Volume 290, number 1,2, 159–161 FEBS 10188 September 1991

ADONIS 001457939100899C

Cloning and sequence analysis of a human type A/B hnRNP protein*

Farhat A. Khan[1], Anil K. Jaiswal[2] and Wlodzimierz Szer[1]

[1]*Department of Biochemistry* and [2]*Department of Cell Biology, New York University School of Medicine, New York, NY 10016, USA*

Received 10 June 1991; revised version received 25 July 1991

A cDNA encoding a 284 residue long type A/B hnRNP protein has been cloned. This protein, previously referred to as type C [(1987) J. Biol. Chem. 262, 17126–17137], is an RNA unwinding protein from HeLa 40S hnRNP with a high affinity for G- followed by U-rich sequences. The N-terminal part of the protein contains two consensus RNA binding domains present in a number of other RNA binding proteins. The C-terminal part is glycine-rich and contains a potential ATP/GTP binding loop. The distribution of charged amino acids is highly uneven and there are multiple potential phosphorylation sites.

hnRNP protein; RNA binding motif; Domain structure

1. INTRODUCTION

A large number of RNA binding proteins, including major core proteins of heterogeneous nuclear ribonucleoprotein particles (hnRNP) A1, A2/B1 and C1/C2, contain one or more loosely conserved putative 'RNA-binding domains', 80–90 amino acids in length. The most highly conserved segments within these domains are two consensus sequences, an octapeptide, RNP-CS1, and a hexapeptide, RNP-CS2 (see [1–3] for review and [4–7] for recent references). While there is no functional assay of hnRNP proteins, they are believed to play a role in the packaging and metabolism of hnRNA including pre-mRNA splicing [8,9]. We have previously described the purification and the RNA binding properties of a novel protein from HeLa 40S hnRNP [10]. This protein is less abundant than the major core proteins and was originally termed type C since it co-migrates on SDS-PAGE with hnRNP protein C2; it corresponds to protein IEF 48m,n of the HeLa protein catalogue [11]. Sequence analysis indicates that this protein is similar to the A and B hnRNP proteins rather than to the C proteins and its name has been changed accordingly. The binding of type A/B protein disrupts the residual secondary structure of RNAs. The protein has the highest affinity for G- followed by U-rich regions; it has little or no affinity for A, and the presence of C residues disrupts the binding [10]. In this paper we describe the cloning of type A/B protein cDNA and analyze the protein sequence in relation to analogous proteins.

*The sequence data in this paper have been submitted to GenBank/EMBL under accession number M65028.

Correspondence address: W. Szer, Department of Biochemistry, New York University School of Medicine, 550 First Avenue, New York, NY 10016, USA. Fax: (1) (212) 263 8166.

2. EXPERIMENTAL

Type A/B protein was purified to homogeneity from 40S HeLa hnRNP [10], digested with trypsin and the tryptic peptides were sequenced as previously described [12]. Two degenerate 17-mer oligonucleotides were used for screening of a human liver cDNA library in λgt11 and overlapping clones containing inserts of 420 (FK2), 600 (FK4) and 1280 bp (FK7) were subcloned into pUC 18 vectors [13] and sequenced by the dideoxy chain termination method [14]. The largest clones from this library (e.g. FK7) extended from bp 261 to bp 1537 (Fig. 1). To obtain cDNA sequences corresponding to the 5' end of the mRNA, a human breast carcinoma MCF-7 cDNA library in λgt11 was screened using a 17-mer probe selected from the 5' end of insert FK7. An overlapping clone (FK6) extending from bp 1 to the end of the coding sequence (Fig. 1) was isolated and sequenced. DNA sequences on the opposite strand were determined by synthesizing 17-mer primers for direct sequencing of the FK7 and FK6 clones. cDNA sequencing yielded deduced amino acid sequences which matched all tryptic peptides from different regions of the protein (a total of 79 amino acids) except that Ile^{274} in peptide was Arg in cDNA (Fig. 1). this microheterogeneity is probably due to the fact that the protein was isolated from HeLa cells and the cDNA libraries were from human liver and human breast carcinoma. Computer analysis of the sequence and protein motifs was carried out using the Genetics Computing Group package (Bestfit and Prosearch) on a Vax6000 computer.

