Construction of a Functional Word Dictionary for Primate Promoter Sequences

Wataru Fujibuchi and Minoru Kanehisa
Institute for Chemical Research
Kyoto University
Uji, Kyoto 611
E-mail: wataru@kuicr.kyoto-u.ac.jp, kanehisa@kuicr.kyoto-u.ac.jp

Abstruct

We constructed a dictionary of sequence motifs for transcription regulation with a heuristic method from a set of DNA sequences upstream of the transcription initiation site. The method first identifies weakly conserved blocks within a given region relative to the initiation site by the search and merge of six-base patterns. Then most conserved portions of these blocks are extracted by calculating the information content after similar blocks are multiply aligned. The procedure was applied to primate promoters and the result was evaluated with the Transcription Factor Database (TFD). The result will give us new biological insights into the DNA signals.

1. Introduction

In the study of protein sequence analysis, sequence motifs are relatively well defined, compiled in libraries like Prosite[1], and utilized for biological interpretation of amino acid sequence data. In the case of nucleic acid sequence motifs, however, it still requires laborious works to collect and properly organize relevant data and to develop computational methods to define and extract biologically important sequence patterns. We are interested here in extracting transcriptional control signals from a set of eukaryotic sequences.

Thus far, certain statistical methods[2,3] were applied to DNA sequences supposed to contain transcriptional signals to detect conserved sequence patterns. None of these methods, however, was able to inform the actual length of the regulatory site. The definition of a regulatory site may better be achieved, when all sequences in a data set are known to contain regulatory sites, by studying the information content[4,5] of the DNA sequence.

Our aim is to identify regulatory sequence patterns which have optimized lengths from a set of sequences which may or may not contain regulatory sites. We have developed a heuristic method which first identifies weakly conserved blocks by the search and merge of short patterns and then extracts most conserved portions of these blocks by a combination of alignment and information content algorithms. Whether our method can reproduce biologically significant patterns is evaluated

藤渕　航, 金久　實　京都大学化学研究所, 〒611 京都府宇治市五ケ庄

by the use of the Transcription Factor Database (TFD)[6] compiled by Ghosh. We present here an example of those motifs and discuss the validity of making a fuctional word dictionary for eukaryotic promoters.

2. Materials and Methods

1) Dataset

First, sequences were selected from the EMBL nucleotide sequence database (rel. 33.0) according to Eukaryotic Promoter Database (EPD)[7] entries, and we obtained 226 primate sequences out of 1,133 entire entries. We examined up to 100 bases upstream of the transcription initiation site, which are numbered in the 5' to 3' direction from -100 to -1 where -1 is the base position immediately preceding the initiation site. For methodological simplicity, we require all sequences which are aligned at the base position -1 be 100 bases long, removing truncated portions.

Next, to avoid using vector-contaminated data, all sequences were checked whether they had vector sequences obtained from genbank.bio.net. If a sequence and any of the vectors contained an identical stretch of more than 16 bases, the sequence was removed.

Lastly, to remove duplicate entries of similar sequences, all pairwise alignments of collected sequences were performed. For each group of sequences with above 80% similarity, only one representative was retained. As the result, there were 196 sequences remained in the dataset.

2) Putative Functional Fragment and Block Mapping

In order to search conserved patterns which may play biological roles at specific locations relative to the transcription initiation site, we define a putative functional fragment (PFF) as:

an l-base pattern found with f% probability allowing up to s% substitutions.

The sequence patterns of 6 ($l = 6$) bases are searched in the 100 base long dataset. Probability f can be calculated by using the relative entropy (Waterman *et al.*, 1984)[8]. If we know beforehand the probability q for finding a pattern within one sequence, the probability of finding n occurrences of the same pattern in a set of N sequences is given by:

$$f = \exp\{ - N H(p, q) \}$$

with $p = n/N$, and H is the relative entropy where:

$$H(p, q) = p \log(p / q) + (1 - p) \log\{(1 - p) / (1 - q)\}$$

If we assume the patterns are Poisson distributed, the probability q that the pattern is found at least once in the sequence is given by:

$$q = 1 - \exp(- u)$$

with:

$$u = E(W) (L - l + 1)$$

L and l are the lengths of the sequence and the pattern, respectively, and:

$$E(W) = E(X_1X_2) E(X_2X_3) \ldots E(X_{l-1}X_l) / E(X_2) \ldots E(X_{l-1})$$

is the expected frequency of pattern W caluculated by assuming the first order Markov chain according to Stückle *et al.*[9] and Pesole *et al.*[3], where $E(X_iX_j)$ and $E(X_k)$ are respective frequencies of

dinucleotides and mononucleotides observed in all of the sequences. By taking substitutions into account, E(W') values for all substitution patterns W's of W are added to E(W).

