# Functional classification of transcription factor binding sites: Information content as a metric

D. Ashok Reddy[1], B. V. L. S. Prasad[2] and Chanchal K. Mitra[1*]

[1] Department of Biochemistry, University of Hyderabad, Hyderabad- 500046, India.

[2] Helix Genomics Pvt. Ltd, Habsiguda, Uppal, Hyderabad-500007, India.

### Summary

The information content (relative entropy) of transcription factor binding sites (TFBS) is used to classify the transcription factors (TFs). The TF classes are clustered based on the TFBS clustering using information content. Any TF belonging to the TF class cluster has a chance of binding to any TFBS of the clustered group. Thus, out of the 41 TFBS (in humans), perhaps only 5 -10 TFs may be actually needed and in case of mouse instead of 13 TFs, we may have actually 5 or so TFs. The JASPAR database of TFBS are used in this study. The experimental data on TFs of specific gene expression from TRRD database is also coinciding with our computational results. This gives us a new way to look at the protein classification- not based on their structure or function but by the nature of their TFBS.

## 1 Introduction

The human and mouse genome projects [1,2] revealed that in eukaryotes the coding region is very less than expected before. Human genome contains approximately 30,000 genes that represent less than 2% of the whole genome. Unlike in most prokaryotic genomes that contain packed gene units with few intergenic regions, repeated and non-coding sequences that do not code for proteins make up the remaining part of the human genome. Gene expression and its regulation involve the binding of many regulatory transcription factors (TFs) to specific DNA elements called Transcription Factor Binding Sites (TFBS). The region 200-300 bp immediately upstream of the core promoter is the proximal promoter that has abundant of TFBS. Further upstream is the distal promoter region that usually contains enhancers and few TFBS. TFBS are represented by relatively short (5-10 bp) nucleotide sequences. Specificity of TF is defined by its interaction with TFBS and it is extremely selective, mediated by non-covalent interactions between appropriately arranged structural motifs of the TF and exposed surfaces of the DNA bases and backbone [3]. The ability of the cell to control the expression of genes under different developmental and environmental conditions is still poorly understood. Identifying functional TFBS is a difficult task because most TFBS are short, degenerate sequences occurring frequently in the genome. The non-coding sequences play a crucial role in gene regulation hence the computational identification and characterization of these regions is very important.

Substitution matrices are widely used to score biological sequence similarity and in database search tools like BLAST [4] and FASTA [5]. The elements of these substitution matrices are explicitly calculated from observed frequencies of aligned nucleotides and expected frequencies of the nucleotides. The information in these matrices depends on the quantification approach like evolutionary models, structural properties and chemical

[*] Corresponding author: c_mitra@yahoo.com

properties of aligned sequences [6-10]. The PAM- Point Accepted Mutation [11,12] matrices are based on alignments of closely related sequences and by using these PAM matrices one can estimate observed frequencies to any desired evolutionary distance by extrapolation. In BLOSUM- BLOcks SUbstitution Matrices [13] the observed frequencies are estimated by using the ungapped segments of multiple sequence alignments of protein families and avoids extrapolation to different evolutionary distances. We have developed conventional substitution matrices for the analysis of TFBS and the developed substitution matrices are optimally suitable for TFBS only. The main focus of our study is to find the functional classification of TFBS in human and mouse with the help of average mutual information content, which is calculated by using neighbor-independent (4×4 matrices) and neighbor-dependent (16×16 matrices) nucleotide substitutions. Neighbor-independent and neighbor-dependent substitution matrices have been used to describe the non-coding sequences [14-16] like core promoter region [17] and TFBS [18-21]. TFs have been structurally classified based on sequence features of TFBS [22] and it has been also shown that a pair of TFs may have a co-localized TFBS [23].

Although non-coding DNA constitutes majority of most eukaryotic genomes, relatively little is known about its function or the nature of its functional classification. Here we characterize the functional classification of human and mouse TFBS with the help of information content. Characterizing the pattern of clustering within TFBS allows us to explore the nature of their functional classification.

# 2 Materials and methods

## TFBS data sets

The sequences in JASPAR database [24] are annotated and experimentally demonstrated TFBS profiles for multicellular eukaryotes. It is an open-access TF binding profile database that contains over a hundred TFBS profiles of *Drosophila melanogaster, Arabidopsis thaliana, Zea mays, Homo sapiens, Mus musculus etc.* We have downloaded (Table 1) and studied only the TFBS for human (Table 2) and mouse (Table 3). The sequences in this database are organized in a FASTA format and also contain the frequencies of the four bases for the selected positions.

**Table 1: Database and the corresponding organism with number of TFBS used in the present study**

| S.No | Database | Organism | No of TFBS |
|---|---|---|---|
| 1 | JASPAR | Human | 41 |
| 2 | JASPAR | Mouse | 13 |

Each TFBS is a collection of binding sites with already aligned sequences of different lengths. The lengths vary between 5-20 nucleotides and have been used without modifications. This implies that the longer sequences have better recognition properties and is expected to have a lower noise threshold (and vice-versa).

