Getting Started Guide: TRANSFAC® Match - Frequently Asked Questions

How can it be that in some cases the multiple matrices for one and the same factor are so different from each other, and which one is the best to use?

There are pronounced differences in the experimental proof underlying the data. The matrices with just a number as appendix _01, _02, ... were taken from the cited reference. Often they were derived by random binding site selection (SELEX), where a specific, often recombinant factor or its isolated binding domain was studied. Other matrices were built by us through compilation of genomic (and sometimes artificial) binding sites for orthologous factors of a broad taxonomic group, as for example vertebrates, insects, plants, fungi, etc. Their appendices _Q1, _Q2, ... refer to the 'quality' of the sites used, i.e. the certainty with which it could be concluded that the binding activity shown was identical with the suggested factor. (Other matrices again, e.g. those ending in _C, where built using specific programs, like CONSIND.) Depending on the experimental material and method/conditions on one hand and the choice and alignment of binding sites on the other hand, i.e. which and how many of the possible 'manifestations' of a 'binding site' were selected/compiled, the derived matrices can differ.

There are no general rules as to which matrices are the best ones to use. You can restrict Match™ to the use of so-called "high quality matrices" only, with a lower number of false positive matches. For vertebrates, "non redundant" profiles are available, where for a group of related factors with similar binding specificity only the matrix with the lowest FP rate at the respective cut-off is included.

Are all TRANSFAC® SITE entries represented in (at least) one of the matrices in the matrix library of Match™ ?

No, to build a matrix you need several binding sites for a transcription factor. For some factors there are not yet enough sites available.

Are all matrix core sequences 5 in length? Are the core sequences always consecutive nucleotides?

Yes, we use the five most conserved, consecutive nucleotides as core sequences for all matrices.

How is the score for the core/matrix similarity calculated?

MatchTM searches for subsequences x of an input sequence s, which are good matches to a matrix of TRANSFAC®. The quality of a match is described by two values: the core similarity and the matrix similarity. The score for the matrix similarity of a subsequence x of sequence s with length L is calculated in the following way:

$$mat\_sim = mat\_sim(x) = \frac{Current - Min}{Max - Min}$$

where: $(b1,\ldots,BL)$ are the nucleotides of x

fi,B is the frequency of B to occur at position i and:

$$current = \sum_{i=1}^{L} I(i) f_{i,b_i}$$ current frequency

$$min = \sum_{i=1}^{L} I(i) f_i^{\min}$$ minimum frequency

$$max = \sum_{i=1}^{L} I(i) f_i^{\max}$$ maximum frequency

$$I(i) = \sum_{B \in \{A,T,G,C\}} f_{i,B} \ln(4 f_{i,B})$$ information vector

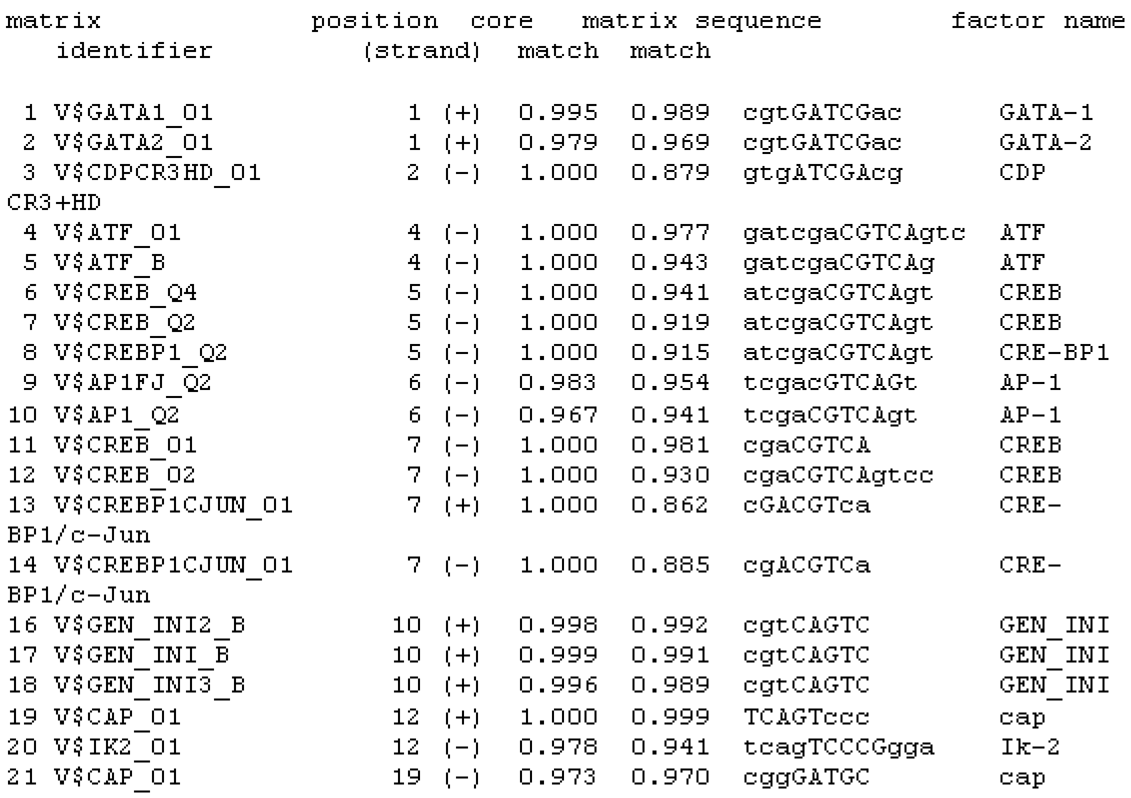
The score for the core similarity is calculated similarly to the matrix similarity.