Once relatively conserved six-base patterns of PFFs are found, the actual locations of the six-base patterns are then examined on the original sequences allowing again up to s% substitutions. If one PFF overlaps another with five bases sharing, the two PFFs are combined. We call a combined region of PFFs a block. By this process which we call block mapping, a set of sequences is converted into a set of blocks which have various lengths.

3) Local Alignment and Information Content of Blocks

Suppose one block is selected for refinement. First it is pairwise aligned with all other blocks. The local homology score of two blocks is defined by:

$$S_{ij} = \max_{i,j,k} \{ \sum_{i,j}^{i+k,j+k} e_{ij} \}, k(\geq l)$$

$$e_{ij} = \{1(\text{base } i = \text{base } j), -1(\text{otherwise})\}$$

where eij is a single base similarity score between position i of one block and position j of the other. Sij is the locally maximal homology score for the segment of length k when two blocks are aligned without any gap. If the base substitution of the k-base segment is equal or below the substitution parameter (s%) of PFF, the aligned block is retained. Thus, the block being considered is multiply aligned with a number of locally homologous blocks. This multiple alignment is converted into the base frequency matrix which has elements of four times the length of the block being considered.

In order to extract a significant portion of the block, the information content was calculated using the base frequency matrix. We adopt the definition of the information content following the work of Iijima and Kanehisa[5]. We consider the portion which satisfies the following two conditions to be significant and only this portion of the block is retained for further analysis.

(a) On both ends of the columns, the information content smoothed with 3 base window is no less than the half of the average information content in all columns.

(b) The information content on both ends of the columns is more than the threshold I_{thres}.

If the selected columns form a segment of more than 2 bases, it is retained as an extracted block for later use.

During the procedure all blocks are sorted and grouped according to the length. Starting with the longest one, each block among the same length group is in turn taken for multiple alignment with all the similar blocks and the resulting base frequency matrix is used to optimize the block length by calculating the information content. The shortened block now belongs to another group of a shorter length and it is later utilized when blocks of that length are examined. To avoid order dependency, all base frequency matrices and information contents are calculated among the same length group before performing the shortening of blocks.

3. Result and Discussion

1) Determination of Parameters

There are three parameters in our method which need be considered first: the probability f of observing sequences containing a given pattern, the pattern substitution rate s, and the threshold

information content I_{thres}. We examined various values of f and s, and blocks were extracted and their lengths were optimized by various values of the information content threshold I_{thres}. To estimate the parameters we prepared a reference set containing those TFD consensus patterns which appear in the original dataset of promoter sequences. We first chose the I_{thres} value in order for the number of perfect matches with the TFD consensus patterns to take the highest value for each set of f and s. Using these parameter sets, extracted blocks were then examined for matching TFD patterns allowing this time the length difference of up to two bases. The sum of the number of matched blocks and the number of TFD patterns corresponded took the highest value when: I_{thres} = 0.1bit, f = 1% and s = 10%. This is the parameter set used in the present study.

2) Characterization of Blocks by TFD

With the parameter set determined above, each block was compared with all TFD patterns allowing their length differences to some extent. The result is shown in Table I, which contains the names of specific transcription factors that matched with any blocks, as well as the numbers of corresponding blocks. Note that a block may correspond to multiple TFD entries. Especially, the entries which contain perfect matches, i.e., matching blocks with exact lengths, are shown in bold-case characters. The numbers of the total extracted blocks and the classified (TFD corresponding) blocks are 162 and 119, respectively, thus the match rate is 73.5%. Most of the unclassifiable blocks seem to appear less frequently; some blocks, howerver, are considerably frequent. For instance, we obtained nineteen 'AAAGCA' and fifteen 'GAAAGT' blocks, but neither could be classified by any TFD entries (see Appendix). These unclassifiable blocks might require further studies on biological roles.

Table I. TFD entry names and the numbers of matching blocks.

AP-2 32, **GCF** 28, **Sp1** 23, **(Sp1)** 19, **TFIID** 17, B-factor 13, **LSF** 12, **BGP1** 12, **(TFIID/TBF)** 10, T-Ag 6, **AP-1** 5, TCF-1 4, LVc 4, GAL4 4, engrailed 3, PU.1 3, **NF-S** 3, **IBP-1** 3, CREB 3, (SRF) 3, Y protein 2, PR 2, EivF/CREB 2, EivF 2, E4TF1 2, E4F1 2, **CTF** 2, CTCF 2, CDF1 2, (TFIID/TBP) 2, (GAL4) 2, b2 1, URSF 1, **TGA1b** 1, TGA1a 1, TCF-2-alpha 1, SRF 1, **Pu.1** 1, Pit-1 1, NFI 1, NF-Y* 1, NF-Y 1, HAP2/3 1, H4TF1 1, H-2RIIBP 1, **GR** 1, GATA-1 1, Ets-1 1, ETFA 1, ER 1, EPBF 1, EF-1A 1, **CTF/CBP** 1, CRF 1, CDF 1, CCAAT-bf 1, CBP 1, CBF 1, C/EBP 1, **ATF/CREB** 1, **ATF** 1, **Unknown** or Undefined 79, Unclassifiable 43

The entries which contain perfectly matching blocks are shown in bold-case.

