

## Letter to the Editor

### 5' Genomic Structure of Human Sp3

Marjorie F. Oleksiak and Douglas L. Crawford

School of Biological Sciences, University of Missouri–Kansas City

Comparative sequence analysis among evolutionarily related genes can provide better insights into gene structures and functions compared with methods that rely solely on canonical rules (e.g., open reading frames, intron positions, or Kozak sequences). Presented here is a comparative sequence analysis of human Sp3 that provides the missing N-terminal portion.

Evolutionarily, the transcription factors Sp1 and Sp3 are related more closely to each other than to other members of the Sp family (Kolell and Crawford 2002), and numerous studies have examined the roles of Sp1 and Sp3 in transcriptional regulation. The Sp transcription factor family members interact with general transcriptional machinery, cell cycle regulators such as retinoblastoma protein, and other transcription factors (Pugh and Tjian 1990; Seto, Lewis, and Shenk 1993; Gill et al. 1994; Udvardia, Templeton, and Horowitz 1995; Kennett, Moorefield, and Horowitz 2002). Sp3 and Sp1 belong to this family of transcription factors and compete for the same binding sites (Hagen et al. 1992; Kingsley and Winoto 1992). Much research has been published because of the widespread dependence of transcription on Sp transcription factors. For example, almost all mammalian TATA-less promoters have multiple Sp-binding sites that have been found to be required for maximal activity (Azizkhan et al. 1993).

Sp1 and Sp3 are well-studied transcription factors. But virtually all studies on Sp3 regulation of gene expression, its interactions with Sp1, or its repression of transcription have used a truncated form of Sp3, a form lacking the N-terminal domain. Thus, the N-terminal portion of the A domain is missing. This domain is critical for protein-protein interactions (Pascal and Tjian 1991). The problem has been the difficulty in cloning the 5' end of the mRNA: several laboratories over the last 10 years have failed to isolate and characterize the full-length cDNA on the basis of the fact that no published Sp3 sequence has an initiating methionine as is found in the closely related Sp1 and Sp4 genes (Kolell and Crawford 2002). In addition, the predicted protein of the original published Sp3 sequence was approximately 80 amino acids shorter than Sp1 or Sp4; yet, the empirical protein's molecular mass is approximately the same in mammalian cell lines (Hagen et al. 1994). Reported here is the 5' human Sp3 sequence including the 5' untranslated sequence of human Sp3 and the first, second, and third introns and exons. The third exon con-

tains nucleotides from the previously published Sp3 sequence.

A combination of genome walking and database mining was used to determine the 5' end of the human Sp3 gene (GenBank: AF494280). Genomic DNA was isolated from human white blood cells using an ammonium hydroxide extraction (Wirgin et al. 1990). Genome walking was done using nested PCR techniques including rapid amplification of genomic DNA ends (RAGE) (Mizobuchi and Frohman 1993), the Human GenomeWalker kit (Clontech, Palo Alto, Calif.), and inverse PCR (Innis et al. 1990, pp. 482). RAGE and the GenomeWalker kit both rely on a linker ligated to the genomic DNA. For RAGE, this linker uses the multiple cloning site of a vector. The vector is cut with a variety of restriction enzymes, and the genomic DNA, cut with the same enzyme, is ligated to the vector DNA. Nested, vector-specific primers are used with nested, gene-specific primers and the polymerase chain reaction to amplify the genomic sequence 5' to the gene-specific primers. The GenomeWalker kit uses a synthetic linker ligated to the ends of cut genomic DNA (rather than a vector sequence), but otherwise the approach is the same. Inverse PCR relies on cut genomic DNA self-ligating to form a circle under conditions favoring such self-ligation (i.e., low concentrations of DNA) and a set of nested, gene-specific primers. Sp3-specific primers used with these techniques are underlined in the supplementary material on the SMBE website ([www.molbioevol.org](http://www.molbioevol.org)). In addition, primers to established Sp3-coding sequences (5'tgtggctgacaaaacctcccatc and 5'gcagcacttggaaatctggactaga) corresponding to the reverse complement of nucleotides 8,983–9,005 and 9,035–9,058 (relative to nucleotide positions in the supplementary material) were used for the first genomic walking step. Fifteen successful walking steps were needed to determine the 5' end of the human Sp3 sequence.