3. RESULTS AND DISCUSSION

The 1537 bp cDNA sequence (Fig. 1) contains a single ORF coding for 284 amino acids. The first in-frame AUG, the presumed initiator site, is placed within a favorable initiation sequence with an A at position −3 and a G at +4 [15]. There is a single polyadenylation signal in the 3' untranslated region. Northern blot hybridization using HeLa RNA detects a single mRNA band of about 1.7 kb and Southern analysis shows

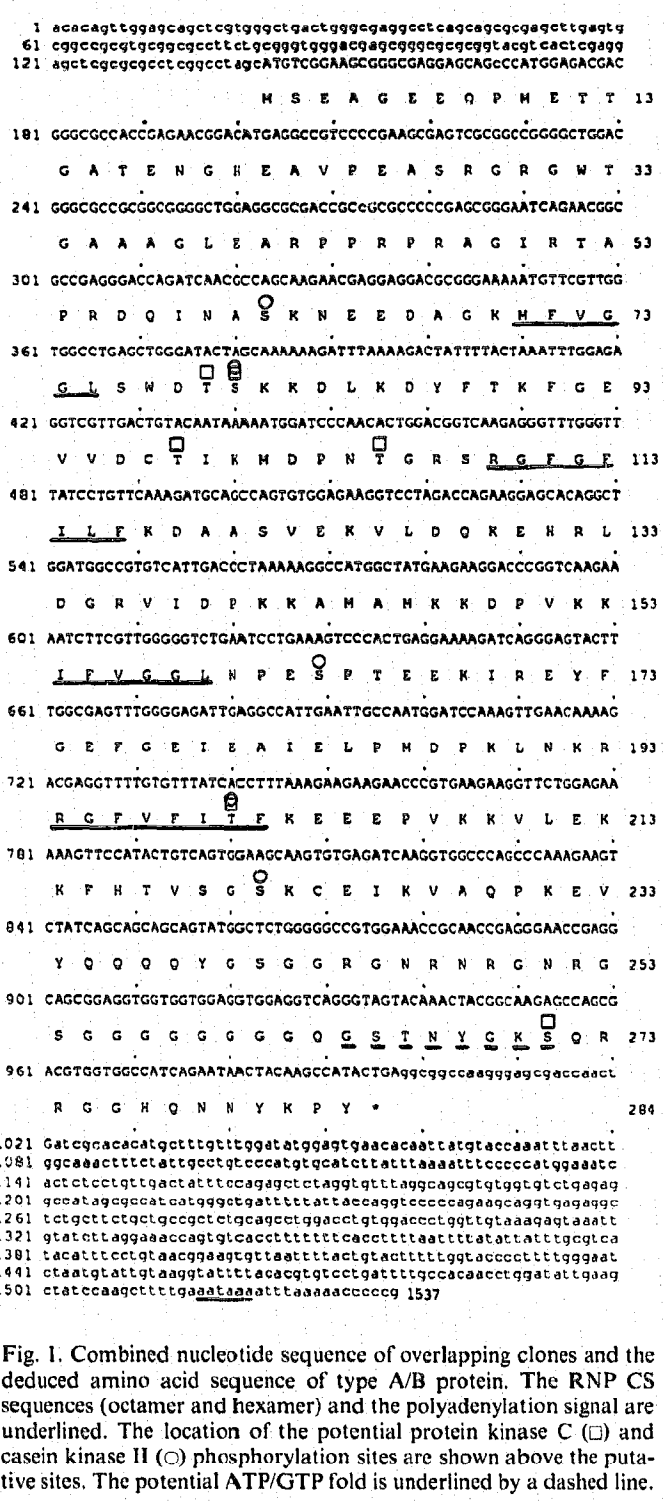

Fig. 1. Combined nucleotide sequence of overlapping clones and the deduced amino acid sequence of type A/B protein. The RNP CS sequences (octamer and hexamer) and the polyadenylation signal are underlined. The location of the potential protein kinase C (□) and casein kinase II (○) phosphorylation sites are shown above the putative sites. The potential ATP/GTP fold is underlined by a dashed line.

