Identification and Characterization of Rat ETV6 Gene by Using Bioinformatics Tools

Lokman VARISLI Osman CEN
Department of Biology, Faculty of Art and Science, Harran University, Sanliurfa 63300, TURKEY

Corresponding Author
E-mail : lokmanv@gmail.com

Received : 26 April 2006
Accepted : 24 July 2006

Abstract

The products of ETS (erythroblast transformation specific domain) family genes are essential for a variety of developmental processes. This group of genes is frequently involved in leukemia and sarcoma as the result of specific chromosomal translocations. The ETV6 (ETS variant gene 6) gene, also called Tel or Tel1, is a member of the ETS family transcriptional regulators. Using *in silico* bioinformatic tools, we identified and characterized the rat ETV6 gene in this study. It has 8 exons and is located on the rat genomic contig NW_047696 of the rat chromosome 4. Its protein product has a predicted nuclear localization signal, one SAM_PNT (SAM-Pointed domain), and one ETS conserved domain. It has high homology to those of other vertebrate species. This is the first report on identification and characterization of the rat ETV6 gene.

Key words: Bioinformatics, comparative genomics, ETV6, rat genome

INTRODUCTION

The ETV6 gene encodes for a nuclear phosphoprotein that belongs to ETS family of transcription factors [1]. Product of this gene is a sequence-specific transcriptional repressor of ETS-binding site-driven transcription promoters [2]. The predicted ETV6 protein contains two putative functional domains, the N terminal PNT and the C terminal ETS domains. The PNT domain is necessary for homodimerization and heterodimerization of ETV6 [3]. The ETS domain is a sequence specific DNA binding domain but it also mediates protein–protein interaction [4]. Genetic alterations in the ETV6 genes are associated with various types of leukemic disorders [5]. In leukemia and lymphoma, specific chromosomal translocations result in the fusion of two genes that sometimes produce chimeric proteins responsible for malignant transformations [6]. Many genes, including those of tyrosine kinases and some transcription factors, have been isolated as the fusion partner of ETV6 [7]. Therefore, characterization of the ETV6 gene in the animal models is important in leukemia-lymphoma research.

The aim of the study was *in silico* identification and characterization of the rat ETV6 gene. Its structure and chromosomal localization, complete coding sequence, amino acid sequence and protein motifs, and domains were investigated. This is the first report on the identification and characterization of the rat ETV6 gene.

MATERIALS AND METHODS

Identification of novel gene fragments in the rat genome

New gene fragments and their corresponding ESTs were identified using *in silico* tools and database searches as described previously [8, 9, 10]. To identify the related genomic, cDNA and EST clones, TBLASTN searches against nr and dbEST on the NCBI server (http://www.ncbi.nlm. nih.gov/blast) were performed using the amino acid sequence of human ETV6 (AC No: P41212) as the query sequence.

Structure and chromosomal localization of the novel gene

The exons and introns were determined based on the gene fragments identified by the TBLASTN program, by examining the corresponding ETS, consensus sequence of exon-intron boundaries (‘gt…ag’ rule of intronic sequence) [11], and the codon usage within the coding region. The chromosomal localization was defined by BLAST the predicted sequence to rat genome using the BLAST program of the NCBI (http://www.ncbi.nlm.nih.gov).

Deducing and analyzing the amino-acid sequence of ETV6 gene

Prediction of coding region as well as translation into amino acid sequence was performed using NCBI’s ORF Finder program (www.ncbi.nlm.nih.gov/gorf/). The homologous protein sequence searches were carried out using the BLASTP program against the protein databases. Amino acid sequence alignments were performed using the ClustalW program (http://www.ebi.ac.uk/clustalw). Domain structure of the novel protein was searched for with the RPS-BLAST program (http://www.ncbi.nlm.nih.gov). Sub-cellular localization of novel protein was searched for with the PSORT II program (http://psort.ims.u-tokyo.ac.jp). The phylogenetic tree was constructed using the neighbor-joining method (NJ) with Jones-Taylor-Thomton (JTT) distances. NJ searches were conducted using the MEGA3 software program [12]. The reliability of internal branches was assessed using 500 bootstrap replicates.