**Table 2: Human transcription factors with the recognized TFBS and their lengths**

| S.No | Name of TF | Class of TF | Total of TFBS | Length of TFBS |
|---|---|---|---|---|
| 1 | Elk-1 | ETS | 28 | 10 |

| | | | | |
|---|---|---|---|---|
| 2 | NRF-2 | ETS | 7 | 10 |
| 3 | SAP-1 | ETS | 20 | 9 |
| 4 | SPI-1 | ETS | 57 | 6 |
| 5 | SPI-B | ETS | 49 | 7 |
| 6 | FREAC-4 | FORKHEAD | 20 | 8 |
| 7 | SOX-9 | HMG | 76 | 9 |
| 8 | SRY | HMG | 28 | 9 |
| 9 | Pbx | HOMEO | 18 | 12 |
| 10 | MEF2 | MADS | 58 | 10 |
| 11 | SRF | MADS | 46 | 12 |
| 12 | COUP-TF | NUCLEAR RECEPTOR | 13 | 14 |
| 13 | PPARgamma-RXRal | NUCLEAR RECEPTOR | 41 | 20 |
| 14 | PPARgamma | NUCLEAR RECEPTOR | 28 | 20 |
| 15 | RORalfa-1 | NUCLEAR RECEPTOR | 25 | 10 |
| 16 | RORalfa-2 | NUCLEAR RECEPTOR | 36 | 14 |
| 17 | RXR-VDR | NUCLEAR RECEPTOR | 10 | 15 |
| 18 | p53 | P53 | 17 | 20 |
| 19 | Pax6 | PAIRED | 43 | 14 |
| 20 | c-REL | REL | 17 | 10 |
| 21 | p50 | REL | 18 | 11 |
| 22 | p65 | REL | 18 | 10 |
| 23 | AML-1 | RUNT | 38 | 9 |
| 24 | Irf-2 | TRP-CLUSTER | 12 | 18 |
| 25 | E2F | Unknown | 10 | 8 |
| 26 | MZF_1-4 | ZN-FINGER, C2H2 | 20 | 6 |
| 27 | MZF_5-13 | ZN-FINGER, C2H2 | 16 | 10 |
| 28 | RREB-1 | ZN-FINGER, C2H2 | 11 | 20 |
| 29 | SP-1 | ZN-FINGER, C2H2 | 8 | 10 |
| 30 | Yin-Yang | ZN-FINGER, C2H2 | 17 | 6 |
| 31 | GATA-2 | ZN-FINGER, GATA | 53 | 5 |
| 32 | GATA-3 | ZN-FINGER, GATA | 63 | 6 |
| 33 | Hen-1 | bHLH | 54 | 12 |
| 34 | Tal 1 beta-E47S | bHLH | 44 | 12 |
| 35 | Thing-E47 | bHLH | 29 | 12 |
| 36 | Max | bHLH-ZIP | 17 | 10 |
| 37 | Myc-Max | bHLH-ZIP | 21 | 11 |

| | | | | |
|---|---|---|---|---|
| 38 | USF | bHLH-ZIP | 30 | 7 |
| 39 | CREB | bZIP | 16 | 12 |
| 40 | E4BP4 | bZIP | 23 | 11 |
| 41 | HLF | bZIP | 18 | 12 |

**Table 3: Mouse transcription factors with the recognized TFBS and their lengths**

| S.No | Name of TF | Class of TF | Total of TFBS | Length of TFBS |
|---|---|---|---|---|
| 1 | SOX17 | HMG | 31 | 9 |
| 2 | Sox-5 | HMG | 23 | 7 |
| 3 | EN-1 | HOMEO | 10 | 11 |
| 4 | Nkx | HOMEO | 17 | 7 |
| 5 | S8 | HOMEO | 59 | 5 |
| 6 | Bsap | PAIRED | 12 | 20 |
| 7 | Pax-2 | PAIRED | 31 | 8 |
| 8 | Brachyury | T-BOX | 40 | 11 |
| 9 | Evi-1 | ZN-FINGER, C2H2 | 47 | 14 |
| 10 | ARNT | bHLH | 20 | 20 |
| 11 | Ahr-ARNT | bHLH | 24 | 24 |
| 12 | n-MYC | bHLH-ZIP | 31 | 31 |
| 13 | Spz-1 | bHLH-ZIP | 12 | 12 |

## Calculation of information content from TFBS

Information content is calculated from the mono and dinucleotide substitution matrices of multiple aligned sequences (already aligned in the database used). Mononucleotide substitutions (Figure 1A) in multiple aligned sequences will give neighbor-independent nucleotide substitution matrices. The replacements are calculated in each column of the block and the summed results of all columns are stored in a 4×4 matrix. The total number of nucleotide pairs (observed frequency, $q_{i,j}$) in a given block is $\frac{\mathrm{w}s(s-1)}{2}$ and the total number of nucleotides (expected frequency, $p_i$) in the block is $ws$, where $s$ is the number of nucleotides in the given position and $w$ is the block width. The resulting 4×4 matrix is used to calculate the “log-odds” (usually logarithm of base 2) and is given by $s_{i,j} = \log_2 \frac{q_{i,j}}{p_i p_j}$ [25,26].