3) Block Length

The distribution of block lengths is shown in Table II. In general 6, 7 and 8 base blocks were most frequent. We did not find 5 base or shorter blocks from the dataset, while more than 12 base blocks also could not be seen. The lack of shorter and longer blocks is the result of statistical infrequency, at least according to our method. It remains to be examined whether, for example, signal blocks constituting of well-conserved bases do not require more than 12 bases because of biological or structural constraints.

Table II. The distribution of block lengths.

Length	6	7	8	9	10	11	12
Blocks	73	46	29	7	5	1	1

4) Motif Dictionary

Toward making a complete dictionary of functional words for interpreting DNA sequence data, we have compiled our collection of blocks in a form shown in the Appendix. Each word contains the observed frequency of the blocks and similar blocks, the organism code, TFD entry names and patterns corresponded. The letter preceding the organism code denotes whether its entry contains the consensus (C) or individual (I) pattern according to the TFD format. The TFD patterns of perfect matches are shown in capital letters. We hope that the interpretation of DNA transcriptional signals will be facilitated by refining and making use of our dicionary.

Acknowledgement

This work was supported by the grant-in-aid for scientific research on the priority area 'Genome Informatics' from the Ministry of Education, Science and Culture, Japan. The computation time was provided by the Supercomputer Laboratory, the Institute for Chemical Research, Kyoto University.

References

[1] A. Bairoch, *Nucl. Acids Res. Suppl.*, **20**, 2013(1991).
[2] D.J. Galas, M. Eggert, and M.S. Waterman, *J. Mol. Biol.*, **186**, 117(1985).
[3] G. Pesole, N. Prunella, S. Liuni, M. Attimonelli, and C. Saccone, *Nucl. Acids Res.*, **20**, 2871(1992).
[4] G.D. Stormo, and G.W. Hartzell, III, *Proc. Natl. Acad. Sci. USA*, **86**, 1183(1989).
[5] T. Iijima, and M. Kanehisa, *Bull. Inst. Chem. Res., Kyoto Univ.*, **69**, 226(1991).
[6] D. Ghosh, *Nucl. Acids Res.*, **18**, 1749(1990).
[7] P.Bucher & E.N.Trifonov, *Nucl. Acids Res.*, **14**, 10009(1986).
[8] M.S.Waterman, R.Arratia and D.J.Galas, *Bull. Math. Biol.* **46**, 515(1984).
[9] E.E.Stückle, C.Emmrich, U.Grob and P.J.Nielsen, *Nucl. Acids Res.*, **18**, 6641(1990).