RESULTS

Identification and chromosomal localization of rat ETV6 gene

Rat genome sequences and rat ESTs homologous to human ETV6 were searched for with the TBLASTN program using the amino acid sequence of human ETV6 as a query sequence. The result showed segmented fragments of rat ETV6 gene on the genomic contig NW_047696 on the chromosome 4 at the nucleotide positions 24,013 kb-24,255 kb. Its precise exon-intron boundaries were determined by examining the ETS on the region (Table 1), the consensus sequences of exon-intron boundaries, and the codon usage within the coding region. The rat *ETV6* gene consisted of 8 exons spanned on a 242 kb genomic DNA (Figure.1). Its cDNA is predicted to be 2041 nucleotide in length.

Table 1. The accession number of ESTs, located on the region 24,013 kb-24,255 kb. of the genomic contig NW 047696, used to construct the rat ETV6 CDS.

BE100111.1	AW251598.1	BF566629.1	BQ189876.1	CO554966.1	BM390643.1
CR460593.1	CB800729.1	AW529320.1	CA505366.1	CO567974.1	BF406537.1
AI044374.1	BI302814.1	CR459037.1	BM384458.1	BF550302.1	BE117896.1
AI714333.1	CB812822.1	CO389595.1	CV076962.1	BM387293.1	CR459969.1
AW254091.1	CB812194.1	CB697939.1	CV118617.1	AW914042.1	AI231434.1
AW253517.1	CB791184.1	BF285055.1	AI547515.1	AW142730.1	AI237182.1
AW251139.1	CV797495.1	BF546884.1	BQ782515.1	CA507645.1	AI013170.1
AW252062.1	BF542553.1	BF524331.1	BF542699.1	BF548331.1	AW522926.1
AW251437.1	AW534087.1	BF523015.1	AI535043.1	BF550172.1	AA899087.1
AW253265.1	CR475211.1	AI556345.1	CB313494.1	BQ199604.1	CO575482.1
AI009381.1	AI547859.1	BM387000.1	CO387335.1	AI008524.1	CO563976.1

Exon No	Nucleotide sequence around exon-intron boundaries		
1	**5′-UTR**	TTTCGC..... ATTAAG	**GT**AACG...
2	... CAAC**AG**	CAGGAA..... ACCTGC	**GT**GAGT...
3	... CAAC**AG**	GTTTGC..... ATTCAG	**GT**GAGA...
4	... CTCC**AG**	GCGATG..... AAGAAG	**GT**ACTG...
5	... CCAC**AG**	ATAATT..... GCAAAG	**GT**TTGG...
6	... CTGC**AG**	ACTGTA..... CATAAG	**GT**AAAA...
7	... AAAC**AG**	AATAGG..... GTTCAG	**GT**AACA...
8	... CCTA**AG**	GTTCAT..... AAAAAA	**3′-UTR**

Figure 1. Exon–Intron boundaries of the rat ETV6 gene. On each line the end of an exon, the beginning and end of the following intron, and the start of the following exon is shown.