Dinucleotide substitutions (Figure 1B) in multiple sequence alignment will give neighbor-dependent substitution matrices. The total number of dinucleotide pairs (observed frequency, $q_{ij,kl}$) in a given block is $\frac{(\mathrm{w}\text{-}1)\, s(s-1)}{2}$ and the total number of dinucleotides (expected frequency, $p_{ij}$) is given by $(w-1)s$, where $s$ is the number of sequences and $w$ is the block width. The resulting 16×16 matrix is used to calculate the log-odds and is given by

$s_{ij,kl} = \log_2 \frac{q_{ij,kl}}{p_{ij} p_{kl}}$. The average mutual information content (H) is the relative entropy of the target and background pair frequencies and can be thought of as a measure of the average amount of information (in bits) available per nucleotide pair [26]. The average mutual information content in a given block of neighbor-independent and neighbor-dependent substitution matrices is given by $H = \sum_{ij} q_{ij} s_{ij} = \sum_{ij} q_{ij} \log_2 \frac{q_{ij}}{p_i p_j}$ and $H = \sum_{ij,kl} q_{ij,kl} s_{ij,kl} = \sum_{ij,kl} q_{ij,kl} \log_2 \frac{q_{ij,kl}}{p_{ij} p_{kl}}$ respectively.

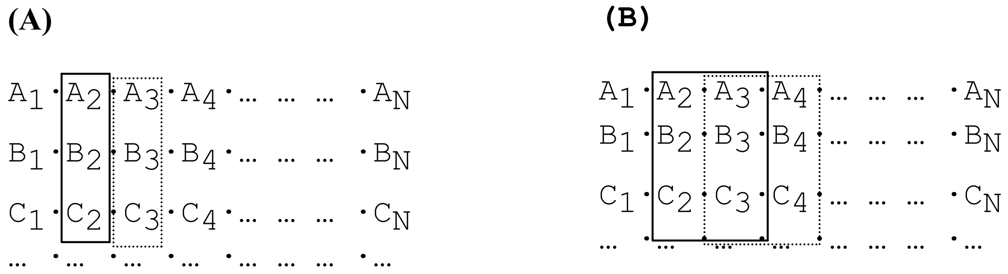
(A)
A1 A2 A3 A4 ... ... ... AN
B1 B2 B3 B4 ... ... ... BN
C1 C2 C3 C4 ... ... ... CN
... ... ... ... ... ... ... ...
(B)
A1 A2 A3 A4 ... ... ... AN
B1 B2 B3 B4 ... ... ... BN
C1 C2 C3 C4 ... ... ... CN
... ... ... ... ... ... ... ...

**Figure 1: The principle of counting the frequencies illustrated diagrammatically. (A) The left side diagram shows the counting principle for neighbor-independent frequency determination. The three lines show the nucleic acid bases corresponding to the TFBS already aligned in the database. The solid box is used for determination of the actual frequencies and the counts for $A_2$-$B_2$, $A_2$-$C_2$ and $B_2$-$C_2$ are put in a 4×4 matrix. Then the counting box is shifted by one position (dotted box) and the process is repeated. (B) In the right side illustration, we indicate the counting principle for neighbor-dependent (pair-wise) determination of frequencies. In this illustration, we get the actual counts for $A_2A_3$-$B_2B_3$, $A_2A_3$-$C_2C_3$, $B_2B_3$-$C_2C_3$ and these are placed in a 16×16 matrix. The counting box is next moved right by one base position (shown by the dotted box) and the process continued till the TFBS region is completed. See text for the details.**

## Noise computations

To ascertain the reliability of the results, we have computed the information content based on a sample sequence (of length 20 nucleotides) selected at random. The sample sequence was subjected to a BLAST search against the respective genome (NCBI sample BLAST with default parameters) and BAC clone sequences were excluded from the results. Finally, 18 best matches for human genome and 14 best matches for mouse genome were taken. The information content was computed based on neighbor-independent and neighbor-dependent procedures as described above. These values are indicated in the histograms as horizontal dotted lines. These values reflect the typical random sequences present in the respective genome to be considered as a reference for comparison. The statistical errors (standard errors) are also indicated in the histograms in the conventional way.