PCR products were cloned into pGEM vector (Promega, Madison, Wis.) and sequenced using BigDye reaction mix (Perkin-Elmer, Foster City, Calif.). This sequence was used to query GenBank, resulting in the identification of a similar genomic sequence and the expressed sequence tags (ESTs) shown in figure 1. Sequences used for clustering included a human cDNA identified as IMAGE: 2115371 3' similar to SW:SP2\_HUMAN Q02086 TRANSCRIPTION FACTOR SP2 (accession number AI474184), an unidentified rat cDNA (accession number BF388711) in the opposite orientation, *Gallus gallus* cDNA for Sp3 (accession number AJ317961), a partial mouse Sp3 mRNA (accession number AF062567), a portion of *Homo sapiens* BAC clone RP11-394I13 from chromosome 2 (accession number AC016737), mouse Sp4 (accession number NM\_009239), human Sp4 (acces-

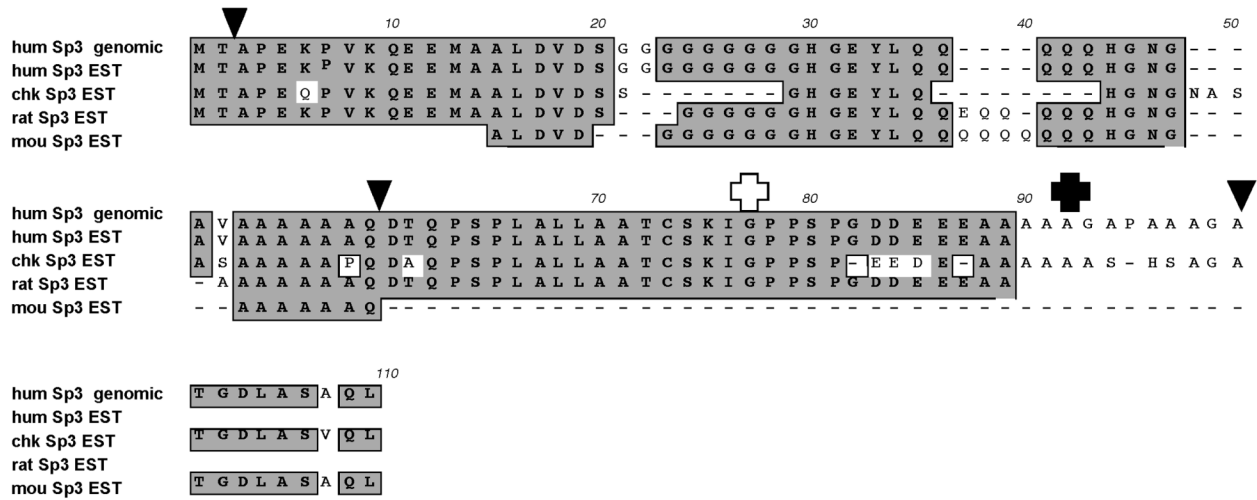
Key words: human, Sp3, Sp1, genomic structure, transcription.

Address for correspondence and reprints: Douglas L. Crawford, School of Biological Sciences, University of Missouri–Kansas City, 5007 Rockhill Road, Kansas City, Missouri 64110. E-mail: [crawforddo@umkc.edu](mailto:crawforddo@umkc.edu).

