Volume 243, number 2, 115–118 FEB 06715 January 1989

Hypothesis

Prediction of terminal protein and ribonuclease H domains in the gene P product of hepadnaviruses

Yu.E. Khudyakov and A.M. Makhov

D.I. Ivanovsky Institute of Virology, Academy of Medical Sciences of the USSR, Gamaleya st. 16, Moscow 123098, USSR

Received 17 November 1988

By means of comparative analysis of primary and secondary structures, and hydropathy plots of hepadnavirus P proteins new functional domains were revealed additionally to the polymerase domain which had been found earlier in these proteins. The C-terminal part of P proteins was revealed to be significantly similar to ribonuclease H of *E. coli*. The ribonuclease H functional domain is known to be an integral entity of retrovirus reverse transcriptase as a rule. Availability of this domain indicates once more the putative reverse transcriptase properties of the P products. The proteins of hepadnaviruses were compared to terminal proteins of picornaviruses, adenoviruses and bacteriophages. The data obtained suggested that a conservative N-terminal region of P proteins functions as protein primer for DNA synthesis in hepadnaviruses.

Reverse transcriptase; Ribonuclease H; Terminal protein; Domain structure; Protein homology

1. INTRODUCTION

In the hepadnavirus family the genome DNA replication mechanism includes a reverse transcription step as well as a terminal protein for priming of DNA synthesis [1]. Nevertheless neither activity has been directly associated with a particular viral protein. There is a hypothesis that the reverse transcriptase is coded by the P gene [2]. A region of the P protein of hepatitis B virus is homologous with the polymerase domain of retrovirus reverse transcriptase [2]. Since the pol proteins of retroviruses were shown to contain ribonuclease H, endonuclease and, in some cases, protease as well [1,3], we looked for additional functional domains in the P protein of hepadnaviruses.

Correspondence address: Yu.E. Khudyakov, D.I. Ivanovsky Institute of Virology, Academy of Medical Sciences of the USSR, Gamaleya St. 16, Moscow 123098, USSR

2. MATERIALS AND METHODS

The matrix homology program DNASEQ (Institute of Protein, Academy of Sciences of the USSR) was used for primary analysis of sequences. The statistical significance of sequence alignment was evaluated according to [4]. The correlation coefficient (K) of hydropathy plots [5] was calculated by the method described in [6]. Calculation of protein secondary structure was carried out according to the method in [7]. Sources of hepadnavirus protein and terminal protein sequences were the articles quoted in [8] and [9], respectively. Other sequences used were ribonuclease H of *E. coli* [10] and terminal protein of PRD1 [11].

3. RESULTS AND DISCUSSION

A comparison of retrovirus pol proteins with hepadnavirus P proteins failed to find additional sequence homology, excluding those that had been identified for the polymerase functional domain [2]. However, by direct comparison of hepadnavirus proteins with *E. coli* ribonuclease H, fairly long homologous regions were discovered. Gene P

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/89/$3.50

polypeptides of hepatitis B virus (HBV) and ground squirrel hepatitis virus (GSHV) appeared to be the most similar to ribonuclease H (fig.1). The percentage of coinciding amino acid residues in *E. coli* ribonuclease H and appropriate segments of P protein of HBV subtype adr or the GSHV protein was 24.1% or 25.4% ($P < 0.001$), respectively.

Analysis of hydropathy plots and secondary structure demonstrated essential similarity of these parameters for the compared proteins. The correlation coefficient (K) of hydropathy plots for the C-terminal region of HBV adr P protein and ribonuclease H was 0.2850 ($P < 0.01$). For hydropathy plots of HBV ayw P protein and ribonuclease H domain of Rous sarcoma virus polymerase [3] the K value was 0.2794 ($P < 0.002$). Thus our results suggest that the C-terminal region of gene P polypeptide of hepadnaviruses is the ribonuclease H functional domain. This finding indicates once more that the gene P product is an RNA-dependent DNA polymerase.

