

Original Research Article

<https://doi.org/10.20546/ijcmas.2017.606.091>

## Few selected Human Transcription Factors Sequence Analysis and their Phylogenetic Relationship

Karmveer Yadav\*

Department of Biochemistry, University of Hyderabad, Hyderabad, 500046, India

\*Corresponding author

### ABSTRACT

Transcription factors (TFs) are DNA-binding proteins responsible for initiating transcription of particular genes upon interacting with specific DNA sequences located in their promoter or enhancer regions. Most of the transcription factors act as dimers (homodimer or heterodimer) or as higher order multimers. Therefore, it is useful to investigate associated structural and functional relationship between TFs and transcription factor binding sites (TFBSs). With increasing computational power, massive experimental databases available for DNA and proteins and recent data mining techniques, it becomes feasible to study the phylogenetic relationship between TFs and TFBSs. In the present study, we have made a framework to predict the evolutionary relationship between TFs and TFBSs patterns in the most explicit and interpretable form using JASPAR and Phylogeny.fr. We have collected 20 human TFBSs from JASPAR database. We have constructed and analysed phylogenetic relationships between TFBS using Phylogeny.fr web service, and conclude that specific domain ( $\alpha$ -helix) of nearly all transcription factor families interact in a same way to regulate gene expression pattern. We propose that transcription factors interact in different ways to regulate the expression of different pathways and are phylogenetic ally related, they may have evolved from common ancestor indicating functional divergence.

#### Keywords

Human transcription, DNA, Proteins, Phylogenetic relationship.

#### Article Info

##### Accepted:

14 May 2017

##### Available Online:

10 June 2017

### Introduction

All cells in a multi cellular organism contain the same genetic information; however, each organism consists of a large array of cell types that perform diverse biological functions. This cell type diversity results from differences in gene expression, the process by which information encoded in DNA is transcribed to RNA, and then in many cases, translated into proteins that are used by the cell to perform specific functions. Importantly, gene expression is tightly regulated to produce the right protein at the right time in each cell. The interaction

between transcriptions factors (TFs) and their transcription factor binding sequences (TFBSs) is essential for many biological processes. For instance, the interaction at the core promoter regions determines the assembly of the pre-initiation complex and the initiation of transcription (Wang and Hannenhalli, 2006; Wang, *et al.*, 2007), whereas interactions in the distal promoter/enhancer region determine the rate of transcription in cell type, tissue and developmental stage-specific manner (Juven-Gershon, *et al.*, 2008). Therefore, the study of

TF-TFBS interaction is critical to our understanding of the transcriptional regulatory network of gene expression. The regulation of the tissue specific gene expression in response to specific stimulus is a critical process in the efficient functioning of the cell. All genes which show similar pattern of gene regulation will show similar type of consensus sequences, *e.g.* Heat shock elements are common for all those genes which are expressed in the elevated temperature conditions. The transcription factors selectively bind to these DNA sequences and regulate the expression of these genes either positively or negatively.

The most common step in the regulation of transcription is the initiation step. It is energetically most efficient step to regulate because no energy and materials are wasted when regulation focused on the initiation. Another reason is that it is very easy to regulate at the initiation because only single DNA molecule has to be regulated and expressed. The binding of RNA polymerase to promoter region is the most committed step in the gene expression so that most of the regulation is focused on the initiation step. In addition to activating transcription factors, inhibitory transcription factors are also there. Inhibitory transcription factors interfering with the activity of positively acting factors there by blocking the stimulatory effect of the transcription. This can be achieved by (1) Preventing positively acting factors from binding to DNA either via (a) by binding to DNA binding sites of positively acting factors. (b) By formation of Non-DNA binding protein complexes between the positively acting factors and negatively acting factors. (2) Alternatively negatively acting factors could act by interacting with positively acting factors to block the activity of activating domain in a phenomenon is called quenching (Eugenio Vazquez *et al.*, 2003). Sequence analysis has become an

