ELSEVIER

Clinical Biochemistry, Vol. 31, No. 8, 687–688, 1998

0009-9120/98 $19.00 + .00

PII S0009-9120(98)00073-3

A Candidate New Gene on Human Chromosome 5q12 Contains a Motif That Is Found in Transcriptional Co-Activators

KATERINA ANGELOPOULOU,[1] CATHY PRODY,[2,3] and ELEFTHERIOS P. DIAMANDIS[1,2]

[1]Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, Toronto, Ontario, Canada, [2]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada, and [3]Skye Pharma Tech Inc., Mississauga, Ontario, Canada

Introduction

During our studies on prostate specific antigen (PSA) gene expression in lung carcinomas we have cloned and sequenced a 450 base pair (bp) novel sequence (1). Through genomic library screening and subcloning, we were able to extend the length of this sequence to 4.8 kilobases (kb). This entire sequence, which resides on chromosome 5q12, is now deposited in GenBank (accession number AF038385). In this article, we describe analysis of this genomic sequence with various computer programs and show that it contains an open reading frame encoding a proline and leucine-rich protein. The protein sequence contains a motif, LXXLL (L: leucine; X: any amino acid), that was found in the sequence of transcriptional co-activators (2). This is an important new family of nuclear proteins, which co-operate with steroid hormone receptors to regulate gene expression. The presence of this motif and the limited homology of the putative protein with another nuclear protein, which is a transcription factor, allow us to speculate that the newly cloned sequence is part of a gene encoding either a nuclear transcription factor or a transcriptional co-activator.

Materials and methods

In this report, all nucleotide numbers relate to GenBank accession number AF038385 (nucleotides 1 to 4809). This 4.8 kb novel sequence was first subjected to the program Repeat Masker 2W, available at the Web site of the Washington University, St. Louis, MO, USA, in order to identify known DNA repeats. The repeats were masked from any additional homology searching and open reading frame analysis. The masked sequence was then subjected to homology searching by using the BLAST software (3). Open reading frames were identified by using the GENE-ID exon prediction program available by the BioMolecular Engineering Research Center, Boston University (geneid@darwin.bu.edu) (4) and the FGENEH program available at the Web site of the Baylor College of Medicine (http://mbcr.bcm.tmc.edu) (5).

Results

Homology searching

Initial analysis of the 4.8 kb sequence revealed a number of repeats at positions 2-274 (MER 6), 1411-1707 (AluSq), 1719-1747 (AT rich), 1788-2101 (MLT1A2), and 2228-3028 (L2). These sequences were masked from additional homology searching analysis.

BLAST homology searching of the remaining DNA sequence with GenBank deposited sequences revealed no major homologies with the exception of short stretches, which did not exceed 28 nucleotides (data not shown). Therefore, we concluded that the 4.8 kb genomic DNA represents a novel sequence.

Open reading frames

Putative open reading frames were searched using two computer programs (4,5). One putative open reading frame was identified by both GENE-ID and FGENEH programs from nucleotides 2581–2845, yielding the predicted aminoacid sequence shown in Figure 1. An SLL motif was identified four times in this frame and notably, this putative protein sequence is proline-rich (16/88 aminoacids are prolines) and leucine-rich (15/88 aminoacids are leucines). Protein homology searching using BLAST

Correspondence: Dr. E.P. Diamandis, Dept. of Pathology & Laboratory Medicine, Mount Sinai Hospital, 600 University Ave., Toronto, ON M5G 1X5, Canada. E-mail: ediamandis@mtsinai.on.ca

Manuscript received June 24, 1998; revised and accepted August 24, 1998.

```
5′ –  TCC  CTC  TTG  ACC  ACA  TGT  GCC  ATA  TCC  TGG
      S    L    L    T    T    C    A    I    S    W
      CCC  ACC  TCC  TTG  CTG  ATC  CGC  GAA  TAT  GGT
      P    T    S    L    L    I    R    E    Y    G
      TAT  GCT  CTT  GCC  TCA  GGG  CCT  TTG  CAC  ATT
      Y    A    L    A    S    G    P    L    H    I
      CCT  CTG  TCT  GAA  GCT  CTT  CCT  CCC  AAA  TGT
      P    L    S    E    A    L    P    P    K    C
      CCA  CAC  AAC  ACA  ACT  TAC  TCC  CTC  CTT  CCC
      P    H    N    T    T    Y    S    L    L    P
      TTG  GGG  ATT  TAC  TCC  ACT  GTC  ACC  CTT  CCA
      L    G    I    Y    S    T    V    T    L    P
      ATC  AGG  CCT  ACC  CTG  ACA  TCC  CTA  CTT  AAA
      I    R    P    T    L    T    S    L    L    K
      ATG  GTG  ATG  CCC  CCA  CCA  TTC  CCC  ATC  TCC
      M    V    M    P    P    P    F    P    I    S
      CCA  CCC  TTG  CCT  TAT  ATC  TTT  CAT  -    3′
      P    P    L    P    Y    I    F    H
```