*Mol. Biol. Evol.* 19(11):2026–2029. 2002

© 2002 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

(A)



(B)

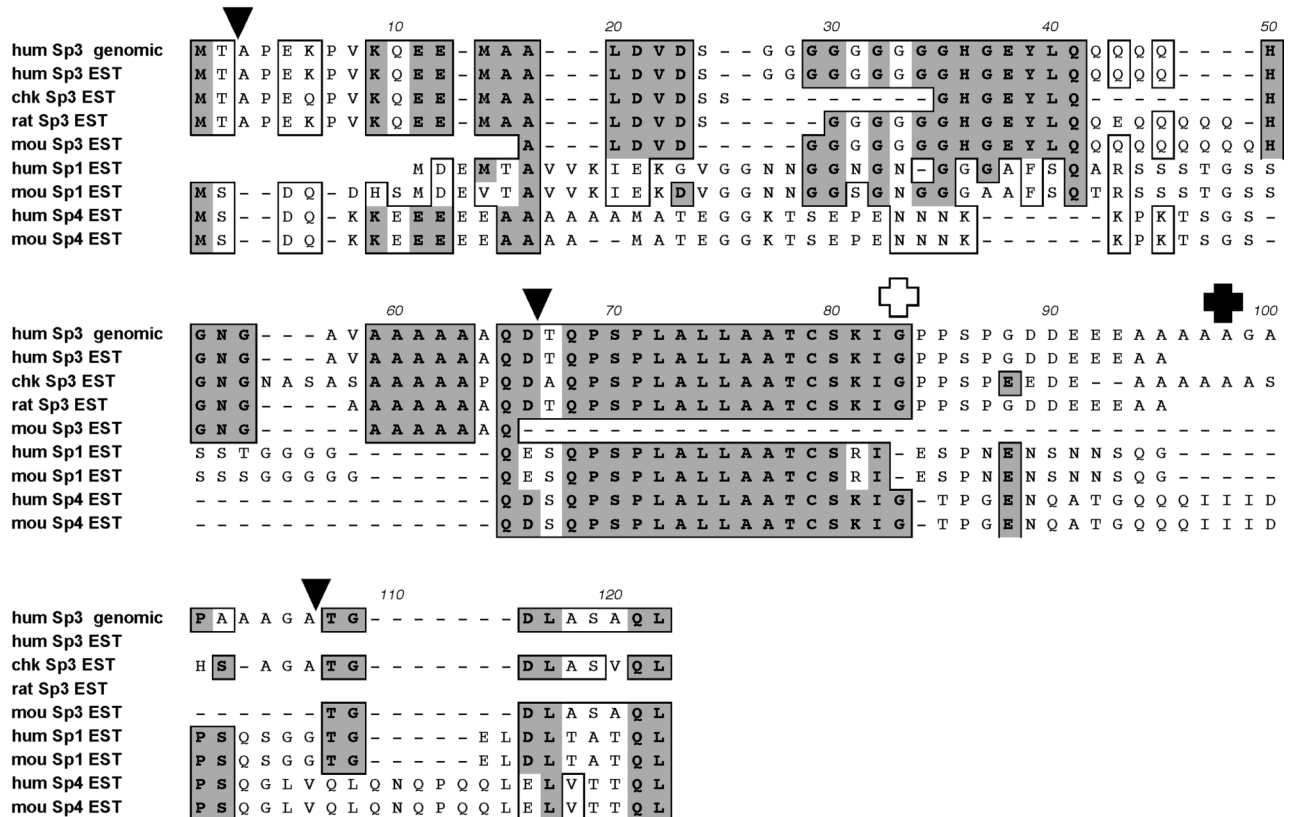


FIG. 1.—Sp3 and Sp1 amino-terminal alignments. A, Sp3 deduced protein alignment: human (hum), chicken (chk), rat, and mouse (mou). Boxes enclose similar amino acids. Shaded areas denote identical amino acids. Wedges indicate intron positions. The filled cross indicates the originally (Hagen et al. 1992) and the plain cross indicates the recently (Hernandez et al. 2002) defined 5' ends of human Sp3. B, Sp3, Sp1, and Sp4 deduced protein alignment. Marks are the same as in (A).

sion number NM\_003112), human Sp1 (accession number XM\_028606), and mouse Sp1 (accession number NM\_013672).