important field in computation biology as the sequence of DNA or protein itself carries a lot of information about the structural, function and evolutionary features of biological sequences. In particular, we usually assume that genes or protein which has similar function and structure; they possibly will have a similar evolutionary history. The tools developed in the field of sequence analysis were aimed to determine the degree of similarity between two sequences. Potential benefits of sequencing the human genome expanded through many fields from molecular medicine to better understanding of human evolution. We can study all the genes in a genome, for example, all the transcripts in a particular tissue or organ or how tens of thousands of genes are proteins work together in interconnected network to orchestrate the chemistry of life. Phylogenetic analysis of nucleic acid and protein sequences is an important area of sequence analysis. Phylogenetic relationships among the genes can help to predict which ones might have an equivalent function that has been conserved during evolution of the corresponding organism. Such functional predictions can then be tested by genetic experiments. The focus of this paper is the transcription factors binding sequence analysis of human transcription factor and their phylogenetic relationship, using computational methods.

## **Materials and Methods**

### **TFBSs Sequence (*Homo sapiens*) collected from JASPAR**

#### **JASPAR**

JASPAR is an open-access database of annotated high-quality, matrix-based transcription factor binding site outlines for multi cellular eukaryotes (Sandelin *et al.*, 2004). The profiles derived in this database were exclusively from sets of nucleotide

sequences that were experimentally demonstrated to bind transcription factor either from SELEX experiments, experimentally determined binding regions of actual regulatory regions and high-throughput technologies like Chip-seq experiments. JASPAR is a web interface for browsing, searching and subset collection, online sequence analysis and tools for genome-wide and comparative genomic analysis of regulatory regions. JASPAR database is available at <http://jaspar.genereg.net/>.

### **TFBSs collection**

The TFBS sequences were collected from the frequency matrices are given for each transcription factor in JASPAR database. For Homo sapiens, 75 transcription factors were reported in the database. The following figures illustrate the collection of TFBS sequences from JASPAR.

As an example, transcription factor CREB1 (ID: MA0018.2) has been described here as reported from the database. The sequence logo and frequency matrix for transcription factor binding site of CREB1 were shown in panel A and B (Fig. 1A). Panel A depicts the sequence logo of transcription factor binding site for CREB1 available from JASPAR database. From the frequency matrix (Fig. 1B), each TFBS sequence was represented based on the high frequency of nucleotide occurring at each position in the matrix (number highlighted in color). TFBS sequence for CREB1 factor can be represented as TGACGTCA. The frequency matrix was constructed based on 11 experimentally verified TFBS sequences (total of each column in the matrix is 11) that were collected from literature and experimental evidence. All the 11 different TFBS sequence were considered as another set of TFBS sequences for CREB1 factor (Fig. 2).

### **Phylogenetic tree construction using Phylogeny.fr**

#### **Phylogeny.fr**

Phylogeny.fr is a free, simple to use web service devoted to reconstructing and analyzing phylogenetic relationships between molecular sequences. Phylogeny.fr runs and connects various bioinformatics programs to reconstruct a robust phylogenetic tree from a set of sequences (Dereeper A., Guignon V, 2008). In 'One Click' mode targets users who do not wish to deal with program and parameter selection. By default, the pipeline is already set up to run and connect well recognized programs: MUSCLE for multiple alignments, blocks for automatic alignment curation, PhyML for tree building and TreeDyn for tree drawing. PhyML is run with the aLRT statistical test of branch support. This test is based on an approximation of the standard Likelihood Ratio Test and is much faster to compute than the usual bootstrap procedure.

We upload the transcription factor binding sequences in FASTA format file and to click the Submit button. The system will do all the rest work; all the parameters are those of programs by default. At the end of the analysis, the server displays an image of the phylogenetic tree.

### **Results and Discussion**

#### **Study of transcription factors binding sequences**

JASPAR website is matrix-based transcription factor binding site (TFBS) profiles and provides the information of experimentally verified TFBS of multicellular eukaryotes. From that website 20 human TFBS and sequences logo were downloaded.