Analyses of deduced amino-acid sequence

The nucleotide sequence of rat ETV6 mRNA was determined by combining nucleotide sequences of its predicted exons. Rat ETV6 mRNA was found to consist of at least 14 bp of 5' UTR, 1428 bp of coding region**,** and a 599 bp 3' UTR region (Figure.2). The mRNA of the *ETV6* gene is predicted to encode a peptide of 476 amino acids.

```
tttcgctgtgagacatgtctgagactcctgctcagtgtagcattaagcaggaacggatttcatctacacccccagagagc    80
            M  S  E  T  P  A  Q  C  S  I  K  Q  E  R  I  S  S  T  P  P  E  S     22
ccagtggcgagctacggtccctcgactccacttcatgttccagtgcctcgggcgctcaggatggaggaagactcgatcca  160
P  V  A  S  Y  G  P  S  T  P  L  H  V  P  V  P  R  A  L  R  M  E  E  D  S  I  H    49
cctgccagcacacctgcgtttgcagcccatttactggagccgagatgacgtagcccagtggctcaaatgggcagaaaatg  240
 L  P  A  H  L  R  L  Q  P  I  Y  W  S  R  D  D  V  A  Q  W  L  K  W  A  E  N  E   76
agttttccttaaggcccatcgagagcaacacattcgaaatgaatggcaaggccctgctgctgctgaccaaagaggatttc  320
  F  S  L  R  P  I  E  S  N  T  F  E  M  N  G  K  A  L  L  L  L  T  K  E  D  F    102
cgctaccgatctcctcattcaggcgatgtgctctatgaactccttcagcatatcctgaagcagaggaagcctcgaattct  400
R  Y  R  S  P  H  S  G  D  V  L  Y  E  L  L  Q  H  I  L  K  Q  R  K  P  R  I  L   129
cttctcaccattcttccaccctgggaactctatccacaccaagccagaggtcctactgcatcagaaccatgaagaagata  480
 F  S  P  F  F  H  P  G  N  S  I  H  T  K  P  E  V  L  L  H  Q  N  H  E  E  D  N  156
attgtgtccagaggacgcccaggacgccagcggagagcctgcaccacaaccctcccactattgaactcttacatcgcccc  560
  C  V  Q  R  T  P  R  T  P  A  E  S  L  H  H  N  P  P  T  I  E  L  L  H  R  P    182
aggtcacccatcaccacaaaccacaggccttctccggacccagagcagcagcggcccctgcggtccccccctggacaacat  640
R  S  P  I  T  T  N  H  R  P  S  P  D  P  E  Q  Q  R  P  L  R  S  P  L  D  N  M   209
gatccgccgcctctcgcctgcagagagagcccaggggccaaggctacagcaggaaaacaaccaccaggaatcctaccccc  720
 I  R  R  L  S  P  A  E  R  A  Q  G  P  R  L  Q  Q  E  N  N  H  Q  E  S  Y  P  L  236
tgtcagtgtctcctatggagaataatcactgcccaccgtcctcggagtccaacccgaagccctcaagcccctggcaggag  800
  S  V  S  P  M  E  N  N  H  C  P  P  S  S  E  S  N  P  K  P  S  S  P  W  Q  E    262
agcacacgagtgatccagctgatgcctagccccatcatgcacccctgatcctgaacccccggcactcggtagatttcaa  880
S  T  R  V  I  Q  L  M  P  S  P  I  M  H  P  L  I  L  N  P  R  H  S  V  D  F  K   289
acagtcccgaatctctgaagatgggatgcatcgggaagggaagcccatcaacctgtctcaccgagaggacctggcttaca  960
 Q  S  R  I  S  E  D  G  M  H  R  E  G  K  P  I  N  L  S  H  R  E  D  L  A  Y  M  316
tgaaccacatcatggtctctgtgtccccacccgaagagcacgccatgcccattggaagaatagcaggtgagtgggctcag 1040
  N  H  I  M  V  S  V  S  P  P  E  E  H  A  M  P  I  G  R  I  A  G  E  W  A  Q    342
cccccagctcctcgctgcccagcggcagccatgaccatgttatgtgcccgcaaagactgtagactgctttgggattacgt 1120
P  P  A  P  R  C  P  A  A  A  M  T  M  L  C  A  R  K  D  C  R  L  L  W  D  Y  V   369
ctaccagttgctctctgacagccggtacgaaaacttcatccgatgggaggacaaggaatccaaaatattccggatagtgg 1200
 Y  Q  L  L  S  D  S  R  Y  E  N  F  I  R  W  E  D  K  E  S  K  I  F  R  I  V  D  396
atcccaacggactggctcgactctggggaaaccataagaataggacaaacatgacctatgagaaaatgtccagagctctc 1280
  P  N  G  L  A  R  L  W  G  N  H  K  N  R  T  N  M  T  Y  E  K  M  S  R  A  L    422
cgtcactactacaaactaaacatcatcaggaaggagcccggacaaaggcttttgttcaggttcatgaaaaccccggatga 1360
R  H  Y  Y  K  L  N  I  I  R  K  E  P  G  Q  R  L  L  F  R  F  M  K  T  P  D  E   449
gatcatgagtggccggacagaccgtctagaacacctcgagtctcaggagctggatgaacaagcgtaccaagaggatgaat 1440
 I  M  S  G  R  T  D  R  L  E  H  L  E  S  Q  E  L  D  E  Q  A  Y  Q  E  D  E  C  476
gttaagggaacctaccacagcctcagcggtgggctggccaagacgagacctgcccacagggaccgcaggaggcatgatgg 1520
  *
agcaggcaagctagggatggcttgaaaggaagagacccaggaatagcaggaacacttctcctgcagatcaagagggaccc 1600
agagcaccttagacaagccacccagcaatggcagggctggaattctggcggagggcacaagcctgagacacacgtcatgt 1680
ttgctcctccttcgactctcttgtctgtaaagcctcaccctcaccctgcacctgttgttagtctcatggtgtttctggtt 1760
ttgttttcgtttctttttttttttaagaacatgcagtttgactattcattgttcatacagggaagacatcacatgttgct 1840
ttcctatggaaatatatctattatatatattattaatttttttttcgttgttgcaaatctcaccaagtacgaccagctct 1920
gctggtcaggaaaaagaaaacttgcagaagagatcaggttcctcttttcctgccgcatagatctggtttgtcctatcca 2000
agtcaggtctttgatgagaaaggaaaacaaaaaaaaaaaaa                                        2041
```

Figure 2. Predicted rat ETV6 mRNA and rat ETV6 protein. Nucleotide sequence of ETV6 mRNA and its encoding amino acid sequences are shown. Nucleotides and amino acids are numbered on the right side, respectively.