## Functional classification of TFBS

We have used the information content as a basis for classification of the results obtained. We stress that the actual protein sequences were not involved in this computations-only their TFBS. The plotting was done using the PHYLIP suite of software [27]. We have used only the UPGMA and plotting packages from this suite. UPGMA -Unweighted Pair Group Method with Arithmetic mean [28] is a simple data clustering method used for the creation of phylogenetic trees. Here the input data is a collection of information content values of TFBS and the output is a clustered tree. Initially, each object is in its own cluster. At each step, the nearest two clusters are combined into a higher-level cluster. The distance $d_{ij}$ between any two clusters $C(i)$ and $C(j)$ is taken to be the average of all distances between pairs of objects from each cluster. $d(ij)=\frac{1}{|C(i)||C(j)|}\sum_{p\ in\ C(i), q\ in\ C(j)} d(pq)$. Where $|C(i)|$ and $|C(j)|$ denote the number of sequences in clusters *i* and *j*, respectively. Distance matrix is developed for drawing the tree with DRAWGRAM of PHYLIP. Note that the labels in the graphs have been taken from the class-names of the proteins (as given in the database) involved and therefore can occur at multiple places.

# 3 Results

The information content calculated for the 41 and 13 TFBS of human and mouse respectively is presented as a histogram in Figure 2. The dotted line shows typical values of information content for random sequences as a reference of comparison. The error bars (standard errors) calculated on the basis of the elements of the information content matrix (4×4 matrix for the neighbor-independent and 16×16 matrix for the neighbor- dependent) are shown on the histograms in the usual way (they are sufficiently small to be invisible for the graphs on the right Figure2: 1B and 2B).

One interesting pattern that is noticed in the two graphs is that they are quite similar but not same. In particular when we compare graphs in Figure 2, we find that neither the largest peaks nor the smallest peaks correspond with each other. We conclude that the consideration of the neighbor dependence provides additional information but the broad features are similar (strong peaks remain big and weak peaks are also weak in both).

We also note another aspect with respect to the random sequence information content. The random sequence chosen is not expected to correspond to a TFBS. If we consider the neighbor-independent plot (Figure 2: 1A and 2A), the information content of the random sequence is 0.3211 and 0.3863 bits for human and mouse respectively (this corresponds to the dotted horizontal line). We note that these random sequence information content values are nearly mean to the actual TFBS information content values (here random sequence representing actual TFBS). When we consider the neighbor-dependent graphs (Figure2: 1B and 2B) the information content of the random sequence is 2.101 and 2.227 bits for human and mouse respectively (this corresponds to the dotted horizontal line) we note that only one (for humans) or two (for mice) are above the line. This suggests that there exists a strong and specific correlation between the neighbor nucleotides in the TFBS regions. This correlation is significantly different from the typical genome regions (as represented by the dotted line).