What do I need for promoter analysis?

TRANSFAC® Professional and the tool Match™ provided with it allow a first/initial promoter analysis. The Match™ program uses a library of mononucleotide weight matrices from TRANSFAC® Professional to search in sequences submitted by the user for potential transcription factor binding sites. More refined promoter analysis, including comparative analysis of a set of co-regulated promoters against a background set or search for composite modules, is provided within the ExPlain™ Analysis System.

How can I confirm that my sequence is actually a TF binding site?

Match™ is defined to identify transcription factor binding sites in uncharacterized sequences by comparing them to a library of distribution matrices that are linked to the respective entries in the TRANSFAC® matrix table. As a result set you get a list of binding matrices indicating where they match your sequence and how good the match is. These distribution matrices have been derived from sites within the DNA for which binding of a specific transcription factor was shown. The binding sites for a group of orthologous transcription factors were aligned and then, at each position, the frequency of the four nucleotides (A,C,G,T) was counted. The derived distribution matrix contains more information than a simple IUPAC consensus, where nucleotides that are found at lower frequencies are neglected. (i.e. at a position where 60% of the sites showed an A, 20% T and 20% C, an A would appear in the IUPAC consensus, the same as for a position where in 100% of all sites an A was found, thus pretending both positions within the site to be equally conserved.) As matrices contain more information than an IUPAC consensus, sequence comparisons based on matrices are usually slower. To enhance performance, sequence comparison is done in two steps by Match™: In the first step, the 'core' of a matrix is compared to the sequence given by the user, and only where the core similarity is higher than the initially chosen threshold, the whole matrix is compared. The matrix-'core' used by Match™ consists of the five consecutive nucleotide positions, which together yield the highest conservation value. Thus, the 5bp-core (capital letters in the result set) serves to speed up the calculations, but it cannot define the whole binding matrix/sites sufficiently on its own. Let's say you are looking for potential binding sites in the following 26-bp sequence: cgtgatcgacgtcagtcccgggatgc. Scanning this sequence for matches to the subset of vertebrate matrices in the library using MATCH™ (with matrix similarity cut-off = 0.86, core similarity cut-off = 0.96) will give the following result set:



```
matrix               position   core   matrix sequence        factor name
   identifier           (strand)  match   match

  1 V$GATA1_01            1 (+)   0.995  0.989  cgtGATCGac       GATA-1
  2 V$GATA2_01            1 (+)   0.979  0.969  cgtGATCGac       GATA-2
  3 V$CDPCR3HD_01         2 (-)   1.000  0.879  gtgATCGAcg       CDP
CR3+HD
  4 V$ATF_01              4 (-)   1.000  0.977  gatcgaCGTCAgtc   ATF
  5 V$ATF_B               4 (-)   1.000  0.943  gatcgaCGTCAg     ATF
  6 V$CREB_Q4             5 (-)   1.000  0.941  atcgaCGTCAgt     CREB
  7 V$CREB_Q2             5 (-)   1.000  0.919  atcgaCGTCAgt     CREB
  8 V$CREBP1_Q2           5 (-)   1.000  0.915  atcgaCGTCAgt     CRE-BP1
  9 V$AP1FJ_Q2            6 (-)   0.983  0.954  tcgacGTCAGt      AP-1
 10 V$AP1_Q2              6 (-)   0.967  0.941  tcgaCGTCAgt      AP-1
 11 V$CREB_01             7 (-)   1.000  0.981  cgaCGTCA         CREB
 12 V$CREB_02             7 (-)   1.000  0.930  cgaCGTCAgtcc     CREB
 13 V$CREBP1CJUN_01       7 (+)   1.000  0.862  cGACGTca         CRE-
BP1/c-Jun
 14 V$CREBP1CJUN_01       7 (-)   1.000  0.885  cgACGTCa         CRE-
BP1/c-Jun
 16 V$GEN_INI2_B         10 (+)   0.998  0.992  cgtCAGTC         GEN_INI
 17 V$GEN_INI_B          10 (+)   0.999  0.991  cgtCAGTC         GEN_INI
 18 V$GEN_INI3_B         10 (+)   0.996  0.989  cgtCAGTC         GEN_INI
 19 V$CAP_01             12 (+)   1.000  0.999  TCAGTccc         cap
 20 V$IK2_01             12 (-)   0.978  0.941  tcagTCCCGgga     Ik-2
 21 V$CAP_01             19 (-)   0.973  0.970  cggGATGC         cap
```