Gene P polypeptides appear to differ from pol proteins by the lack of an endonuclease domain which locates on the C-end of pol [3]. Absence of this sequence seems to reflect differences in genome replication of these two virus groups [1]. In retroviruses a polymerase domain locates as a rule on the N-end of pol [1,3]. But in hepadnaviruses the domain has been found in the C-terminal half of P protein [2]. Probably this region occupies position 400–700. A comparative analysis of the gene P products of hepadnaviruses demonstrated that the N-terminal region 1–200

```
                 ++++++ +++ ++ + + +  ++ ++ +  +    +++++  +     +
RNase H          LGNPGPG GYG AILRYRGREKTFSAGYTRTTNNRMELMAAIVA   LEA
                    *  * * **   * *  ** *     *    ** **  *     *
HBV adr          FADATPT GWGLAIGHRRMR GTFVAPLPIHT   AELLAACFARSRSGAK
HBV ayw          FADATPT GWGLVMGHQRMR GTFSAPLPIHT   AELLAACFARSRSGAN
WHV              FADATPT GWGIATTCQLLS GTFAFPLPIAT   AELIAACLARCWTGAR
GSHV             FADATPT GWGICTTCQLIS GTFGFSLPIAT   AELIAACLARCWTGAR
DHBV             ATDATPTHGAISHITGGSAV FAFSKVRDIHV   QELLMSCLAKIMIKPR

                 +      ++++          +     ++  ++  +     +    +    +
                 L KEHCEVILSTDSQYVRQGITQWIHNWKKRGWKTADKKPVKNVDLWQRLD
                 *      * **                ** **  *     *    *    *
                 LIGTDNSVVLS PKYTSFPWLLGCAANWILRG  T SFVYV PSALNPADD
                 IIGTDNSVVLS RKYTSFPWLLGCAANWILRG  T SFVYV PSALNPADD
                 LLGTDNSVVLS GKLTSFPWLLACVANWILRG  T SFCYV PSALNPADL
                 LLGTDNSVVLS GKLTSFPWLLACVANWILRG  T SFCYV PSADNPADL
                 CLLSDSTFVCH KRYQTLPWHFAMLAKQLLKP   IQLYFV PSKYNPADG

                 ++ + +   +  +   + ++   +    + ++++ + + +
                 AALGQHQIKWEWVKGHAGHPENERCDELARAAAMNPTLED  (14-148)
                    *              *   *    *        * *
                 PSRG RLGLYRPLLLLPFRPTTGRTSLYAVSPSVPSHLPD  (698-828)
                 PSRG RLGLSRPLLRLPFRPTTGRTSLYADSPSVPSHLPD  (687-817)
                 PSRG LLPVLRPLPRLRFRPPTSRISLWAASPPVSPRRPV  (732-862)
                 PSRG LLPALRPLPLLRFRPVTKRISLWAASPPVSTRRPV  (734-864)
                 PSR      HKPPDWTAF PYT PLSKAIYIPHRLCGT..  (714-836)
```

Fig.1. Comparison of sequences of *E. coli* ribonuclease H and hepadnavirus DNA polymerases. Asterisks indicate identical residues in the ribonuclease H and the P protein of subtype adr hepatitis B virus (HBV adr). Symbol + marks residues which belong to the same group of amino acids in these sequences. The groups were formed as follows: A, S, T, P and G; N, D, E and Q; H, R and K; M, L, I and V; F, Y and W. HBV ayw, sequence of DNA polymerase of HBV subtype ayw; WHV, woodchuck hepatitis virus; GSHV, ground squirrel hepatitis virus; DHBV, duck hepatitis B virus.

was conserved to the same degree as ribonuclease H and polymerase functional domains. Such obvious evolutionary preservation of the structure testifies apparently to the functional significance of this segment. In some retrovirus pol proteins the N-terminal part has been known to be protease [12]. However, we failed to find sequence homology between retrovirus proteases and P proteins, in agreement with published data [8]. What is the role of the N-terminal domain of the protein? We assume that this region functions as a protein-primer.