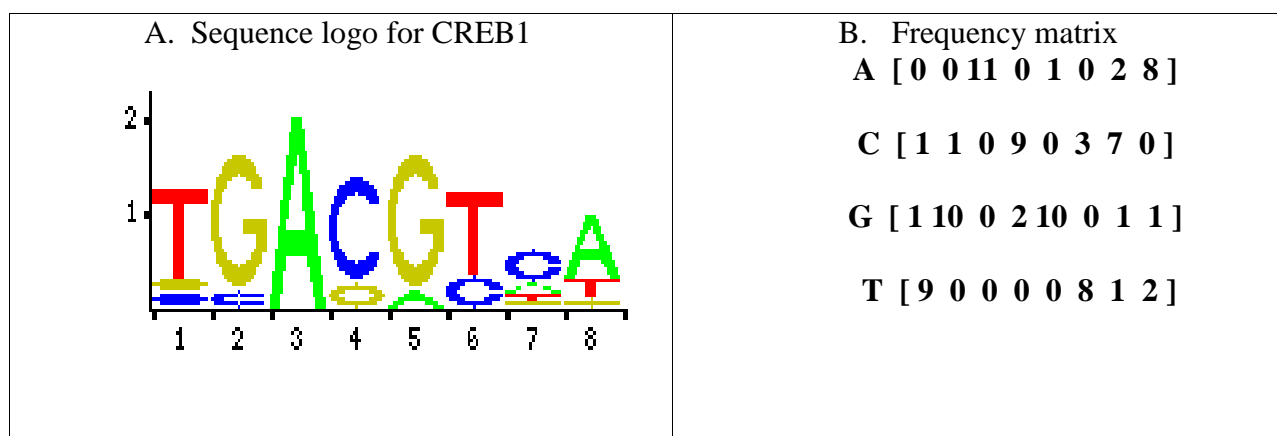
List of 20 human transcription factor binding site sequences collected from JASPAR database (Table 1). ID represents the ID of TF from JASPAR. Column 3 represents the total number of experimentally verified transcription factor binding site (TFBS) sequences that were reported for each TF. TFBS sequences listed here were collected based on the highest frequencies of nucleotides reported from frequency matrices of binding profiles for each TF. Y= C or T, S= G or C.

**Structural and functional correlation between transcription factor binding sites**

We have constructed phylogenetic relationship between transcription factor binding site sequences using Phylogeny.fr web service (Table 2). Closely related sequences are aligned first and then the resulting groups of sequences, which may be less related to one another but have a common ancestor. So there is a possibility that these

sequences may align accurately. Pax6 and ETS1 transcription factors belong to helix-turn-helix (HTH) family. The recognition helix ( $\alpha$ -helix) of these factors inserts in the major groove of DNA and makes several contacts with the bases and phosphate backbone. ETS1 HTH has a role in Hematopoietic cell differentiation and development of lymphoid tissue. It is Proto-oncogene and helps to regulate the invasive behavior of tumor cells, including the expression of VEGF in endothelial cells. NFATC2 (Ig fold) is a nuclear factor for activating T-cells. It is involved in Breast cancer (Pro-invasive and Pro-migratory). BRAC1 is nuclear phosphoprotein that helps to maintain genomic stability and acts as a tumor suppressor. BRAC1 mutation is responsible for Breast and Ovarian cancer. Pax6 (HTH) and SP1 (Zinc finger) they recognized major groove; therefore both the transcription factors have structurally similar recognition site on DNA.

**Fig.1** Sequence logo and frequency matrix for CREB1 (ID. No: MA 0018.2) from JASPAR database. Panel A represents the sequence logo for the binding site of CREB1. Panel B represents the frequency matrix of the binding site sequence for CREB1. From the matrix, the binding site sequence for CREB1 can be represented as TGACGTCA