Appendix

The functional word dictionary for primate promoter sequences.

```
format:
Word [# of observed blocks(substituted)]: <C/I + organism code> TFD entry name {pattern, ...}...

  GGCCGG[29(0)]: <CMAMM> AP-2{cccmnsss} GCF{scgsssc} <IMAMM> unknown{ggccg}
  CAGAGG[22(0)]: <IMAMM> unknown{agagg}
  CCCGCCC[21(0)]: <CMAMM> AP-2{cccmnsss} Sp1{gggcgg} <CMAMM/AVIAN> Sp1{krggckrrk} <CMAMM>
        unknown{gggnggrr, yyccgccc} <IAVIAN> BGP1{gggcgg} unknown{ggggcggg} <IMAMM>
        (Sp1){ccccgccc, gggcgg} LSF{ccgccc} Sp1{cccgcc, ccgccc, ggcggg, gggcgg, ggggcggg}
        unknown{ggcggg}
  CAGCCT[20(0)]: <CEUK> unknown{CANYYY}
```

```
GGGCAG[20(0)]: <CMAMM> AP-2{gsswgscc}
GCCGCG[19(0)]: <CMAMM> GCF{scgsssc} <IMAMM> unknown{cgccgcg}
AAAGCA[19(0)]:
GCCTCC[19(0)]: <CMAMM> AP-2{gsswgscc} unknown{cwkkanny} <IMAMM> T-Ag{gaggc}
CCCTCC[17(0)]: <CMAMM> AP-2{cccmnsss} unknown{gggnggrr} <IAVIAN> CTCF{ccctc}
GAAAGT[15(0)]:
GATTGG[15(0)]: <CMAMM> GATA-1{mywatcwy} <CMULTIPLE> unknown{ccaatna} <IAVIAN> unknown{ccaat}
      <IECHINO> CBF{attgg} CDF{attgg} <IMAMM> CBF{ccaat} CBP{ccaat} CCAAT-bf{ccaat} CRF{attgg}
      CTF{GATTGG} CTF/CBP{GATTGG} CTF{attgg, ccaat} EPBF{ccaat} GATA-1{ccaatct} NF-Y{attgg}
      NF-Y*{ccaat} NFI{gccaatc} Y protein{ctgattgg} unknown{ccaat} <IYEAST> HAP2/3{tgattggt}
CGGGCG[14(0)]:
TATAAAT[14(0)]: <CMAMM> TFIID{TATAWAW} <IDROS> B-factor{tataaata} <IECHINO>
      (TFIID/TBP){tataaatag} <IEUK> TFIID{tataaa} <IMAMM> (TFIID/TBF){tataaa} unknown{tataaa}
      <IYEAST> TFIID{tataaa}
GGCGCG[12(0)]: <CMAMM> GCF{scgsssc}
GCAGGG[12(0)]: <CMAMM> AP-2{cccmnsss} <IMAMM> LVc{cctgc}
AGGACA[12(0)]: <CMAMM> unknown{cwkkanny} <IMAMM> GR{TGTCCT} PR{tgtcctct}
CCCTCCC[11(0)]: <CMAMM> AP-2{cccmnsss} unknown{gggnggrr} <IAVIAN> CTCF{ccctc} <IMAMM>
      H4TF1{gggggaggg}
CCTCCC[11(0)]: <CMAMM> AP-2{cccmnsss, gsswgscc} unknown{gggnggrr} <IPLANT> unknown{cacctccc}
CCCGCC[11(0)]: <CMAMM> AP-2{cccmnsss} unknown{gggnggrr, yyccgccc} <IAVIAN> unknown{ggggcggg}
      <IMAMM> (Sp1){ccccgccc} Sp1{CCCGCC, GGCGGG, ggggcggg} unknown{GGCGGG, gccccgcc}
TGACCA[11(0)]: <IAVIAN> ER{ggtca}
CCTGAC[11(0)]: <CMAMM> unknown{ryywsgtg} <IMAMM> H-2RIIBP{ggtcaggg}
ACAGAG[10(0)]:
CTGATT[10(0)]: <CMAMM> unknown{cctgawwa} <IMAMM> Y protein{ctgattgg} unknown{ctgatta}
AAAACAG[10(0)]: <CMAMM> TCF-1{mamag}
CCGCCC[10(0)]: <CMAMM> AP-2{cccmnsss} GCF{scgsssc} Sp1{GGGCGG} unknown{gggnggrr, grgsggtg,
      yyccgccc} <IAVIAN> BGP1{GGGCGG} unknown{ggggcggg} <IMAMM> (Sp1){GGGCGG, ccccgccc,
      gggggcgg} LSF{CCGCCC} Sp1{CCGCCC, GGGCGG, gggcggag, ggggcggg}
GTCAGC[10(0)]: <CMAMM> NF-S{ygtcagc}
CGCGAC[9(0)]:
GGCAGGG[9(0)]: <CMAMM> AP-2{cccmnsss} <IMAMM> LVc{cctgc}