For each match to a matrix position (within the above 26bp-sequence), orientation and similarity (of the core and of the whole matrix) are given. In the second to last column the fragment of the sequence which matched the matrix (orientation!) is shown (with the 'core' in capital letters). Capital letters within the sequence indicate the position of the core string within the matching matrix. Clicking on the matrix name (ID) gets you to the matrix entry in TRANSFAC®, where you can get information about the matrix and its binding factor. When you apply a stringent cut-off it is likely that you can prove binding of factors, belonging to the matching matrices, to the sequence in vitro. To draw any conclusions for the regulation in vivo of a promoter containing the above sequence is a bit more problematic, however, as this is dependent on the presence of other sites and the interaction of the factors binding to them.

The program ignores the parameters that I am selecting?

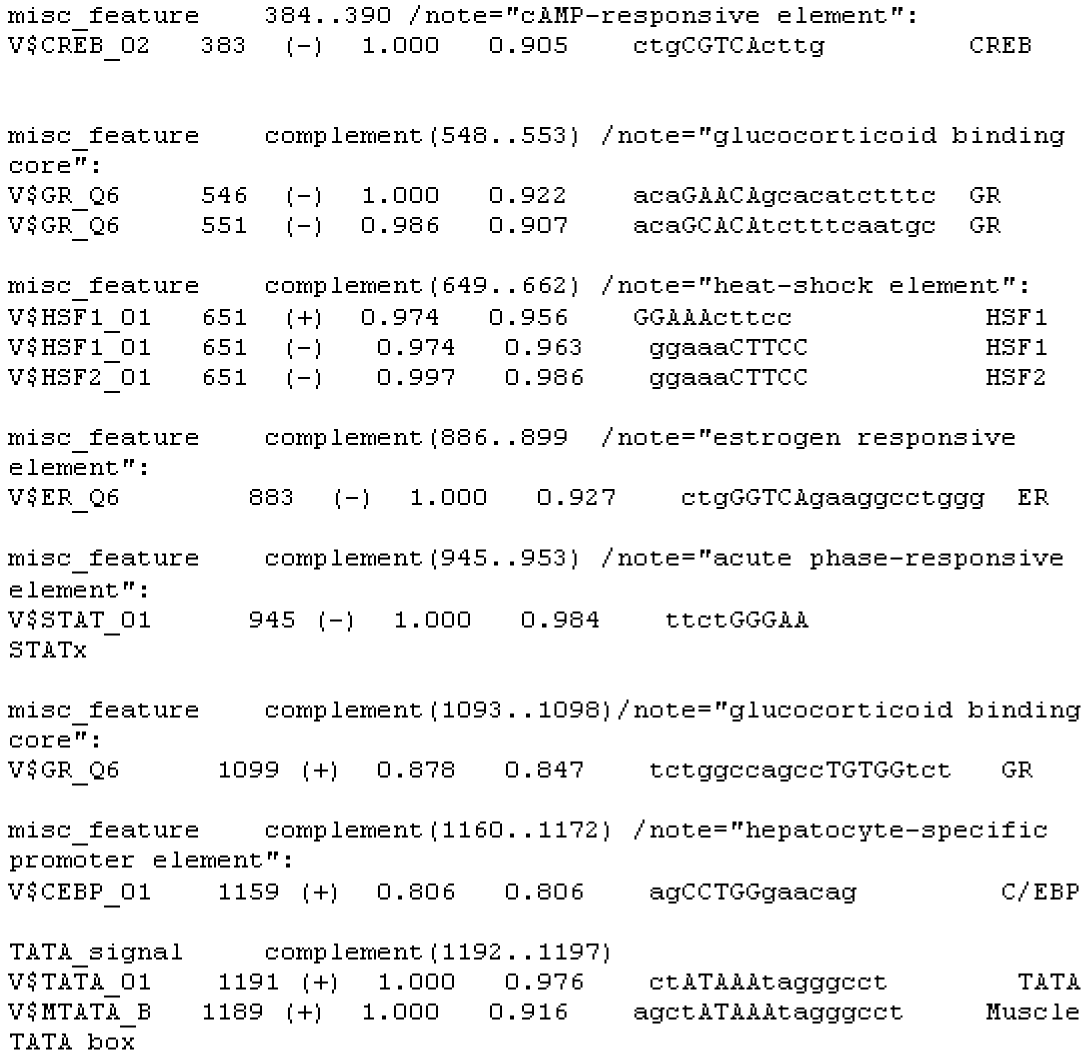
One reason could be that the radio buttons on the Match input page are not set at the correct position.