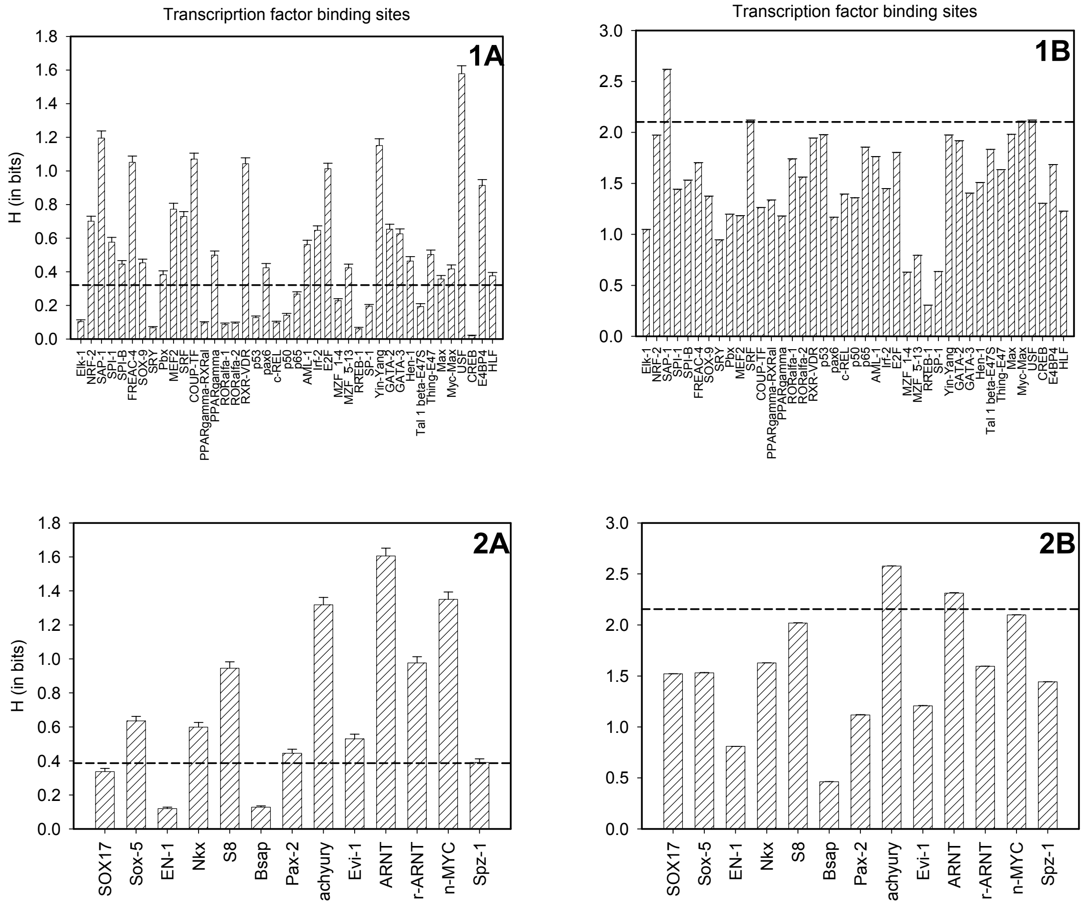

**Figure 2: The average mutual information content H (in bits) of TFBS (calculated by (A; left) neighbor-independent and (B; right) neighbor-dependent nucleotide substitutions) of human (1A and 1B) and mouse (2A and 2B). The dotted line represents information content of random sequence of their genome. The bars of the histograms represent the standard errors of the 16 $H_{ij}$ neighbor-independent substitution matrices. In case of neighbor dependent substitution matrices, (256 $H_{ij,kl}$) the error bars have been calculated as the standard errors of these 256 elements. The standard errors have been actually plotted but cannot be seen, as they are too small in case of neighbor-dependent.**

Even though the neighbor-dependency is observed in TFBS, the information content values of both neighbor-independent and neighbor-dependent substitution matrices are used for clustering analysis. The trees (Figure 3 and Figure 4) represent the clustering of TFBS. In a tree, each node with descendants represents the functional group of TFBS that has close information content values. We believe that one factor may bind to multiple TFBS and cause initiation of transcription of a group of proteins. The results of the clustering suggest that this is likely to be the possible event.