Clustering of deduced amino acid sequences were performed in Clustal (Higgins, Bleasby, and Fuchs 1992) and, for figure 1B, adjusted by eye to maximize chemical similarities among amino acids, to conserve amino acid alignments at exon-intron borders, and to minimize the number of derived characters. The adjusted alignment produced a maximum parsimony tree (exhaustive search, gaps treated as missing) that was 24 steps shorter than the original Clustal alignment (122 vs. 146 steps).

The supplementary material on the SMBE website ([www.molbioevol.org](http://www.molbioevol.org)) shows sequences from the first three introns and exons of Sp3 (AF494280). The third exon overlaps the same sequence found in previously identified Sp3 cDNA. The splicing junctions between introns and exons all conform to the GT-AG rule (Breathnach and Chambon 1981). The length of the entire coding sequence is 2,346 bp (~3.9 kb with 5'- and 3'-untranslated sequences), and the calculated molecular weight of the resulting deduced protein is about 82 kDa. This is similar to the calculated molecular weights of human Sp1 and Sp4 proteins, which are 80 and 82 kDa, respectively. The positions of the first and second introns are identical to these positions in mouse Sp4 (Supp et al. 1996). The third intron in Sp3 is not present in Sp4, but the fourth intron in Sp3 is in the same position as the third intron in Sp4 (data not shown). The first methionine (coding nucleotides 1–3) has an A in the position –3, which is sufficient for most ribosomes to select it irrespective of the rest of the surrounding sequence (Kozak 1991). The second methionine (coding nucleotides 37–39) has a G three base pairs upstream (position 34) and a G four base pairs downstream (position 43), suggesting that this AUG could be used to initiate translation.

Support that this is the correct identification of the 5' end of human Sp3 is based on (1) an initiation methionine codon, (2) the similarity to the genomic structure of Sp4, (3) sequence similarities among human and rat ESTs (fig. 1A), (4) similarities among the deduced amino acid sequences of chicken and mouse Sp3's, (5) similarities in protein length, and (6) regions of sequence similarity to other members of the Sp family (fig. 1B: amino acids 1–15, glycine-rich regions in exon 2, and exon 3 between amino acids 65 and 83).

Since they were first described, Sp3 and Sp1 have been shown to affect a wide variety of mammalian genes in transfection assays. Much can be learned about transcriptional control using transfection assays: they address binding site strength, protein-protein interactions, and transcriptional activation and repression. But although the full-length Sp1 sequence has been published for mammals, the full-length Sp3 sequence has not been identified. Incomplete Sp3 sequences were first published in 1992 (Hagen et al. 1992; Kingsley and Winoto 1992). Kingsley and Winoto (1992) identified putative non-AUG start sites, and Kennett, Udvardia, and Horowitz (1997) have identified two truncated Sp3 isoforms

produced by internal translational initiation that, in contrast to full-length Sp3, act as potent transcriptional repressors. A recent publication suggests another non-AUG start site initiated *in vitro* (Hernandez et al. 2002). However, gel-shift assays with mammalian nuclear extracts suggest that endogenous Sp3 is similar in size to Sp1 (though this is uncertain due to known posttranslational modifications of Sp1), and Northern analyses show a transcript of 4.2 kb in a variety of mammalian cell lines (Kingsley and Winoto 1992). These data and the fact that the closely related Sp1 and Sp4 (Kolell and Crawford 2002) both initiate with a methionine and have roughly 70–90 more amino acids at their 5' ends suggest that full-length Sp3 transcripts also would code for a larger protein and initiate with a methionine. Thus, it is likely that no transient transfection studies using human Sp3 sequence to examine transcriptional regulation of human genes have used the full-length Sp3 sequence. In fact, they have been completed without the first three exons of Sp3. The 5' end of Sp1 is thought to be important in protein-protein interactions that affect transcription (Pascal and Tjian 1991; Murata et al. 1994). Because Sp3 belongs to the same family of proteins, the 5' end of Sp3 might have similar transcriptional importance.