Figure 1 — Translation of the nucleotide sequence shown, which was identified as an open reading frame. Aminoacids are depicted with the single letter code. The motif SLL is shown in bold and the motif LXXLL, present on known nuclear co-activators, is underlined. Sixteen and fifteen aminoacids out of a total of 88 aminoacids are prolines (P) and leucines (L), respectively.

software (4) revealed some restricted homologies with two proteins: (a) the retinoblastoma binding protein, RIZ (pir I38902) or transcription factor (prf/2206335A) or zinc-finger DNA binding domain (gn1/P1D/d100870). The homology spans within amionoacids 22–60 of our putative protein and the identity is 40%; (b) the enzyme arginine methyltransferase (gi/1710263). The homology is with aminoacids 5–61 of our putative protein and the identity is 35%. The sequence of Figure 1 contains the known motif LXXLL (aminoacids 75–79; X: any aminoacid), which was found to be part of transcriptional co-activators (2).

Discussion

In this report, we describe analysis of a novel 4.8 kb human genomic DNA sequence, which resides on chromosome 5q12. We first identified part of this sequence during our studies on PSA gene expression in lung carcinomas (1). After isolating a P1-derived artificial chromosome (PAC) clone from a human PAC library (6), we were able to extend the length of the original 450 bp sequence to 4809 nucleotides. The sequence, which has now been deposited in GenBank, was analyzed for open reading frames in order to examine if it encodes a novel protein or a protein of known structure or function. We here report preliminary analysis of this genomic sequence.

We have identified one open reading frame by two different exon prediction programs (4,5). The 88 aminoacid putative protein shown in Figure 1 is novel and it has a number of distinguishing features: (a) it has an SLL motif which appears 4 times; (b) it is proline and leucine rich; and (c) part of the putative protein has 40% identity with a transcription factor known as retinoblastoma binding protein and 35% identity with the enzyme arginine methyltransferase. An interesting feature of the putative protein sequence is the motif LXXLL (X: any aminoacid), which was found to be part of the sequence of a new class of proteins known as transcriptional co-activators (2,7). These new proteins cooperate with nuclear receptors to activate gene transcription. The presence of this motif allows us to speculate that the novel DNA sequence on chromosome 5q12 may encode a new transcriptional co-activator. We are now in the process of cloning the cDNA of this putative new gene and study its tissue expression and regulation.

Acknowledgements

We would like to thank S. Scherer for valuable discussions and C. Platt and J. Herbrick for technical assistance.

References

1. Zarghami N, Levesque M, D'Costa M, Angelopoulou K, Diamandis EP. Frequency of expression of prostate specific antigen mRNA in lung tumors. *Am J Clin Pathol* 1997; **108**: 184–90.
2. Heery DM, Kalhoven E, Hoare S, *et al*. A signature motif in transcriptional co-activators mediates binding to nuclear receptors. *Nature* 1997; **387**: 733–6.
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990; **215**: 403–10.
4. Guigo R, Kundsen S, Drake N, Smith T. Prediction of gene structure. *J Mol Biol* 1992; **245**: 45–56.
5. Solovyev VV, Salamov AA, Lawrence CB. Predicting internal exons by oligonucleotide composition and discriminant ananlysis of spliceable open reading frames. *Nucleic Acids Res* 1994; **22**: 5156–63.
6. Ioannou PA, Amemiya CT, Garnes J, *et al*. A new bacteriophage P1-derived vector for propagation of large human DNA fragments. *Nat Genet* 1994; **6**: 84–9.
7. Anzick SL, Kononen J, Walker RL, *et al*. AIB1, a steroid receptor coactivator amplified in breast and ovarian cancer. *Science* 1997; **277**: 965–8.