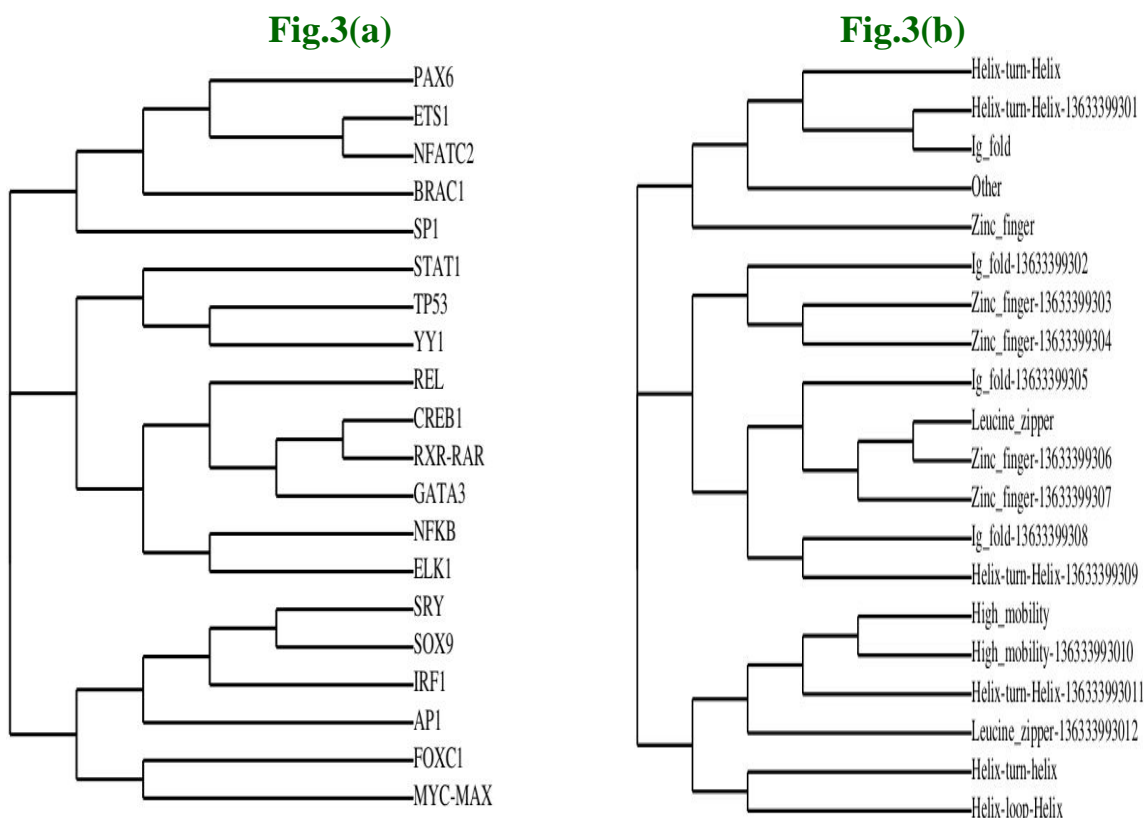


**Fig.2** Experimentally verified binding site sequences that were used to construct the frequency matrix of CREB1 from JASPAR

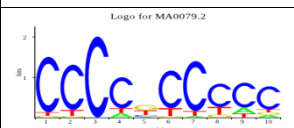


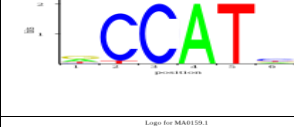
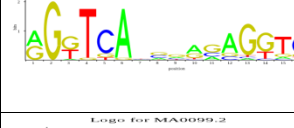