CCTCTC[9(0)]: <CMAMM> AP-1{gagagga} <IMAMM> unknown{agagg}
TATTTA[9(0)]: <CPLANT> CDF1{ctaaatac} <IDROS> B-factor{tataaata} <IMAMM> unknown{swatwwag}
GGGCAGGG[9(0)]:
AAGTGA[9(0)]: <CMAMM> IBP-1{AAGTGA} unknown{AARKGA} <IMAMM> unknown{AAGTGA}
GGGGCGGGG[9(0)]: <CMAMM> AP-2{cccmnsss} GCF{scgsssc} <CMAMM/AVIAN> Sp1{KRGGCKRRK} <CMAMM>
      Sp1{kgggcggrry, krggcgkrry} unknown{gggnggrr, yyccgccc} <IAVIAN> unknown{ggggcggg}
      <IMAMM> (Sp1){ccccgccc} Sp1{ccccgcccc, ggggcgggc, ggggcggggg, ggggcggg, ggggcgggt}
TTAATA[8(0)]: <IMAMM> unknown{swatwwag, tattaaaa}
GGCGGG[8(0)]: <CMAMM> AP-2{cccmnsss} unknown{gggnggrr, yyccgccc} <IAVIAN> unknown{ggggcggg}
      <IMAMM> (Sp1){ccccgccc} Sp1{CCCGCC, GGCGGG, ggggcggg} unknown{GGCGGG, gccccgcc}
GGCGCGG[8(0)]: <CMAMM> GCF{SCGSSSC}
GTATAAAAG[8(0)]: <CMAMM> TFIID{tatawaw} <IDROS> B-factor{tataaaa} <IMAMM> TFIID{tataaaa}
CGCGGC[8(0)]: <CMAMM> GCF{scgsssc} <IMAMM> unknown{cgccgcg}
CGCGCG[8(0)]:
TGTCAG[8(0)]: <CMAMM> NF-S{ygtcagc}
AAACAC[8(0)]: <CMULTIPLE> unknown{aaacaca}
ATTATT[8(0)]: <CMAMM> unknown{cwkkanny}
TGTTAT[7(0)]:
GGGCAGG[7(0)]: <CMAMM> AP-2{gsswgscc} <IMAMM> LVc{cctgc}
TTTAAG[7(0)]: <CMAMM> unknown{cwkkanny}
CATAAAAG[7(0)]: <CMAMM> (SRF){ccwwwwwwgg} <IMAMM> SRF{ccttttatgg}
AAATAG[7(0)]: <IMAMM> unknown{swatwwag}
GGGGCGGG[7(0)]: <CMAMM> GCF{scgsssc} Sp1{gggcgg, kgggcggrry, krggcgkrry} <CMAMM/AVIAN>
      Sp1{krggckrrk} <IAVIAN> BGP1{gggcgg} unknown{GGGGCGGG} <IMAMM> (Sp1){gggcgg} LSF{ccgccc}
      Sp1{GGGGCGGG, ccccgcccc, cccgcc, ccgccc, ggcggg, gggcgg, ggggcgggac, ggggcgggat,
      ggggcggggc, ggggcggggg, ggggcggggt, gggggcgggc} unknown{ggcggg}
TCCTCT[7(0)]: <CMAMM> AP-1{gagagga} <IMAMM> PR{tgtcctct} unknown{agagg} <IYEAST> GAL4{tcctc}
TATAAAA[7(0)]: <CDROS> engrailed{hcwathaaa} <CMAMM> TFIID{TATAWAW} <IDROS> B-factor{TATAAAA}
      <IEUK> TFIID{tataaa} <IMAMM> (TFIID/TBF){tataaa} TFIID{TATAAAA} unknown{tataaa} <IYEAST>
      TFIID{tataaa} unknown{tttttata}
TTCCTC[7(0)]: <CMAMM/AVIAN> Ets-1{smggawgy} <CMAMM> TCF-2-alpha{saggaagy} unknown{gggnggrr}
      <IMAMM> PU.1{GAGGAA} <IYEAST> GAL4{tcctc}
GCCTCCC[7(0)]: <CMAMM> AP-2{gsswgscc} <IMAMM> T-Ag{gaggc}
ACGTCA[6(0)]: <CMAMM> ATF/CREB{ACGTCA} ATF{TGACGT, tgacgymr} CREB{cgtca, kwcgtca, tgacgtyw}
      E4F1{acgtmac} ETFA{tgacgtrr, yyacgtca} EivF{gtkacgt} EivF/CREB{gtkacgw}
      unknown{cwkkanny} <IMAMM> ATF/CREB{tacgtcat} CREB{tgacgtca, tgacgtct, tgacgtc, tgacgtgt}
      E4F1{acgtcac, acgtcag, acgtcat} E4TF1{rtgacgt} b2{tgacgtca} unknown{TGACGT, gtgacgt}
      <IPLANT> TGA1a{tgacgtaa} TGA1b{TGACGT}
TATAAAG[11(0)]: <CMAMM> unknown{cwkkanny} <IEUK> TFIID{tataaa} <IMAMM> (TFIID/TBF){tataaa}