The BLASTp program revealed that rat ETV6 protein has 90%, 90%, 88%, 88%, 87%, 81%, 69%, 52% and 51% total amino acid identity with homologous sequences of *M. mulatta* (XP_001083470), *H. sapiens* (P41212), *C. familiaris* (XP_543812), *B. taurus* (NP_001015514), *M. musculus* (NP_031987), *G. gallus* (AAC97200), *X. tropicalis* (NP_001016880), *T. rubripes* (AAK54061) and *D. rerio* (AAK49950), respectively. ClustalW alignment displayed a very high homology of these genes in these species (Figure.3A). RPS-BLAST results revealed that the rat ETV6 protein, like those of other organisms, had two conserved functional domains, SAM_PNT and ETS, spanning from amino acid 38 to 123 and 362 to 448, respectively. PSORT II results revealed a predicted nuclear localization signal at the amino acid position 124-127 (RKPR) (Figure.3A, B).

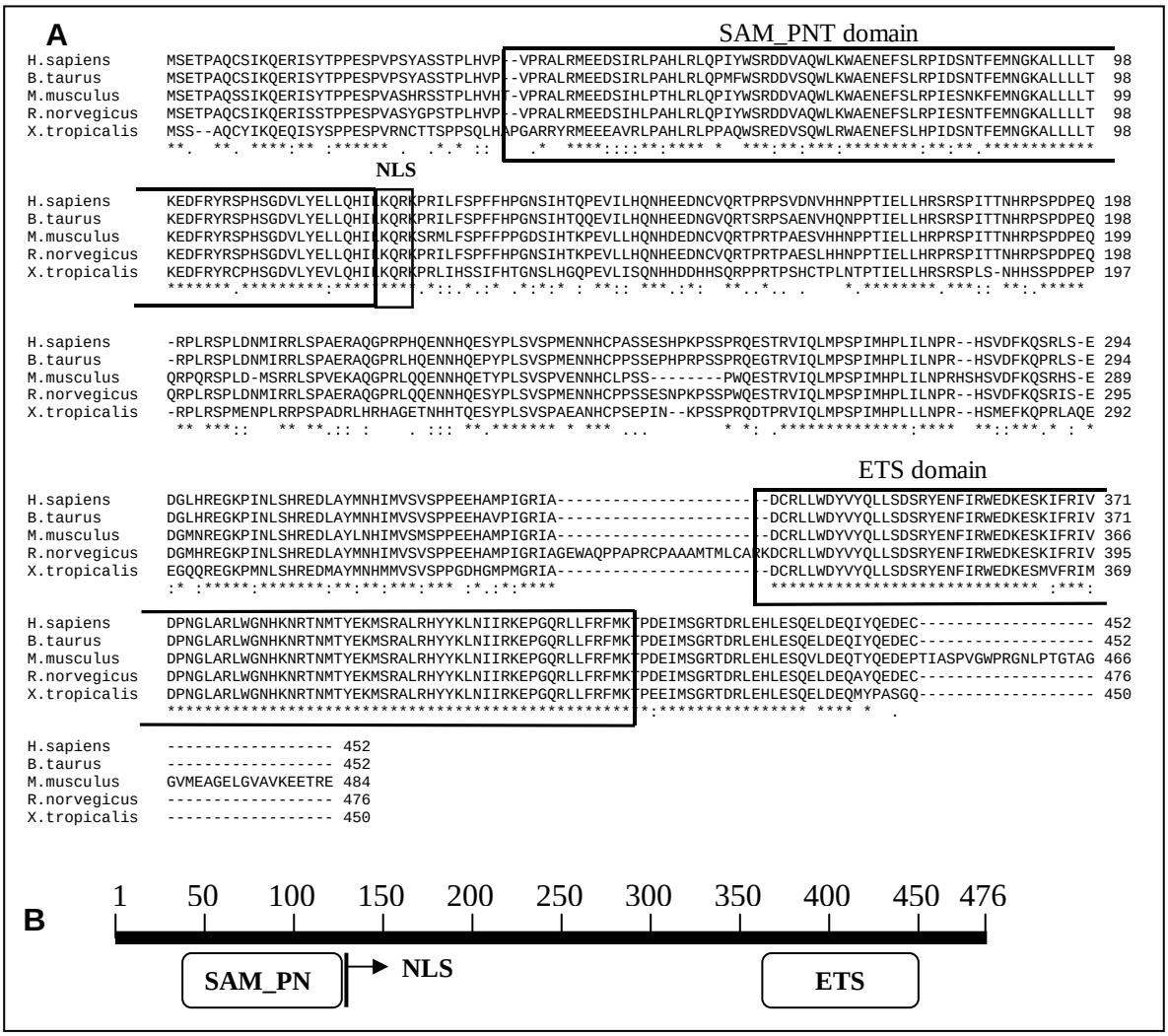

Figure 3. ClustalW alignment of ETV6 proteins from different species and their conserved domains. (A), Alignment of some homologous proteins to rat ETV6 protein. Amino acid residues are numbered on the right. Amino acid residues conserved among these species are shown by an asterisk below the alignment. All of these proteins include SAM_PNT and ETS domains and nuclear localization signal (NLS). (B), Schematic representation of rat ETV6 protein and its domains.

The phylogenetic tree of the ETV6 genes in different species was constructed using Mega3 program [12]. This tree indicates that the ETV6 gene is evolutionarily highly conserved among all organisms investigated. The most diverse are ETV6 genes of *T. rubribes* and *D. rerio* according to internal branch lengths of the tree (Figure 4).

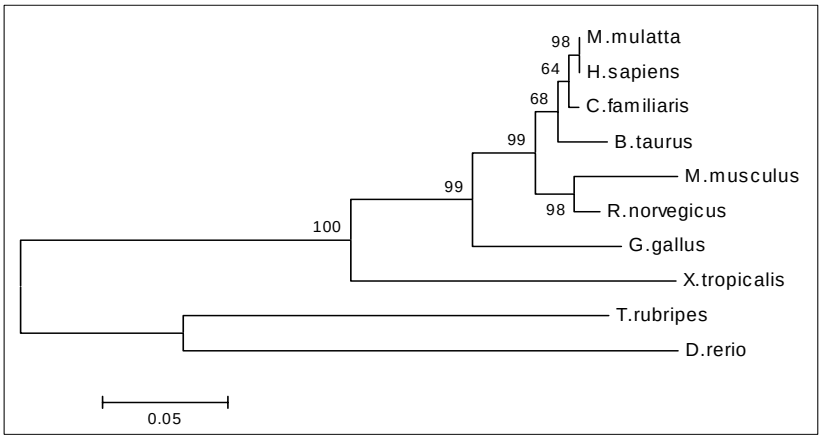