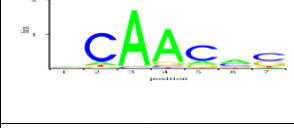

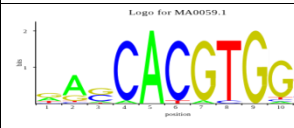
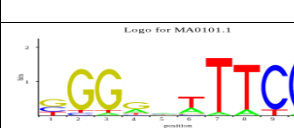
>MA0018.2	CREB1	1	gtcc <b>ATGCGTCA</b> ttag
>MA0018.2	CREB1	2	actga <b>TGACGTCC</b> atg
>MA0018.2	CREB1	3	ggctt <b>TGACGTCA</b> gcct
>MA0018.2	CREB1	4	ctcttt <b>CCAGGTAT</b> ctc
>MA0018.2	CREB1	5	gccc <b>gTGACGCGG</b> ccg
>MA0018.2	CREB1	6	ggcat <b>TGACGTCA</b> aacggcage
>MA0018.2	CREB1	7	gctc <b>gTGACGTCA</b> ccaaga
>MA0018.2	CREB1	8	tcccc <b>gTGACCTCA</b> ctcga
>MA0018.2	CREB1	9	aa <b>TTGCGTCA</b> tttc
>MA0018.2	CREB1	10	tcatact <b>gTGACGTCT</b> ttcag
>MA0018.2	CREB1	11	tctct <b>gTGACGTCA</b> cgac

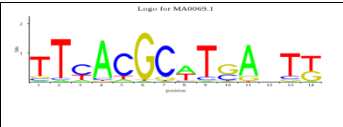
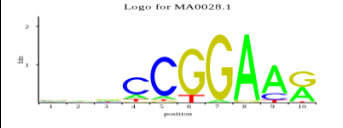


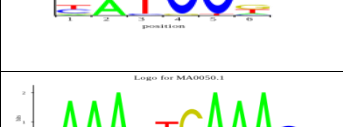

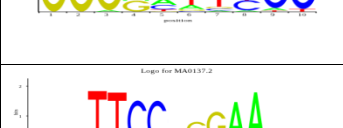
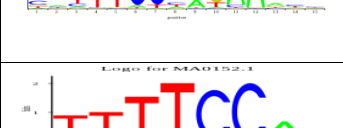
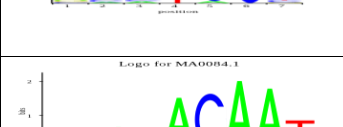

Fig. 2 depicts list of experimentally verified binding site information available for CREB1 in JASPAR database. The 11 sequences were used for generation of frequency matrix and sequence logo of binding site sequence for CREB1 shown in Fig. 2.1. All the binding site sequences were represented in Fig. 2.2. The text highlighted in color was extracted from each sequence and studied as another set of transcription factor binding site sequence for CREB1. In JASPAR database, out of 75 TFs, binding site information for 65 TFs is available.

**Fig.3a** Phylogenetic relationship among the transcription factor binding sequences; **b** The structural relationship among the transcription factor



**Table.1** List of 20 human transcription factor binding site sequences collected from JASPAR database. ID represents the ID of TF from JASPAR. Column 3 represents the total number of experimentally verified transcription factor binding site (TFBS) sequences that were reported for each TF. TFBS sequences listed here were collected based on the highest frequencies of nucleotides reported from frequency matrices of binding profiles for each TF.  
Y= C or T, S= G or C

S. No.	ID No.	TF NAME	No. of TFBS	RECOGNITION SEQUENCE	RECOGNITION SEQUENCE LOGO
1.	MA0079.2	SP1	35	CCCCGCCCC	
2	MA0037.1	GATA-3	63	AGATAG	
3	MA0106.1	TP53		CCGGACATGCCCGGGCATGT	
4	MA0095.1	YY1	17	GCCATC	
5	MA0159.1	RXR-RAR-DR5	23	AGGTCAYGGAGAGGTCA	
6	MA0099.2	AP1	18	TGACTCA	
7	MA0133.1	BRAC1	43	ACAACAC	
8	MA0018.2	CREB1	11	TGACGTCA	
9	MA0059.1	MYC-MAX	21	GASCACGTGGT	
10	MA0101.1	REL	17	GGGGATTTC	

11	MA0069.1	Pax-6	43	TTCACGCATGAGTT	
12	MA0028.1	ELK-1	28	GAGCCGGAAT	
13	MA0032.1	FoxC1	-	GGTAAGTA	
14	MA0098.1	ETS1	40	YTTCCG	
15	MA0050.1	IRF-1	-	GAAAGSYGAAACC	
16	MA0061.1	NF-KB	38	GGGAATTTCC	
17	MA0137.2	STAT1	2085	CATTTCCTGGAAACC	
18	MA0152.1	NFATC2	26	TTTTCCA	
19	MA0084.1	SRY	28	GTAAACAAT	
20	MA0077.1	SOX9	76	GAACAATGG	