      unknown{swatwwag, tataaa} <IYEAST> TFIID{tataaa}
TTTAAGA[6(0)]:
```

```
AAACAG[6(0)]: <CMAMM> TCF-1{mamag}
GGGCGG[6(0)]: <CMAMM> AP-2{cccmnsss} GCF{scgsssc} Sp1{GGGCGG} unknown{gggnggrr, grgsggtg,
      yyccgccc} <IAVIAN> BGP1{GGGCGG} unknown{ggggcggg} <IMAMM> (Sp1){GGGCGG, ccccgccc,
      gggggcgg} LSF{CCGCCC} Sp1{CCGCCC, GGGCGG, gggcggag, ggggcggg}
GGCAGG[6(0)]: <CMAMM> AP-2{cccmnsss, gsswgscc} <IMAMM> LVc{cctgc}
GCGGGG[6(0)]: <CMAMM> AP-2{cccmnsss} GCF{scgsssc} unknown{gggnggrr, yyccgccc} <IMAMM>
      (Sp1){ccccgccc} unknown{gccccgcc}
TATATAAG[6(0)]: <CMAMM> TFIID{tatawaw} <IYEAST> unknown{atataa}
GGCGGC[6(0)]:
CGGGGC[6(0)]: <CMAMM> GCF{scgsssc} <IMAMM> T-Ag{ggggc} unknown{gccccgcc}
TATATA[6(0)]: <CMAMM> TFIID{tatawaw} <IMAMM> unknown{swatwwag}
GCGCGG[6(0)]: <CMAMM> AP-2{cccmnsss} GCF{scgsssc}
TTAAGA[5(0)]: <IMAMM> C/EBP{tcttaagc}
TAATAT[5(0)]:
CCTGACT[5(0)]: <CYEAST> unknown{tgact}
TAAATA[5(0)]: <CPLANT> CDF1{ctaaatac} <IDROS> B-factor{tataaata} <IMAMM> unknown{swatwwag}
CTGACT[5(0)]: <CMAMM> AP-1{stgactma} <CYEAST> unknown{tgact} <IMAMM> AP-1{ctgactcg, ctgactgg,
      tgagtcag, ttagtcag}
GGGGCG[5(0)]: <CMAMM> GCF{scgsssc} <IAVIAN> unknown{ggggcggg} <IMAMM> (Sp1){gggggcgg}
      Sp1{ggggcggg} T-Ag{ggggc}
TAAAAG[5(0)]: <CMAMM> TCF-1{mamag}
GCGGCG[5(0)]: <CMAMM> GCF{scgsssc} <IECHINO> undefined{gcggcgca} <IMAMM> unknown{cgccgcg}
TTCCTCTC[4(0)]: <CMAMM> AP-1{gagagga} <IMAMM> PU.1{gaggaa}
TGTCACA[4(0)]:
ATAAAT[4(0)]: <CMAMM> TFIID{tatawaw} <IDROS> B-factor{tataaata} <IMAMM> unknown{aataaat}
TATAAAAG[5(0)]: <CMAMM> TFIID{tatawaw} <IDROS> B-factor{tataaaa} <IEUK> TFIID{tataaa} <IMAMM>
      (TFIID/TBF){tataaa} TFIID{tataaaa} unknown{tataaa} <IYEAST> TFIID{tataaa}
GTCACA[4(0)]:
AAGTGAC[4(0)]: <CMAMM> IBP-1{aagtga} unknown{aarkga} <CYEAST> unknown{awgtgactc} <IMAMM>
      unknown{aagtga}
ATAAAA[4(0)]: <CMAMM> TFIID{tatawaw} <IDROS> B-factor{tataaaa} <IMAMM> TFIID{tataaaa} <IYEAST>
      unknown{tttttata}
GGGCGGGG[4(0)]: <CMAMM> AP-2{CCCMNSSS, ycsccmnsss} Sp1{gggcgg, kgggcggrry, krggcgkrry}
      <CMAMM/AVIAN> Sp1{krggckrrk} <CMAMM> unknown{GGGNGGRR, YYCCGCCC} <IAVIAN> BGP1{gggcgg}
      <IMAMM> (Sp1){CCCCGCCC, gggcgg} CTF{accccgccca} LSF{ccgccc} Sp1{accccgccca, cccccgcccc,
      cccgcc, ccgccc, gccccgcccg, ggcggg, gggcgg, ggggcggggc, ggggcggggg, ggggcggggt,
      tgggcggggc, tgggcggggt} unknown{ccccgcccag, ggcggg}
TGTCAC[4(0)]:
TCCTCTC[4(0)]: <CMAMM> AP-1{GAGAGGA} unknown{ancctctcy} <IMAMM> unknown{agagg} <IYEAST>
      GAL4{tcctc}
CTATAAAA[8(0)]: <CDROS> engrailed{hcwathaaa} <CMAMM> TFIID{tatawaw} <IDROS> B-factor{tataaaa}