Although no Sp3 sequence has been published with a 5' methionine, the Sp3 gene has been mapped to chromosome 2 (Kalff-Suske et al. 1996). Blast searches with putative intron and exon sequences identified by genome walking from known Sp3 sequence match *H. sapiens* BAC clone RP11-394I13 from chromosome 2. These are contiguous with published Sp3 sequences. Blast searches also match unidentified rat and human ESTs in the database and a chicken Sp3 (accession number AJ317961). All these code for similar amino-terminal ends containing methionines (fig. 1A). The sequence also is similar to an incomplete mouse Sp3 (Supp et al. 1996). Mouse Sp3 sequence lacks approximately fourteen 5' amino acids and the entire third exon. But it does have the fourth exon, suggesting that mouse Sp3 might be alternatively spliced. These sequences are similar to Sp1 and Sp4 proteins (fig. 1B) with respect to length and protein sequences.

In summary, the identification of this sequence as the full-length Sp3 sequence is supported for the following reasons. First, genome walking and database searching has identified contiguous sequence upstream of the published Sp3 sequence. Exons identified in this sequence code for a protein more similar in size to Sp1, as is expected from gel-shift assays with endogenous Sp1 and Sp3 proteins. This protein has a calculated molecular weight of 82 kDa, similar to the calculated molecular weight of human Sp1 (80 kDa), and the deduced transcript (3.9 kb) is more similar in size to the Sp3 transcript (4.2 kb), based on Northern analyses. Sequence similarities with amino termini for Sp1 and Sp4 sequences also suggest that exons identified in this sequence code for the amino terminus of Sp3. In particular, this sequence has an initiating methionine with a favorable Kozak sequence. This sequence also has an intron-exon structure, similar to that of Sp4. Notably,

the first exons in both proteins are only seven nucleotides long, coding for only two amino acids. Finally, matches in the database to chicken Sp3 and unidentified human and rat ESTs suggest that this transcript is expressed *in vivo*.

The intron and exon sequence for the 5' end of human Sp3 is shown in the supplementary material on the SMOBE website ([www.molbioevol.org](http://www.molbioevol.org)). This sequence has been deposited in GenBank and assigned the accession number AF494280.

### Acknowledgments

This research benefited from a discussion with Kevin Kolell and was supported by NSF Bioinformatics postdoctoral fellowship 0074520 to M.F.O. and NSF IBN grant # 9986602 to D.L.C.