**Table.2** Functional relationship between transcription factors

S. No.	TFs NAME	FUNCTION
1	PAX6	Development of sensory system (eye, nasal and olfactory tissue).
2	ETS1	Hematopoietic cell differentiation and lymphoid tissue development.
3	NFATC2	Activation of T-cells and breast cancer (pro-invasive and pro-migratory).
4	BRAC1	Tumor suppressor, DNA repair and DNA recombination (mutation leads breast cancer)
5	SP1	Chromatin remodeling, apoptosis, immune response and DNA damage
6	STAT1	Anti-viral activity binding with IFN- $\gamma$ , EGF, PDGF, IL-6. STAT-1 interact with p53 and enhance its growth arrest and apoptosis- inducing properties (Latchman and Stephanou, 2004)
7	TP53	Cell cycle arrest, Apoptosis, DNA repair and senescence.
8	YY1	Directs Histone deacetylase and Histone acetyl transferase.
9	REL	Immuno & inflammatory response, developmental process and apoptosis.
10	NF- $\kappa$ B	Inflammation, immunity, cell growth, tumor genesis and apoptosis.
11	CREB1	Induce the transcription of gene in response to cAMP pathway.
12	RXR-RAR	Activation of steroid and thyroid receptor
13	GATA3	Regulation of T- cell development.
14	ELK1	Long term memory formation, drug addiction and Alzheimer.
15	SRY	Testis differentiation.
16	SOX9	Sertolli cell differentiation.
17	IRF1	Activator of $\alpha$ and $\beta$ -interferon transcription and inducible of MHC I.
18	API1	Regulates gene expression in response cytokines, growth factor and bacterial & viral.
19	FOXC1	Regulation of embryonic and ocular development.
20	MYC-MAX	Oncoprotein control cell proliferation, differentiation and Apoptosis.

STAT1 (Ig fold) is inducing the cellular antiviral activity binding with Interferon- $\gamma$ . STAT-1 interacts with p53 and enhances its growth arrest and apoptosis- inducing properties (Latchman and Stephanou, 2004). REL (Ig fold) and NF- $\kappa$ B (Ig fold) proteins a family of structurally related transcription factors that are involved in the control of a large number of normal cellular and organ specific processes such as immune, inflammatory responses, developmental process, cellular growth and apoptosis.

TP53 (Zinc finger) and YY1 (Zinc finger) transcription factors are sequence- specific

DNA recognition is achieved by presentation of  $\alpha$ -helix into the major groove of the double helix, where it comes into contact with the 3-4 base pair-long site. YY1 inhibits the activation of TP53 tumor suppressor in response to genetic stress. GATA3 (Zinc finger), RXR-RAR (Zinc finger) and CREB1 (Leucine zipper) they recognized major groove, therefore, these transcription factors are structurally similar recognition of DNA sequences. MYC-MAX (helix-loop-helix) and FOXC1 (helix-turn-helix) are structurally related transcription factors. MYC-MAX is on coprotein implicated in cell proliferation, differentiation and apoptosis. FOXC1 play



role in the regulation of embryonic and ocular development. It promotes Breast cancer invasive by inducing matrix metalloprotease-7 expression. AP1 (Leucine zipper) and IRF1 (helix-turn-helix) are structurally related transcription factors. AP1 regulates gene expression in response to the verity of stimuli, including cytokines, growth factors, bacterial and viral infection. It controls the cellular processes including differentiation, proliferation and apoptosis. IRF1 serves as an activator of  $\alpha$  and  $\beta$ -interferon transcription. It plays role in regulating apoptosis and tumor suppression. SRY (high mobility) and SOX9 (high mobility) are structurally and functionally related transcription factors. SRY initiate testis differentiation by activating male specific transcription factors. SRY accomplishes this by up regulating SOX9. SOX9 is a transcription factor that up regulates fibroblast growth factor (fgf-9). Fgf-9 necessary for proper sertolli cell differentiation.