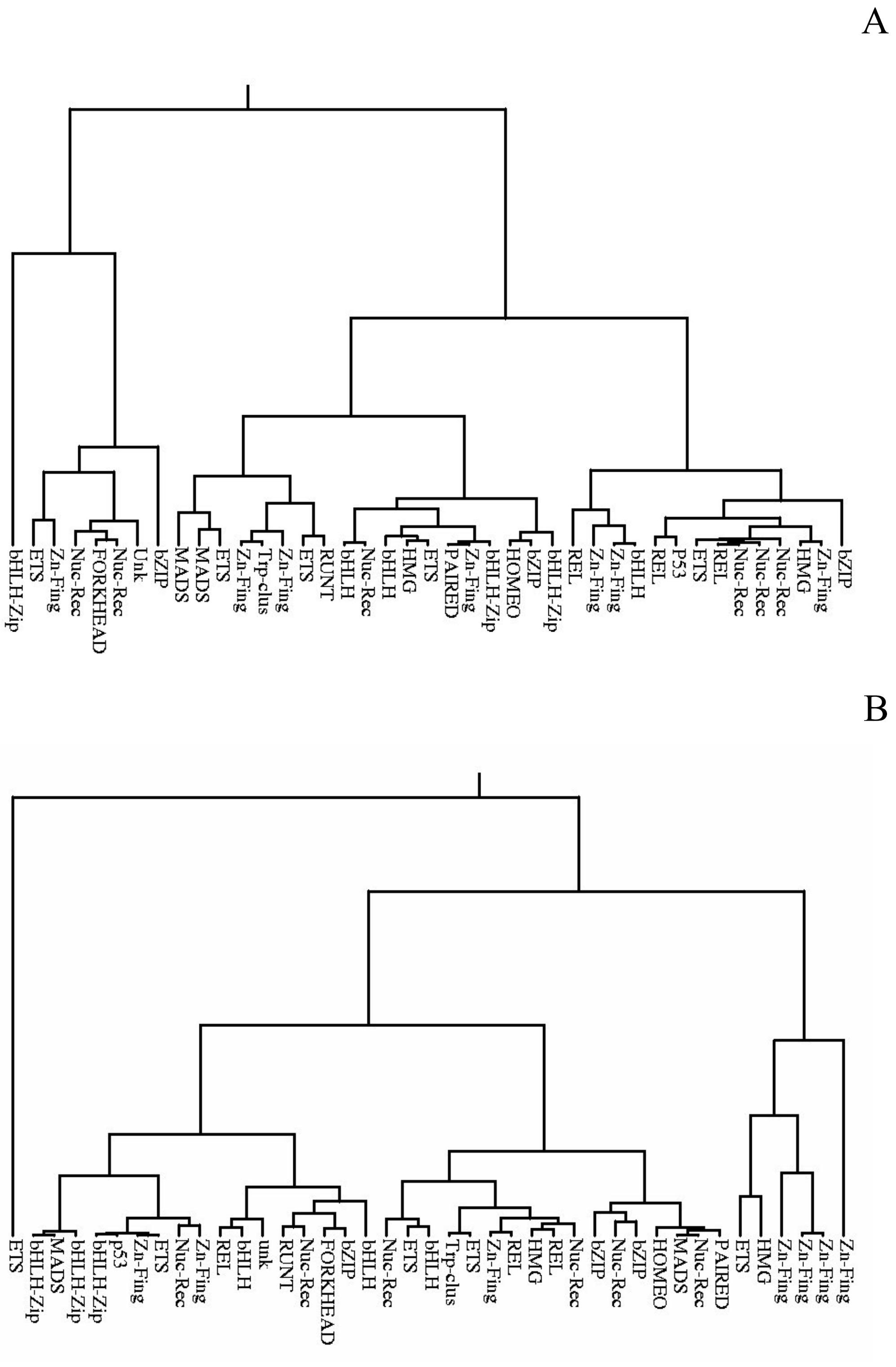

**Figure 3: Functional classification of TFBS in Human; information content is calculated from nucleotide (A) neighbor-independent and (B) neighbor-dependent substitution matrices**