### LITERATURE CITED

- AZIZKHAN, J. C., D. E. JENSEN, A. J. PIERCE, and M. WADE. 1993. Transcription from TATA-less promoters: dihydrofolate reductase as a model. *Crit. Rev. Eukaryot. Gene Expr.* **3**:229–254.
- BREATHNACH, R., and P. CHAMBON. 1981. Organization and expression of eukaryotic split genes coding for proteins. *Annu. Rev. Biochem.* **50**:349–383.
- GILL, G., E. PASCAL, Z. H. TSENG, and R. TJIAN. 1994. A glutamine-rich hydrophobic patch in transcription factor Sp1 contacts the dTAFII110 component of the Drosophila TFIID complex and mediates transcriptional activation. *Proc. Natl. Acad. Sci. USA* **91**:192–196.
- HAGEN, G., S. MULLER, M. BEATO, and G. SUSKE. 1992. Cloning by recognition site screening of two novel GT box binding proteins: a family of Sp1 related genes. *Nucleic Acids Res.* **20**:5519–5525.
- . 1994. Sp1-mediated transcriptional activation is repressed by Sp3. *EMBO J.* **13**:3843–3851.
- HERNANDEZ, E. M., A. JOHNSON, V. NOTARIO, A. CHEN, and J. R. RICHERT. 2002. AUA as a translation initiation site *in vitro* for the human transcription factor Sp3. *J. Biochem. Mol. Biol.* **35**:273–282.
- HIGGINS, D. G., A. J. BLEASBY, and R. FUCHS. 1992. Clustal V improved software for multiple sequence alignment. *Comput. Appl. Biosci.* **8**:189–191.
- INNIS, M. A., D. H. GELFAND, J. J. SNINSKY, and T. J. WHITE. 1990. PCR protocols: a guide to methods and applications. Volume 1. Academic Press, San Diego.
- KALFF-SUSKE, M., J. KUNZ, K. H. GRZESCHIK, and G. SUSKE. 1996. Human Sp3 transcriptional regulator gene (SP3) maps to chromosome 2q31. *Genomics* **37**:410–412.
- KENNETT, S. B., K. S. MOOREFIELD, and J. M. HOROWITZ. 2002. Sp3 represses gene expression via the titration of promoter-specific transcription factors. *J. Biol. Chem.* **277**:9780–9789.
- KENNETT, S. B., A. J. UDVADIA, and J. M. HOROWITZ. 1997. Sp3 encodes multiple proteins that differ in their capacity to stimulate or repress transcription. *Nucleic Acids Res.* **25**:3110–3117.
- KINGSLEY, C., and A. WINOTO. 1992. Cloning of GT box-binding proteins: a novel Sp1 multigene family regulating T-cell receptor gene expression. *Mol. Cell. Biol.* **12**:4251–4261.
- KOLELL, K. J., and D. L. CRAWFORD. 2002. Evolution of Sp transcription factors. *Mol. Biol. Evol.* **19**:216–222.
- KOZAK, M. 1991. Structural features in eukaryotic mRNAs that modulate the initiation of translation. *J. Biol. Chem.* **266**:19867–19870.
- MIZOBUCHI, M., and L. A. FROHMAN. 1993. Rapid amplification of genomic DNA ends. *Biotechniques* **15**:214–216.
- MURATA, Y., H. G. KIM, K. T. ROGERS, A. J. UDVADIA, and J. M. HOROWITZ. 1994. Negative regulation of Sp1 trans-activation is correlated with the binding of cellular proteins to the amino terminus of the Sp1 trans-activation domain. *J. Biol. Chem.* **269**:20674–20681.
- PASCAL, E., and R. TJIAN. 1991. Different activation domains of Sp1 govern formation of multimers and mediate transcriptional synergism. *Genes Dev.* **5**:1646–1656.
- PUGH, B. F., and R. TJIAN. 1990. Mechanism of transcriptional activation by Sp1: evidence for coactivators. *Cell* **61**:1187–1197.
- SETO, E., L. LEWIS, and T. SHENK. 1993. Interactions between transcription factors Sp1 and YY1. *Nature* **365**:462–464.
- SUPP, D. M., D. P. WITTE, W. W. BRANFORD, E. P. SMITH, and S. S. POTTER. 1996. Developmental Sp4, a member of the Sp1-family of zinc finger transcription factors, is required for normal murine growth, viability, and male fertility. *Dev. Biol.* **176**:284–299.
- UDVADIA, A. J., D. J. TEMPLETON, and J. M. HOROWITZ. 1995. Functional interactions between the retinoblastoma (Rb) protein and Sp-family members: superactivation by Rb requires amino acids necessary for growth suppression. *Proc. Natl. Acad. Sci. USA* **92**:3953–3957.
- WIRGIN, I. I., M. D'AMORE, C. GRUNWALD, A. GOLDMAN, and S. J. GARTE. 1990. Genetic diversity at an oncogene locus and in mitochondrial DNA between populations of cancer-prone Atlantic tomcod. *Biochem. Genet.* **28**:459–475.

CLAUDIA KAPPEN, reviewing editor

Accepted July 17, 2002