Figure 4. The phylogenetic tree of ETV6 proteins from different organisms. Branch lengths and species names are shown. Branch lengths indicate evolutionary relationship.

DISCUSSION

Bioinformatics have an important role in confirming laboratory-based work and in *de novo* analysis [13]**.** In this study, we identified and characterized the rat ETV6 gene using bioinformatic tools. It overlaps with the LOC312777 gene, recently predicted by the GNOMON, NCBI's gene prediction program. Even though both genes partially overlap, some overlapping ESTs seem to have been omitted when LOC312777 was predicted. Therefore, we think the LOC312777 gene may have an alternative splicing form. By taking into account these excluded ESTs, we predicted the structure of the rat ETV6 gene, its RNA and protein sequence. Nucleotide and amino acid sequences of LOC312777 transcript are predicted to be 1900 bp and 453 amino acids, respectively. Our predicted transcript isoform is 2041 bp and its encoded protein is 476 amino acids.

The rat ETV6 gene encodes for 476 amino acid peptide with an N-terminal PNT-SAM and a C-terminal ETS domain. These features suggest that ETV6 protein may function as a transcriptional factor. N-terminal PNT-SAM domain exists in more than 250 regulatory proteins including receptor tyrosine kinases, serin-threonin kinases, and transcription factors [14]. This domain is very likely to mediate in dimerization of ETV6. Homodimerization is essential to the oncogenic activation of ETV6 derived fusion proteins [2]. The ETS domain specifically recognizes DNA sequences that contain a GGAA/T core element [15] but it also involves in multiple protein-protein interactions [16].