What can I do if MatchTM does not find all promoter elements listed in the "misc_features" of a Genbank report?

If you are looking for a binding site for a particular factor, please make sure that there is a matrix in TRANSFAC® for this factor. If matrices exist for this factor, this can be a problem of the profile and cut-off selection. If you use fairly high cut-offs, e.g. cut-offs to minimize false positive matches, you might miss sites. Cut-offs to minimize false positive matches try to filter out all possible random matches, but they do not guarantee that all "real sites" are found. If you want to make sure that no real site is missed, you should use cut-offs to minimize false negatives for your analysis. A cut-off that finds all "real" binding sites and filters out all random matches would be optimal. But in most cases it is not that easy to separate these two sets of sites. Cut-offs to minimize the sum of both error rates are just the best possible approximation.

Therefore, to make sure that you do not miss any real sites, use cut-offs to minimize false negative matches.

Here is one example, which shows that it is possible to find all known promoter elements with Match™. The promoter of the human angiotensinogen gene (Genbank Accession: X15323) was searched with MatchTM using cut-offs to minimize false negative matches. The list below shows the misc_features of the Genbank entry and the respective matches found by Match™ . For each matching matrix identifier, position and orientation, core similarity score, matrix similarity score, the matching sequence and the name of the binding factor are given.



```
misc_feature    384..390 /note="cAMP-responsive element":
V$CREB_02   383  (-)  1.000   0.905    ctgCGTCActtg         CREB


misc_feature    complement(548..553) /note="glucocorticoid binding
core":
V$GR_Q6     546  (-)  1.000   0.922    acaGAACAgcacatctttc  GR
V$GR_Q6     551  (-)  0.986   0.907    acaGCACAtctttcaatgc  GR

misc_feature    complement(649..662) /note="heat-shock element":
V$HSF1_01   651  (+)  0.974   0.956    GGAAActtcc            HSF1
V$HSF1_01   651  (-)   0.974   0.963    ggaaaCTTCC           HSF1
V$HSF2_01   651  (-)   0.997   0.986    ggaaaCTTCC           HSF2

misc_feature    complement(886..899  /note="estrogen responsive
element":
V$ER_Q6        883  (-)  1.000   0.927    ctgGGTCAgaaggcctggg  ER

misc_feature    complement(945..953) /note="acute phase-responsive
element":
V$STAT_01      945 (-)  1.000   0.984    ttctGGGAA
STATx

misc_feature    complement(1093..1098)/note="glucocorticoid binding
core":
V$GR_Q6       1099 (+)  0.878   0.847    tctggccagccTGTGGtct   GR

misc_feature    complement(1160..1172) /note="hepatocyte-specific
promoter element":
V$CEBP_01     1159 (+)  0.806   0.806    agCCTGGgaacag          C/EBP

TATA_signal     complement(1192..1197)
V$TATA_01     1191 (+)  1.000   0.976    ctATAAAtagggcct         TATA
V$MTATA_B   1189 (+)  1.000   0.916    agctATAAAtagggcct     Muscle
TATA box
```

The disadvantage of this approach is that one gets a huge number of false positive matches. To reduce this number you can restrict MatchTM to use high quality matrices only.

How can I reduce the number of false negative matches and make sure that I do not lose any "real" sites at the same time?