The present work laid some light on a computational study of transcription factors. We analyzed the phylogenetic relationship of 20 Human transcription factors, at two different levels: Structural binding of TFs and Function of TFs. In the first level, it is identified the presence of structurally similar recognition motif ( $\alpha$ -helix) in different transcription factors (Fig. 3b) indicating an evolutionary relationship of human TFs. These motifs suggested that these TFs are diverged (evolved) from a common ancestral gene and certain amino acid residues in the structure of motifs seem to be crucial to maintaining the helix-turn-helix or zinc finger or leucine zipper structure of the motif. In the second level, we observed that transcription factors, interact via different ways to regulate the expression of same or different pathways, are phylogenetically related (Fig. 3a). So that it can be concluded that all these transcription factors may have evolved from common ancestor indicating functional divergence.

## Acknowledgements

The authors would like to thank Head of Department Biochemistry, University of Hyderabad for providing research facility.

## References

- Dereeper, A., Guignon, V., *et al.*, 2008. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. *Nucleic Acids Res.*, 36: 465-469.
- Harrison, A., Binder, H., *et al.*, 2013. Physico-chemical foundations underpinning microarray and next-generation sequencing experiments. *Nucleic Acids Res.*, 41 (5): 2779-2796.
- Jolma, A., Yan, J., *et al.*, 2013. DNA-Binding Specificities of Human Transcription Factors. *Cell*, 152: 327-339.
- Latchman, D. S. and Stephanou A., 2004. STAT1 and STAT3: Closely Related Transcription Factors with Antagonistic Effects on Cell Proliferation and Apoptosis. *Curr. Genomics*, 5(5): 453-457.
- Latchman, D. S., 1997. Transcription factor overview. *Int. J. Biochem. Crrl Biol.*, 29(12): 1305-1312.
- Narlikar L., and Hartemink A. J., 2006. Sequence features of DNA binding sites reveal structural class of associated transcription factor. *Bioinformatics*, 22(2): 157-163.
- Portales-Casamar, E., Thongjuea, S., *et al.*, 2010. JASPAR the greatly expanded open-access database of transcription factor binding profiles, *Nucleic Acids Res.*, 38: 105-110.
- Reddy, D. A., Prasad, S. B. V. L. and Mitra, C. K., 2006. Functional classification of transcription factor binding sites: Information content as metric. *J. Integrative Bioinformatics*, 3 (1).
- Vazquez, M. E., Caamano A. M. and Mascarenas J.L., 2003. Transcription

- factor to designed sequence-specific DNA-binding peptides. *Chem. Soc. Rev.*, 32: 338-349.
- Wingender, E., Schoeps, T. and Donitz J., 2013. TFClass: an expandable hierarchical classification of human transcription factors. *Nucleic Acids Res.*, 41: 165-170.
- Yang, S., Yalamanchili, H. K., *et al.*, 2011. Correlated evolution of transcription factors and their binding sites. Oxford University Press, 27 (21): 2972-2978.
- Wang, J. and Hannehalli, S., 2006. A mammalian promoter model links cis elements to genetic networks. *Biochem. Biophys. Res. Commun.*, 347(1): 166-177.
- Wang, J., Hannehalli, S. and Ungar, L., 2007. MetaProm: a neural network based metapredictor for alternative human promoter prediction. *BMC Genomics*, 8: 374-386.
- Juven-Gershon, T., Hsu, Jer-Yuan, *et al.*, 2008. The RNA polymerase II Core Promoter – the Gateway to Transcription. *Curr. Opin. Cell Biol.*, 20(3): 253–259.
- Sandelin, A., Alkema, W., *et al.*, 2004. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, 32: 91-94.

**How to cite this article:**

Karmveer Yadav. 2017. Few Selected Human Transcription Factors Sequence Analysis and their Phylogenetic Relationship. *Int.J.Curr.Microbiol.App.Sci.* 6(6): 776-785.  
doi: <https://doi.org/10.20546/ijcmas.2017.606.091>