Rat ETV6 gene was mapped to rat chromosome 4q.43. In human, studies have shown that chromosomal region of ETV6 gene (12p.13) is frequently implicated in both lymphoid and myeloid hematological malignancies [17]. The ETV6 gene is the most common target for rearrangements in 12p13 chromosomal region [18]. Characterization of the rat genes may provide a ground for genetic models to study the related human disorders.

The result of BLASTp clearly indicates that ETV6 proteins are evolutionarily well conserved. The phylogenetic tree indicates that the mouse ETV6 is the closest to the rat ETV6 gene.

Data mining techniques are powerful in terms of discovering novel genes and analyzing them [19]. Even though the results from bioinformatics studies are very helpful in directing and designing the experiments, they generally need to be experimentally confirmed.

REFERENCES

[1]. Poirel H, Oury C, Carron C, Duprez E, Laabi Y, Tsapis A, Romana SP, Mauchauffe M, Le Coniat M, Berger R, Ghysdael J, Bernard OA. 1997. The TEL gene products: nuclear phosphoproteins with DNA binding properties. Oncogene. 14: 349–357.

[2]. Lopez RG, Carron C, Oury C, Gardellin P, Bernard O, Ghysdael J. 1999. TEL Is a Sequence-specific Transcriptional Repressor**.** J Biol Chem. 274 (42): 30132–30138.

[3]. Kim CA, Phillips ML, Kim W, Gingery M, Tran HH, Robinson MA, Faham S, Bowie JU. 2001. Polymerization of the SAM domain of TEL in leukemogenesis and transcriptional repression. EMBO J. 20: 4173–4182.

[4]. Arai H, Maki K, Waga K, Sasaki K, Nakamura Y, Imai Y, Kurokawa M, Hirai H, Mitani K. 2002. Functional regulation of TEL by p38-induced phosphorylation. Biochem Biophys Res Commun. 299 (1): 116-125.

[5]. Bohlander SK. 2005. ETV6: a versatile player in leukemogenesis. Semin Cancer Biol. 15 (3): 162-174.

[6]. Gilliland DG, Jordan CT, Felix CA. 2004. The molecular basis of leukemia. Hematology. 80-97.

[7]. Oikawa T. 2004. ETS transcription factors: Possible targets for cancer therapy. Cancer Sci. 95(8): 626-633.

[8]. Katoh M. 2002. Paradigm shift in gene–finding method: From bench-top approach to desktop approach. Int J Mol Med. 10: 677-682.

[9]. Wei L, Liu Y, Dubchak I, Shon J, Park J. 2002. Comparative genomics approaches to study organism similarities and differences. J Biomed Inform. 35(2): 142-150.

[10]. Varisli L, Cen O. 2005. Identification and Characterization of Rat GMDS Gene by Using Bioinformatics Tools. Turk J Biochem. 30: 306-309.

[11]. Mount SM. 1982. A catalogue of splice junction sequences. Nucleic Acids Res. 10: 459-472.

[12]. Kumar S, Tamura K. Nei M. 2004. MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform. 5: 150-163.

[13]. Gale CP, Grant PJ. 2004. The characterization and functional analysis of the human glyoxalase-1 gene using methods of bioinformatics. Gene. 340: 251–260.

[14]. Schultz J, Ponting CP, Hofmann K, Bork P. 1997. SAM as a protein interaction domain involved in developmental regulation. Protein Sci. 6(1): 249-253.

[15]. Dittmer J. 2003. The biology of the Ets1 proto-oncogene. Mol Cancer. 2: 29-49.

[16]. Sharrocks AD. 2001. The ETS-domain transcription factor family. Nat Rev Mol Cell Biol. 2(11): 827-837.

[17]. Vieira L, Marques B, Cavaleiro C, Ambrosio AP, Jorge M, Neto A, Costa JM, Junior EC, Boavida MG. 2005. Molecular cytogenetic characterization of rearrangements involving 12p in leukemia. Cancer Genet Cytogenet. 157(2): 134-139.

[18]. Odero MD, Carlson K, Lahortiga I, Calasanz MJ, Rowley JD. 2003. Molecular cytogenetic characterization of breakpoints in 19 patients with hematologic malignancies and 12p unbalanced translocations. Cancer Genet Cytogenet. 142: 115-119.

[19]. Hu Z, Chen K, Wang L, Yao Q. 2005. Identification and characterization of *Bombyx mori* eIF5A gene through bioinformatics approaches**.** In Silico Biol. 5: 573-580.