Cut-offs to minimize false positive matches try to filter out all possible random matches, but they have the disadvantage that also some "real sites" are also missed. "Real sites" do not naturally have the highest matrix similarity score, because the binding of a factor does not only depend on the sequence of its binding site. An optimal cut-off would find all "real" binding sites and it would filter out all random matches. But it is rather infrequent that it is possible to separate these two sets of sites so easily. Cut-offs to minimize the sum of both error rates are the best possible approximation. If you do not want to lose any real sites, you should use cut-offs to minimize false negative matches. To reduce the number of false negative matches, you can restrict your search to the use of high quality matrices only, i.e. excluding matrices with particularly high false positive rates. So, for example, matrices with a short matrix length, which therefore have a high amount of random matches, are filtered out.

But the high amount of false positive matches is in fact the general limitation of this type of analysis, when one just tries to identify all possible subsequences that might be potential transcription factor binding sites.

The analysis of the overall structure of promoters to understand the promoter context seems to be a more promising approach than just searching a promoter for single binding sites. First of all we are talking about certain combinations of TF sites that are specific for definite types of promoters. Searching for such combinations is much more specific and produces less false positives. You may want to take a look at our paper: "Kel et al., JMB (1999)288,353-376" concerning composite elements in immune responsive genes.

When you have a set of co-regulated genes, it would be best to apply a comparative analysis for over-represented sites or composite modules characteristic for your set of genes in comparison to a background set. Such comparative analysis is available in the ExPlain™ Analysis System.

How do I search a promoter DNA sequence for potential transcription factor binding sites?

To search a DNA sequence for potential transcription factor binding sites, you can use the Match tool provided with TRANSFAC® professional. Match™ compares the sequence with the nucleotide distribution matrices from TRANSFAC® professional. As a result you get a list of those matrices that matched your sequence.

Please keep in mind that the significance of single potential binding matrices for one or the other factor in a promoter or other sequence can be low, but that they need to be seen in context.

When you have a set of co-regulated genes, it would be best to apply a comparative analysis for over-represented sites or composite modules characteristic for your set of genes in comparison to a background set. Such comparative analysis is available in the ExPlain™ Analysis System.

I would like more information about the tissue-specific profiles: How were they constructed?

We have selected a number of genes described in the TRANSFAC® Gene table, which are highly inducible upon response in different cells of a certain tissue. Both human and mouse genes have been selected.

We have created a list of transcription factors (TFs) that have been experimentally shown to bind specific DNA sites in promoters of those genes and regulate their transcription. Thus, widely expressed TFs, which play an important role in the transcriptional regulation of genes within a certain tissue, are also included in the list. TRANSFAC® matrices for those TFs were selected. For some of the TFs there are several matrices in TRANSFAC® (for instance, GATA, Oct). In these cases, only the best matrix in a group was selected for the profile. Cut-offs that are given by default (for command line use) are to minimize false negative matches..

I would like more information about the tissue-specific profiles: Which matrices have you included in it?

You can easily find a list of all matrices that are used to construct each profile in the following way. The button "Create Profile" is located on the bottom of the first page of Match™. Press this button and you will find the next page. On the top of the page you can select the provided profiles for viewing. The matrices from the selected profile with their FP rates at particular cut-offs will be listed at the end of the page.

Based on the profile, the user can generate his own profile, by removing (unchecking) or adding additional matrices. After matrix selection, click "Proceed to Cut-off selection".

On the next page you can select cut-offs, choose a name for your profile and save it. When you open the Match start page, after you have saved your profile, then you will find your profile in the profile selection list under "user defined profiles".

I would like more information about the tissue-specific profiles: Is it a specific human matrix in this profile? Why don't you make a human specific profile?

For profile construction we have used vertebrate matrices from TRANSFAC® (V$*). Many of them are built on the base of human, mouse and rat DNA sites and because of this they are mammalian matrices.

DNA-binding domains of orthologous mammalian factors (for example, mouse and human E2F-1) are homologous and are able to recognize the same binding sites on DNA. Moreover, rat or mouse recombinant factors are often used to study transcriptional regulation of human genes, and vice versa. Therefore, our suggestion is that mammalian matrices are useful for searching DNA sequences of any mammalian species.

What workflow would you recommend to analyze differential gene expression data with Match™?

The Match tool on its own provides only individual analysis of (a list of) single sequences. Differential gene expression data can be analysed in the ExPlain™ Analysis System, which allows analysis for over-represented binding sites in the differentially expressed genes vs. the (unaltered) background set.