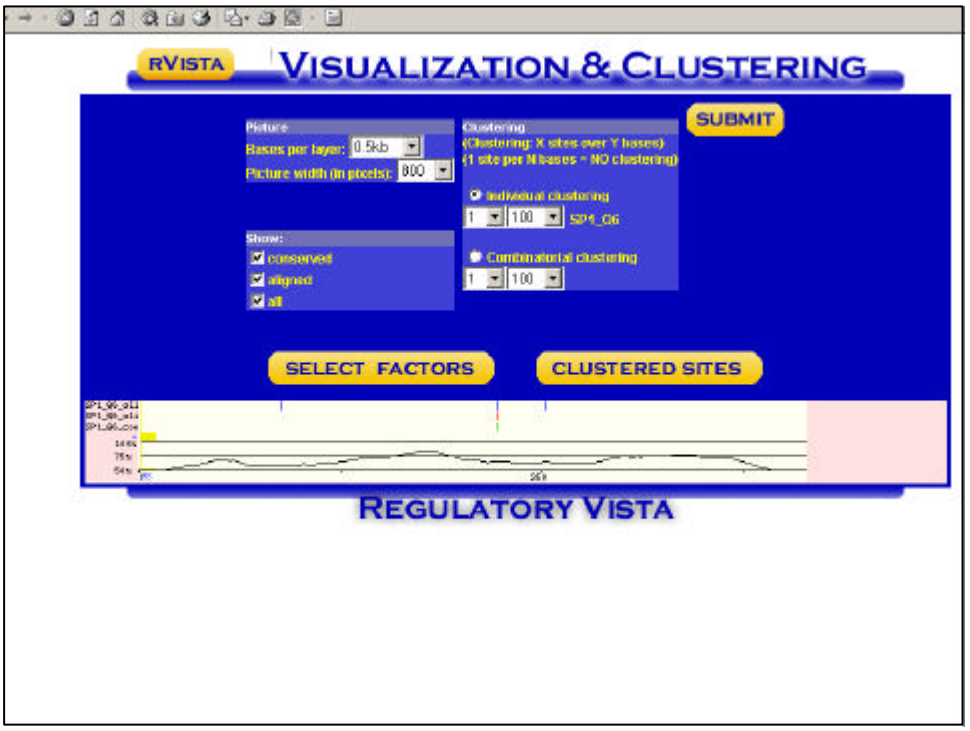
rVista

- 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/




rVISTA
RESULTS
Request ID: ebi021020040032568827
Summary tables: aligned and conserved sites (4 kb), aligned sites (5 kb)
Visualization and clustering module
TFBS families in the alignment
Positions of conserved, aligned, and all TFBS
Binding sites in base and top sequences: seq1.tf and seq2.tf
Sequences: seq1.fa and seq2.fa
Annotation: anno1 and anno2
Questions or comments? Write to Gabriela Loots and Ivan Ovcharenko
Genome Conservation Browser (GCB Browser) has rVista portal now
You can use it to fetch human/mouse or human/fugu alignments
for any region in the human genome
REGULATORY VISTA


rVISTA
ALIGNMENT
Conserved KCND binding sites are in blue:
ALIGNMENT 1/1
SCORE: 29797
STRAND: +
base: 1-414 (414 bps)
second: 1-738 (738 bps)
CONSERVATION (base: stats)
Matches = 70.77% (293/414)
Mismatches = 18.60% (77/414)
Gapped nucleotides = 10.63% (44/414)
base: 1-414 (414 bps)
second: 362-738 (377 bps)


rVista
Visualization & Clustering
Picture
Bases per layer: 0.5kb
Picture width (in pixels): 800
Show:
conserved
aligned
all
Clustering
(Clustering: X sites over Y bases)
(1 site per N bases = NO clustering)
Individual clustering
1 100 SP1_06
Combinatorial clustering
1 100
SUBMIT
SELECT FACTORS
CLUSTERED SITES
Regulatory Vista