cross-hybridization to genomic DNA from mouse, chicken, Drosophila and yeast (not shown). Computer searches identified multiple potential serine and threonine phosphorylation sites for protein kinase C and casein kinase II within the protein (Fig. 1). Some of these are apparently phosphorylated in vivo since the purified protein separates into several isoelectric species within a pI range of 6.0–6.7 [10] although the amino acid sequence shown an excess of basic over acidic residues. Analysis of the derived amino acid sequence suggests the presence of several distinct domains (Fig. 2A and B) and an uneven distribution of charged amino acids. The N-terminal region of 25 amino acids is acidic; it contains 7 Glu residues and no other charged amino acids. The C-terminal part is basic, containing 8 basic and no acidic amino acids within a 40 amino acid segment. In contrast, the C1/C2 proteins have an acidic C-terminal and a basic N-terminal region [16]. There are 4 highly polar regions within type A/B protein, each containing nearly 50% of charged amino acids (residues 78–90, 117–153, 163–179 and 202–232). The presence of the polar segments and the asymmetric charge distribution may account for the retarded mobility of the protein in SDS-PAGE since the estimated M_r is 40–42 000 [10] while the cDNA encodes a protein with a M_r of 31 233. Retarded migration of a number of proteins with unevenly distributed charged residues has been observed, e.g. hnRNP proteins C1/C2 and initiation factor eIF-4B [4,7].

As seen from Fig. 2, type A/B protein contains 2 typical RNA binding domains (RNP BD), each about 80 amino acids long [1–7]. These domains are characterized by the presence of a conserved octapeptide motif, RNA-CS1, and a less conserved hexapeptide, RNA- CS2, located about 30 amino acids upstream from RNP-CS1. In addition, several amino acids are often conserved at specific positions throughout the 80 amino acids RNP BD (vertical arrows in Fig. 2A). There is 38% amino acid identity between the two domains and 55% similarity if conservative replacements are taken into account; note the positional conservation of aromatic and numerous charged and hydrophobic amino acids between the two domains (Fig. 2A). Two RNA binding domains are also present in hnRNP A and B proteins while C proteins contain a single RNA BD [7].

The C-terminal part of the protein contains 18 Gly residues, including a stretch of 8 consecutive glycines, within a 37 amino acid segment (res. 240–276). Similar C-terminal Gly-rich domains are present in hnRNP proteins A1 and A2/B1. The C-terminal part also contains a potential ATP/GTP binding fold (A site) identified so far only in hnRNP proteins C1/C2 [16].

The domain structure of type A/B protein and its isolation from 40S hnRNP suggest that the protein is a new member of the HeLa hnRNP protein family. The RNP BD sequence motifs, considered a hallmark of this class of RNA binding proteins [1–3], apparently recognize a wide variety of RNA structures and/or sequences while specific binding may depend on short amino acid sequences within or outside the 80 amino acid BD [3].

A GenBank search identified a recently submitted cDNA sequence encoding a 285 amino acid mouse pro-