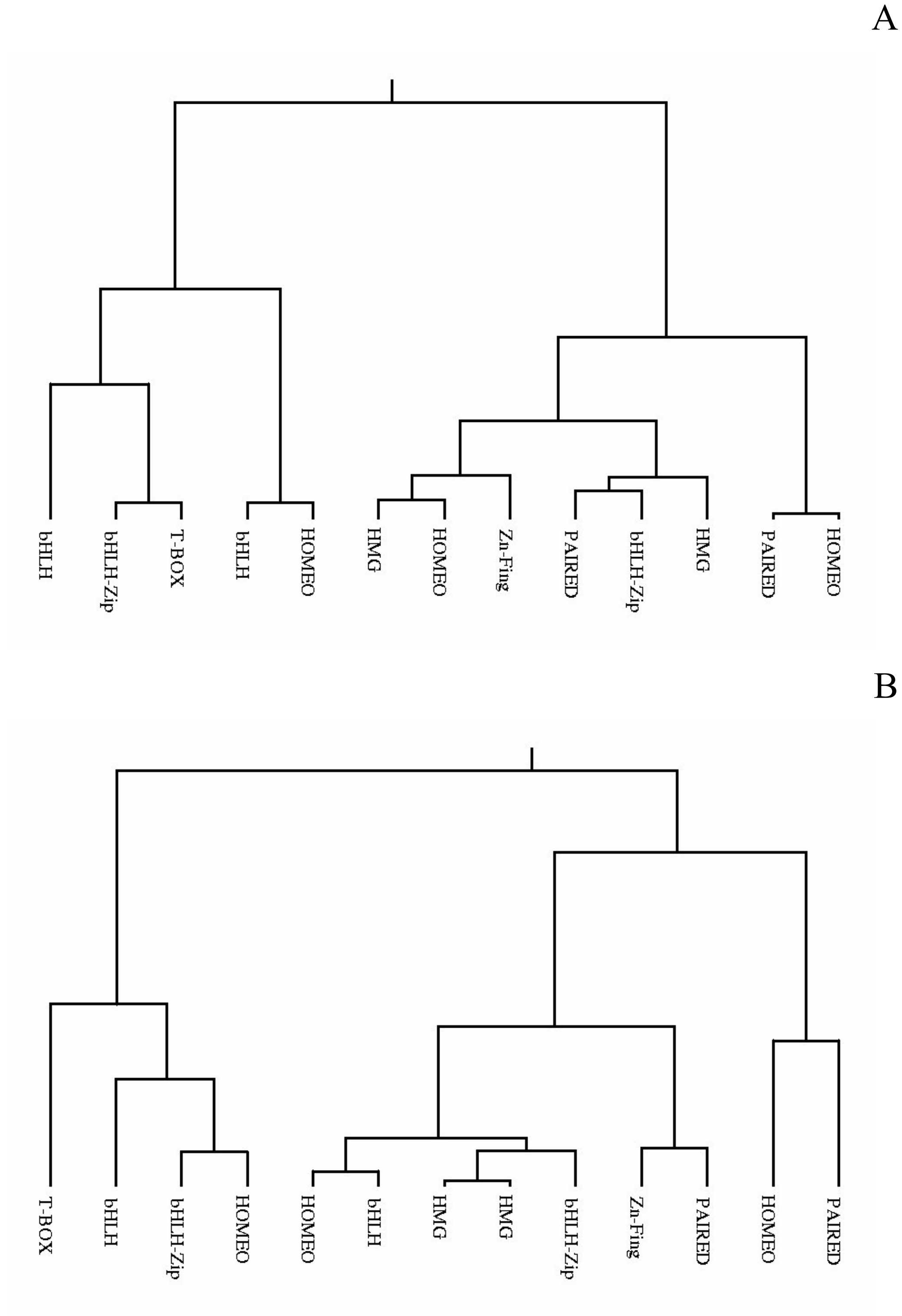

**Figure 4: Functional classification of TFBS in Mouse; information content is calculated from nucleotide (A) neighbor-independent and (B) neighbor-dependent substitution matrices**

We note that the neighbor-dependent and independent results are different in details but are very similar in broad appearance. In case of mouse the two results are practically identical particularly at early times. However, we fully understand that there may be effects due to smaller size of the sample (13 vs 41 in case of human). The information about the TFs that are involved in a specific gene regulation of human (Table 4) and mouse (Table 5) were collected from the Transcription Regulatory Regions Database (TRRD). The results of hierarchical clustering of TFBS of specific TFs were compared with the TFs that are involved in specific gene regulation in TRRD. These results show that the computational results are comparable with the experimental results.

**Table 4: Human gene with number of TFs involved for their regulation and class/family name of TFs (Data is extracted from TRRD database [29])**

| S.No | Gene name | Name of TF | Class/family of TF |
|---|---|---|---|
| 1 | Angiotensinogen (AGT) | HNF-4 | ZN-FINGER, C2H2 |
| | | COUP-TF | NUCLEAR RECEPTOR |
| | | DBP | bHLH-ZIP |
| | | c/EBPdelta | bZIP |
| | | USF1 | bHLH-ZIP |
| 2 | GammaF-crystallin (CRYGF) | RAR/RXR | NUCLEAR RECEPTOR |
| | | Pax-6 | PAIRED |
| | | Prox-1 | HOMEO |
| | | Sox-1 | HMG |
| | | L-Maf | bZIP |
| 3 | Multidrug resistance (MDR1) | HSF1 | HSF family |
| | | NF-IL6 | bZIP |
| | | NF-R1 | bZIP |
| | | NF-kB | bZIP |
| | | NF-Y | CBF family |
| | | YB-1 | cold-shock domain factors |
| | | SP1 | ZN-FINGER, C2H2 |
| | | WT1 | ZN-FINGER, C2H2 |
| 4 | COX5B | GABP | ETS |
| | | SP1 | ZN-FINGER, C2H2 |
| | | YY-1 | ZN-FINGER, C2H2 |
| | | USF2 | bHLH |
| 5 | CDC25C | SP1 | ZN-FINGER, C2H2 |
| | | NF-Y | CBF family |
| | | p53 | P53 |
| | | CDF-1 | CDF family |
| | | YY-1 | ZN-FINGER, C2H2 |

**Table 5**: **Mouse gene with number of TFs involved for their regulation and class/family name of TFs (Data is extracted from TRRD database)**

| S.No | Gene name | Name of TF | Class/family of TF |
|---|---|---|---|

| | | | |
|---|---|---|---|
| 1 | GammaF-crystallin (CRYGF) | RAR/RXR | NUCLEAR RECEPTOR |
| | | Pax-6 | PAIRED |
| | | Sox1 | HMG |
| | | Prox-1 | HOMEO |
| | | Six-3 | HOMEO |
| 2 | CACNA1S | Sox-5 | HMG |
| | | GATA-2 | ZN-FINGER, GATA |
| | | CREB | bZIP |
| 3 | HOXA7 | Antp | HOMEO |
| | | Ftz | HOMEO |
| | | Cad | HOMEO |

## 4 Discussion

The information content (relative entropy) of TFBS is used to identify the TF classes required to regulate a specific gene expression. When we look at two TFBS, we have information in the form of the differences in between these two TFBS, a measure that we can interpret as the distance between the TFBS. If a small distance separates two TFBS then they may have a common TF binding site.

We note that apparently diverse proteins are placed closely in the classification given in this study. This is not surprising as their TFBS are likely to be very similar. This suggests that these groups of proteins may be needed together and they may share the same transcription factors. Thus, out of the 41 TFBS (in humans), perhaps only 5-10 or so transcription factors may be actually needed (instead of 41 different transcription factors). For the mouse TFBS, instead of 13 transcription factors, we may have actually 5 factors. The JASPAR database TFBS are used in this study. The experimental data of TFs of specific gene expression from TRRD database is also coinciding with our computational results. This gives us a new way to look at the protein classification- not based on their structure or function of TFs - but by the nature of their transcription factor binding sites.

