Profiles for Match™ in TRANSFAC® and geneXplain platform

Profile set	Description	Available in TRANSFAC®	Available in geneXplain platform
54 Profiles for specific p-value-thresholds	Error rates computed analytically for species-specific dinucleotide distributions of regulatory sequences. All TRANSFAC® Positional Weight Matrices (PWMs), as well as taxonomic and methodological subsets.		X
31 TRANSFAC® profiles	Error rates calculated empirically on the basis of sequence samples. All TRANSFAC® PWMs, taxonomic as well as biological function-oriented subsets.	X	X

Applying PWM profiles

PWM profiles serve two main purposes, 1. to provide a means to select a subset of weight matrices available in the database and 2. to define PWM score thresholds for analysis tools like Match™.

Lower score thresholds, like *minFN* thresholds or a 5%-*p-value* (described below), yield more binding site predictions than higher score thresholds such as *minFP* or a 0.1%-*p-value*. Certainly also, more PWMs predict more binding sites than fewer PWMs.

For the *manual* localization of binding sites it is preferable to choose a set of motifs a priori, e.g. by focusing on a certain biological context or specific transcription factors of interest. Then higher score thresholds can be applied to firstly identify higher affinity binding sites. The abundance of binding sites of individual PWMs, possibly guided by initial findings, can further be analyzed at lower cutoffs.

When a set of co-regulated sequences is available, approaches like FMatch enable prioritization of motifs on the basis of their statistical overrepresentation and an a priori selection of PWMs is less important. As these methods usually include automatic cutoff optimization, the initial score threshold defined by a profile also becomes less important and it is indeed recommendable to start out from less stringent parameters.

The *minFN* and 5%-pvalue cutoffs define lower, less stringent score thresholds whereas *minFP* and 0.1%-pvalue cutoffs result in stronger selection for high affinity binding sites. The stringency of *minSUM* and 1%-pvalue cutoffs lies between the previously described choices.

Profiles for specific p-value-thresholds (available in the geneXplain platform only)

The geneXplain platform provides profiles with selected *p-value*-thresholds which are, from least to most stringent, *0.05*, *0.01* and *0.001*. The p-values correspond to false positive rates in sequences with a nucleotide composition following one of four dinucleotide distributions. The dinucleotide distributions are estimated from genomic sequences in upstream regions of genes in human, mouse, rat and thale cress. Using calculations described in Stegmaier et al. [1], score thresholds for TRANSFAC® PWMs are adjusted so that scores in either orientation of a motif predict a putative binding site in 5%, 1% or 0.1% of potential binding site locations.

The profiles are prepared for all TRANSFAC® PWMs as well as for vertebrate and plant subsets. For each combination of PWM set, species-specific dinucleotide distribution and false positive rate there is an addition profile that excludes PWMs derived through homology modeling (non3d). Profile names are composed to reflect the choice of

1. PWM set (all, vertebrate or plant)
2. Dinucleotide distribution (human, mouse, rat, thale)
3. P-value cutoff (0.001, 0.01, 0.05)
4. if applicable, *non3d* if homology modeling-based PWMs were excluded.

E.g., the profile *all_human_p0.001.prf* contains all TRANSFAC® PWMs with score thresholds adjusted to a false positive rate of 0.1% in sequences with a dinucleotide composition like human promoter regions. The same applies to the profile *all_human_p0.001_non3d.prf*, but it does not include PWMs derived through homology modeling.

TRANSFAC® profiles (available in both TRANSFAC® and geneXplain platform)

Profiles with score thresholds defined by TRANSFAC®

TRANSFAC® has defined three types of cutoffs to address false positive and false negative errors, *minFN*, *minFP*, and *minSUM*. Error rates are estimated for each PWM on the basis of genomic sequences. The *minFN* cutoffs control false negatives at an error rate of 10%, whereas *minFP* corresponds to a false positive error rate of 0.1% and *minSUM* is adjusted to minimize sum of false positive and negative errors.

The profiles to *minimize false negative errors* (minFN), *minimize false positive errors* (minFP), or the sum (minSUM) cover all TRANSFAC® PWMs with threshold settings according to described cutoff types.

Profiles with taxonomic PWM selections

TRANSFAC® PWMs were assigned to profiles according taxonomic divisions of their binding factors. There are profiles for PWMs of transcription factors in bacteria, fungi, insects, invertebrates, nematodes, plants and vertebrates. If not further specified these profiles apply *minFN* score cutoffs.

Taxonomic profiles with recommended specific matrices

For transcription factors with several PWMs in the database, the best performing PWM was determined on the basis of genome-wide binding site data and a specialized machine learning model. The profiles with *recommended specific* PWMs, available for fungal, insect, plant and vertebrate transcription factors, provide collections with reduced redundancy compared to the full taxonomic subsets.

Non-redundant profiles for vertebrate transcription factors

Due to the large number of vertebrate matrices provided in TRANSFAC®, of which many show similar binding motifs, matrices for related factors have been grouped together and the matrices from each group were clustered based on their similarity. From each cluster then one matrix was selected as "representative" of the matrix cluster and their associated factors. These representative matrices were combined in the "vertebrate non-redundant" profile. Furthermore, profiles were defined with *minFN*, *minFP* and *minSUM* score cutoffs for the non-redundant PWM selection.

Profiles for biological functions

Cell cycle-specific profile

This profile is designed to search for potential binding sites within regulatory regions of genes whose expression is dependent on the stage of cell cycle.

Redox-specific profile

This profile is designed to search for potential binding sites within regulatory regions of genes whose expression is redox-sensitive.

Methylation-specific profile

This profile consists of matrices created by methylation-sensitive SELEX to reflect the DNA binding specificity of transcription factors that prefer CpG-methylated sequences.

Tissue-specific profiles

Adipocyte-specific profile

This profile is designed to search for potential binding sites within regulatory regions of adipocyte-specific genes.

Immune cell-specific profile

This profile is designed to search for potential binding sites within regulatory regions of genes whose transcription is induced upon immune response in T-cells, B-cells, mast cells, myeloid cells, natural killer cells, and macrophages.

Liver-specific profile

This profile is designed to search for potential binding sites within regulatory regions of liver-enriched genes.

Lung-specific profile

This profile is designed to search for potential binding sites within regulatory regions of lung-specific genes.

Muscle-specific profile

This profile is designed to search for potential binding sites within regulatory regions of muscle-specific genes.

Nerve system-specific profile

This profile is designed to search for potential binding sites within regulatory regions of nerve system-specific genes.

Pancreatic beta cell-specific profile

This profile is designed to search for potential binding sites within regulatory regions of pancreatic beta-cell-specific genes.

Pituitary-specific profile

This profile is designed to search for potential binding sites within regulatory regions of pituitary-specific genes.

Other profiles

Profile for Composite Element models (models_CE)

This profile covers PWMs included in the file models_CE.dat to be used as basis for Composite Model Search (CMSearch).

Profile for retinoic acid response elements (models_RARE)

The profile for RARE models (retinoic acid response elements and similar motifs) consists of two half-sites to be used as basis for CMSearch.

References

1. Stegmaier P, Voss N, Meier T, Kel A, Wingender E, et al. (2011) Advanced Computational Biology Methods Identify Molecular Switches for Malignancy in an EGF Mouse Model of Liver Cancer. PLOS ONE 6(3): e17738. https://doi.org/10.1371/journal.pone.0017738