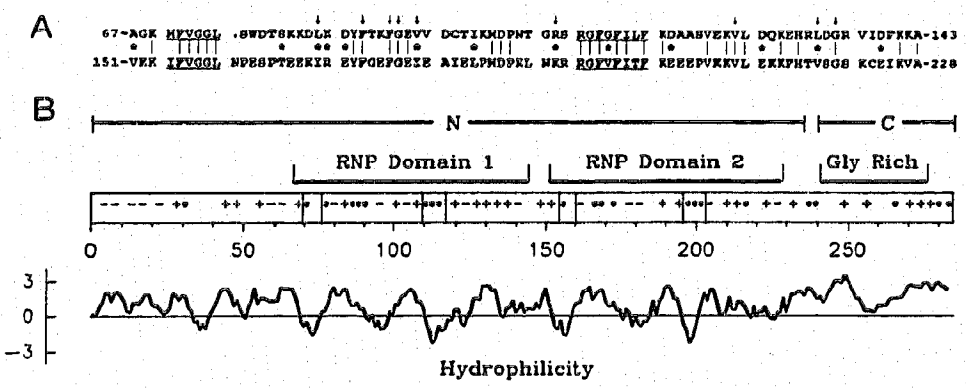

Fig. 2. RNA binding domains and the domain structure of type A/B protein. A. Amino acid sequence homologies between the two RNP BDs. Conservative replacements are shown by asterisks; vertical arrows indicate positions of lesser homologies within the domains found in many analogous proteins (cf. [1,4]). B. Domain structure and hydrophilicity. The distribution of charged (+,−) and aromatic (*) amino acids is shown and the RNP CS1 and 2 are boxed.

tein that is 88% identical to type A/B protein (Kamada, S. and Miwa, T., GenBank accession number D90151, 1991, unpublished). The mouse protein, termed 'CArG binding factor, single stranded DNA binding protein' differs from type A/B protein by a non-homologous region of 37 amino acids starting at residue 26 and contains a 3 Gly segment instead of the 8 Gly segment starting at residue 256 (Fig. 1). It will be of interest to determine how these differences affect the binding properties of the two proteins.

Acknowledgements: We gratefully thank Dr Barbara Merrill for peptide sequencing, Mr Yanggu Shi for assistance in computer analysis and Dr Halina Sierakowska for critical reading of the manuscript. This work was supported by National Institutes of Health Grant GM23705-18; computing was supported by National Science Foundation Grant DIR8908095.

REFERENCES

[1] Bandziulis, R.J., Swanson, M.S. and Dreyfuss, G. (1989) Genes Dev. 3, 431–437.
[2] Mattaj, I.W. (1989) Cell 57, 1–3.
[3] Zamore, P.D., Zapp, M.L. and Green, M.R. (1990) Nature 348, 485–486.
[4] Milburn, S.C., Hershey, J.W.B., Davies, M.V., Kelleher, K. and Kaufman, R.J. (1990) EMBO J. 9, 2783–2790.
[5] Li, Y. and Sugiura, M. (1990) EMBO J. 9, 3059–3066.
[6] Lee, W.-C., Xue, Z. and Melese, T. (1991) J. Cell Biol. 113, 1–12.
[7] Burd, G.C., Swanson, M., Görlach, M. and Dreyfuss, G. (1989) Proc. Natl. Acad. Sci. USA 86, 9788–9792.
[8] Choi, Y.D., Grabowski, P.J., Sharp, P.A. and Dreyfuss, G. (1986) Science 231, 1534–1539.
[9] Sierakowska, H., Szer, W., Furdon, P.J. and Kole, R. (1986) Nucleic Acids Res. 14, 5241–5254.
[10] Kumar, A., Sierakowska, H. and Szer, W. (1987) J. Biol. Chem. 262, 17126–17137.
[11] Celis, J.E., Bravo, R., Arenstorf, H.P. and LeStourgeon, W.M. (1986) FEBS Lett. 194, 101–109.
[12] Kumar, A., Williams, K.R. and Szer, W. (1986) J. Biol. Chem. 261, 11266–11273.
[13] Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) Molecular Cloning, 2nd edn., Cold Spring Harbor Laboratory, New York.
[14] Zhang, H., Scholl, R., Browse, J. and Somerville, C. (1988) Nucleic Acids Res. 16, 1220.
[15] Kozak, M. (1987) Nucleic Acids Res. 15, 8125–8148.
[16] Swanson, M.S., Nakagawa, T.Y., LeVan, K. and Dreyfuss, G. (1987) Mol. Cell. Biol. 7, 1731–1739.