## 5 References

[1] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt *et al.*, The sequence of the human genome. Science, 291: 1304-1351, 2001.

[2] R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, *et al.*, Initial sequencing and comparative analysis of the mouse genome. Nature, 420: 520–562, 2002.

[3] M. E. Vazquez, A. M. Caamano and J. L. Mascarenas. From transcription factors to designed sequence-specific DNA-binding peptides. Chemical Society Reviews Articles, 32: 338-49, 2003.

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman. Basic local alignment search tool. Journal of Molecular Biology, 215: 403-410, 1990.

[5] W. R. Pearson and D. J. Lipman. Improved Tools for Biological Sequence comparision. Proc. Proceedings of the National Academy of Sciences ,USA, 85:2444, 1988.

[6] S. F. Altschul. A protein alignment scoring system sensitive at all evolutionary distances. Journal of Molecular Evolution, 36: 290-300, 1993.

[7] H. B. Nicholas Jr, D. W. Deerfield II and A. J. Ropelewski. Overviw: Strategies for searching sequence databases. BioTechniques, 28: 1174-1191, 2000.

[8] A. R. Panchenko and S. H. Bryant. A comparison of position-specific score matrices based on sequence and structure alignments. Protein Science, 11: 361-370, 2002.

[9] Y. -K. Yu, J.C. Wootton and S.F. Altschul. The compositional adjustment of amino acid substitution matrices. Proceedings of the National Academy of Sciences ,USA, 100: 15688-15693, 2003.

[10] Y. -K. Yu and S.F. Altschul. The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions. Bioinformatics, 21, 902-911, 2005.

[11] M.O. Dayhoff, R.M. Schwartz and B.C. Orcutt. in Atlas of protein sequence and structure (Eds) M.O. Dayhoff, National Biomedical Research Foundation, Washington, DC, 5, pp. 345-352, 1978.

[12] R. M. Schwartz and M.O. Dayhoff. in Atlas of protein sequence and structure; (eds) M.O. Dayhoff. National Biomedical Research Foundation, Washington, DC, 5, pp. 353-358, 1978.

[13] S. Henikoff, J. G. Henikoff. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences ,USA, 89: 10915-10919, 1992.

[14] G. Lunter and J. Hein. A nucleotide substitution model with nearest-neighbor interactions. Bioinformatics, 20: i216-i223, 2004.

[15] G. D. Stormo. DNA binding sites: representation and discovery. Bioinformatics, 16: 16-23. 2000.

[16] P. Arndt and T. Hwa. Identification and measurement of neighbor-dependent nucleotide substitution process. Bioinformatics, 21: 2322-2328, 2005.

[17] D. A. Reddy, B. V. L. S. Prasad and C. K. Mitra. Comparative analysis of core promoter region: Information content from mono and dinucleotide substitution matrices. Computational Biology and Chemistry, 30,58-62, 2006.

[18] M. Stepanova, T. Tiazhelova, M. Skoblov and A. Baranova. A comparative analysis of relative occurrence of transcription factor binding sites in vertebrate genomes and gene promoter areas. Bioinformatics, 21: 1789-1796, 2005.

[19] N. I. Gershenzon, G. D. Stormo and I. P. Ioshikhes. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites. Nucleic Acids Research, 33: 2290-2301, 2005.

[20] R. Staden. Methods to define and locate patterns of motifs in sequences. Computer Applications in the Biosciences, 4: 53-60, 1988.

[21] M. L. Bulyk, P. L. F. Johnson and G. M. Church. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. Nucleic Acids Research, 30: 1255-1261, 2002.

[22] L. Narlikar and A. Hartemink. Sequence features of DNA binding sites reveal structural class of associated transcription factor. Bioinformatics, 22(2),157-163, 2006.

[23] S. Hannenhalli and S. Levy. Predicting transcription factor synergism. Nucleic Acids Research, 30: 4278-4284, 2002.

[24] A. Sandelin, W. Alkema, P. Engström, W. Wasserman and B. Lenhard. JASPAR: an open access database for eukaryotic transcription factor binding profiles. Nucleic Acids Research, 32(1): D91-D94, 2004.

[25] S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proceedings of the National Academy of Sciences ,USA, 87: 2264-2268, 1990.

[26] S. F. Altschul. Amino acid substitution matrices from an information theoretic perspective. Journal of Molecular Biology, 219: 555-565, 1991.

[27] J. Felsenstein, PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle, 2004.

[28] R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. University of Kanas Scientific Bulletin. 28: 1409-1438, 1958.

[29] N. A. Kolchanov, E. V. Ignatieva, E. A. Ananko, O. A. Podkolodnaya, I. L. Stepanenko, T. I. Merkulova, M. A. Pozdnyakov, N. L. Podkolodny, A. N. Naumochkin and A. G. Romashchenko. Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucleic Acids Research, 30: 312-317, 2002.