      <IEUK> TFIID{tataaa} <IMAMM> (TFIID/TBF){tataaa} TFIID{tataaaa} unknown{tataaa} <IYEAST>
      TFIID{tataaa}
AAAACA[4(0)]:
GTGACG[4(0)]: <CMAMM> CREB{cgtca} E4F1{acgtmac} EivF{gtkacgt} EivF/CREB{gtkacgw} <IMAMM>
      E4F1{acgtcac, tcgtcac} E4TF1{rtgacgt} unknown{gtgacgt}
ATAAAAG[4(0)]: <CMAMM> TCF-1{mamag}
GCGGCGGG[3(0)]: <CMAMM> GCF{scgsssc} <IMAMM> Sp1{cccgcc, ggcggg} unknown{ggcggg}
GGCGGGGC[3(0)]: <CMAMM> AP-2{ycsccmnsss} GCF{scgsssc} Sp1{kgggcggrry, krggcgkrry} <IMAMM>
      Sp1{cccgcc, gccccgcccg, ggcggg, ggggcggggc, tgggcggggc} unknown{GCCCCGCC, ggcggg}
GCGGGGCG[3(0)]: <CMAMM> AP-2{ycsccmnsss} GCF{scgsssc}
CATAAAG[3(0)]: <CMAMM> unknown{cwkkanny}
TTATAA[3(0)]:
ATAAAG[3(0)]: <CMAMM> unknown{cwkkanny} <IMAMM> unknown{swatwwag}
ATATTT[3(0)]:
TTATAAG[3(0)]:
GATATA[3(0)]:
AATATT[3(0)]:
TTTATA[3(0)]: <CMAMM> TFIID{tatawaw} unknown{cwkkanny} <IDROS> B-factor{tataaaa, tataaata}
      <IEUK> TFIID{TATAAA} <IMAMM> (TFIID/TBF){TATAAA} TFIID{tataaaa} unknown{TATAAA,
      swatwwag} <IPROK> unknown{tttatatg} <IYEAST> TFIID{TATAAA} unknown{tttttata}
TTCCTCT[3(0)]: <CMAMM> EF-1A{rnmggawgt} <IAVIAN> Pu.1{agaggaact} <IMAMM> PU.1{gaggaa}
      unknown{agagg} <IYEAST> GAL4{tcctc}
CGCGGCG[3(0)]: <IMAMM> unknown{CGCCGCG}
TGTCAGC[3(0)]: <CMAMM> NF-S{YGTCAGC}
CGGCGGG[3(0)]: <CMAMM> AP-2{cccmnsss} <IMAMM> Sp1{cccgcc, ggcggg} unknown{ggcggg}
GGGCGGG[3(0)]: <CMAMM> AP-2{cccmnsss} Sp1{gggcgg} <CMAMM/AVIAN> Sp1{krggckrrk} <CMAMM>
      unknown{gggnggrr, yyccgccc} <IAVIAN> BGP1{gggcgg} unknown{ggggcggg} <IMAMM>
      (Sp1){ccccgccc, gggcgg} LSF{ccgccc} Sp1{cccgcc, ccgccc, ggcggg, gggcgg, ggggcggg}
      unknown{ggcggg}
GGGGCGG[3(0)]: <CMAMM> GCF{SCGSSSC} Sp1{gggcgg} <CMAMM/AVIAN> Sp1{krggckrrk} <IAVIAN>
      BGP1{gggcgg} unknown{ggggcggg} <IMAMM> (Sp1){gggcgg, gggggcgg} LSF{ccgccc} Sp1{ccgccc,
      gggcgg, ggggcggg} T-Ag{ggggc}
AGTGAC[3(0)]:
```

```
AAGTGACG[3(0)]: <CMAMM> IBP-1{aagtga} unknown{aarkga} <IMAMM> unknown{aagtga}
ATATAAG[2(0)]: <IYEAST> (GAL4){acttatat} unknown{atataa}
GGGGCGGC[2(0)]: <CMAMM> GCF{scgsssc} Sp1{gggcgg} <IAVIAN> BGP1{gggcgg} <IMAMM> (Sp1){gggcgg}
      LSF{ccgccc} Sp1{ccgccc, cggggcggcg, gggcgg, ggggcggcgc}
GATATAAG[2(0)]: <IYEAST> unknown{atataa}
CTATAAAG[7(0)]: <CMAMM> (SRF){ccwwwwwwgg} <IEUK> TFIID{tataaa} <IMAMM> (TFIID/TBF){tataaa}
      unknown{SWATWWAG, tataaa} <IYEAST> TFIID{tataaa}
CTATAAG[4(0)]:
AATATTT[2(0)]:
GCGGGGC[2(0)]: <CMAMM> GCF{SCGSSSC} <IMAMM> T-Ag{ggggc} unknown{gccccgcc}
CATAAAA[2(0)]:
GGCGGGG[2(0)]: <CMAMM> AP-2{cccmnsss} <CMAMM/AVIAN> Sp1{krggckrrk} <CMAMM> unknown{gggnggrr,
      yyccgccc} <IMAMM> (Sp1){ccccgccc} Sp1{cccgcc, ggcggg} unknown{gccccgcc, ggcggg}
CCTATAT[2(0)]:
CATAAAT[2(0)]:
CTATAAA[2(0)]: <CDROS> engrailed{hcwathaaa} <IEUK> TFIID{tataaa} <IMAMM> (TFIID/TBF){tataaa}
      unknown{swatwwag, tataaa} <IYEAST> TFIID{tataaa}
CTATAAAAG[2(0)]: <CMAMM> TFIID{tatawaw} <IDROS> B-factor{tataaaa} <IMAMM> TFIID{tataaaa}
CGGCGGC[2(0)]: <IYEAST> URSF{tcggcggca}
CGCGCGG[2(0)]:
CGGCGGGG[2(0)]: <CMAMM> AP-2{CCCMNSSS, ycsccmnsss} <IMAMM> Sp1{cccgcc, ggcggg} unknown{ggcggg}
CGGGCGGC[2(0)]: <CMAMM> Sp1{gggcgg} <IAVIAN> BGP1{gggcgg} <IMAMM> (Sp1){gggcgg} LSF{ccgccc}
      Sp1{ccgccc, gggcgg}
TCTATATA[2(0)]:
GGCGCGGC[2(0)]: <CMAMM> GCF{scgsssc}
CGGGCGGCG[2(0)]:
ATATTTA[1(0)]:
CTATAA[2(0)]: <IMAMM> unknown{swatwwag}
TTTATAA[1(0)]: <IEUK> TFIID{tataaa} <IMAMM> (TFIID/TBF){tataaa} unknown{tataaa} <IYEAST>
      TFIID{tataaa}
TATAAA[2(0)]: <CMAMM> TFIID{tatawaw} unknown{cwkkanny} <IDROS> B-factor{tataaaa, tataaata}
      <IEUK> TFIID{TATAAA} <IMAMM> (TFIID/TBF){TATAAA} TFIID{tataaaa} unknown{TATAAA,
      swatwwag} <IPROK> unknown{tttatatg} <IYEAST> TFIID{TATAAA} unknown{tttttata}
GTATAAG[2(0)]:
TTAATATT[1(0)]:
GTATAT[1(0)]: <CMAMM> unknown{cwkkanny} <IMAMM> unknown{swatwwag}
CGGCGG[1(0)]: <CMAMM> AP-2{cccmnsss} GCF{scgsssc}
CCTATATA[1(0)]: <CMAMM> (SRF){ccwwwwwwgg}
TATATAA[1(0)]: <CMAMM> TFIID{TATAWAW} <IYEAST> (GAL4){ttatatatc} unknown{atataa}
CGGGCGG[1(0)]: <CMAMM> Sp1{gggcgg} <IAVIAN> BGP1{gggcgg} <IMAMM> (Sp1){gggcgg} LSF{ccgccc}
      Sp1{ccgccc, gggcgg}
GGTATA[1(0)]:
TGTATATA[1(0)]: <CMAMM> Pit-1{awwtatncat}
TAATATTT[1(0)]:
ACTATAT[1(0)]:
GGTATATA[1(0)]:
AAAACAC[1(0)]:
CATAAATA[1(0)]:
ATAAATA[1(0)]: <IDROS> B-factor{tataaata} <IECHINO> (TFIID/TBP){tataaatag}
AATATTTA[1(0)]:
CGCGCGGC[1(0)]: <CMAMM> GCF{scgsssc}
AGTGACG[1(0)]: <CMAMM> CREB{cgtca}
CGGGCGGG[1(0)]: <CMAMM> Sp1{gggcgg} <IAVIAN> BGP1{gggcgg} <IMAMM> (Sp1){gggcgg} LSF{ccgccc}
      Sp1{cccgcc, ccgccc, gccccgcccg, ggcggg, gggcgg} unknown{ggcggg}
GGGCGGC[1(0)]: <CMAMM> Sp1{gggcgg} <IAVIAN> BGP1{gggcgg} <IMAMM> (Sp1){gggcgg} LSF{ccgccc}
      Sp1{ccgccc, gggcgg}
GCGGCGGC[1(0)]: <CMAMM> GCF{scgsssc}
GGCGCGGCG[1(0)]: <CMAMM> GCF{scgsssc} <IMAMM> unknown{cgccgcg}
GGGGCGGCG[1(0)]: <CMAMM> GCF{scgsssc} <IMAMM> Sp1{cggggcggcg, ggggcggcgc}
CGGCGGGGC[1(0)]: <CMAMM> AP-2{cccmnsss} GCF{scgsssc} <IMAMM> unknown{gccccgcc}
GCGCGGC[1(0)]: <CMAMM> GCF{SCGSSSC}
GCGCGGCGGC[1(1)]:
CGGGCGGGGC[1(0)]: <CMAMM> AP-2{cccmnsss} unknown{gggnggrr, yyccgccc} <IMAMM> (Sp1){ccccgccc}
      Sp1{GCCCCGCCCG} unknown{gccccgcc}
CCTATA[1(0)]:
GCGGGGCGGC[1(0)]:
CGCGGCGGCG[1(0)]:
GCGGCGGGGC[1(0)]: <CMAMM> AP-2{cccmnsss} <IMAMM> unknown{gccccgcc}
GGCGGGGCGGG[1(0)]: <CMAMM> AP-2{ycsccmnsss}
GCGGCGGGGCGG[1(0)]: <CMAMM> AP-2{ycsccmnsss}
```