The GSHV terminal protein has not been found to be connected with the 5′-end of growing DNA chains through Ser or Thr residues [13]. Another known possible connection of terminal protein to nucleic acids is a phosphodiester linkage through Tyr. The only example of terminal protein of this type studied in detail is picornavirus VPg [9]. When the P proteins of hepadnavirus were compared to VPgs, we identified a conserved region in the N-terminal part of P proteins having marked similarity to the genome linked proteins of picornaviruses. This region contains the Gly-X-Tyr sequence which is the DNA connection site in VPgs and A protein of ϕX179 [9]. No such sequence was found in surface and core antigens, or X protein. The consensus was revealed only in the N-terminal region of P protein (fig.2). Region 61–84 of HBV ayw P protein and VPg of human rhinovirus 14 displayed the greatest similarity (fig.3). There are 32% of identical amino acids and 52% of chemically related residues. For comparison it should be noted that VPgs of picornaviruses belonging to the same genus, for example VPg of poliovirus 1 and hepatitis A virus, can be similar for 27.3% only. The most similar hydropathy profiles were those of picornavirus ECHO9 VPg and region 66–89 of the GSHV P protein. The K value was 0.4699 ($P < 0.01$).

In spite of overall absence of a sequence homology between adenovirus or phage ϕ29 terminal protein and the N-end part of hepadnavirus P protein, hydropathy plots of appropriate regions of these proteins coincided significantly. The K values for region 510–600 of adenovirus 5 terminal protein and region 180–260 of phage ϕ29 gp3 (both contain the functionally essential Ser residue), on the one hand, and region 1–90 of HBV ayw P protein and region 1–100 of GSHV protein, on the other hand, were 0.4053 ($P < 0.001$) and 0.3013 ($P < 0.002$), respectively.

The characteristic peculiarity of adenovirus and phage ϕ29 terminal protein secondary structures is

```
                *                    *   * * *   *       *    *
HBV ayw   FRRLLLLDDE      AGPLEEE L PRLA DEGLNRRVAEDLNLGNLNVS
WHV       FRNIQSLGEEEVQELLGPPEDA L PLLA GEDLNHRVADALNLHLPTAD
GSHV      FRNIQSLGEEEVQELLGPPEDA L PLLA GEGLNHRVADALNLQLPTAD
DHBV      LREAEKQLRVLENLVDSNLEEEKLKPQLSMGEDVQ#VRAATIDLPRLGNK
                  ---------------------------

             *       ***       **  *  *      *                *
          IPWTHKVGNFTGLYSSTVPVFNPHWKTPSFPNIHLHQDIIKKCEQFVGPL
          LQWVHKTNAITGLYSNQAAQFNPHWIQPEFPELHLHNELIKKLQQYFGPL
          LEWIHKTNVITGLYSTQTEKFNCNWKQPVFPKIHLDNNLFQKLENYFGPL
          LPARHHLGKLSGLYQMKGYTFNPEWKVPDISDTHFNLDVVNEC      P
                            ↑

            * *  *     ** * *     * *    * *  **      *       **
          TVNEKRRLQLIMPARFYPKVTKYLPLDKGIKPYYPEHLVNHYFQTRHYLH (8-150)
          TINEKRKLQLNFPARFFPKATKYFPLIKGIKNNYPNFALEHFFATANYLW (8-154)
          TTNEKRRLKLVFPARFFPNATKYFPLLKGIKDKYPNYTIEHFFAAANYLW (8-154)
          SRNWK YL    TPAKFWPKRISYFPVQVGVKPKYPDNVMQHESIVGKYLT (65-222)
```

Fig.2. Comparison of N-terminal parts of the polypeptide P sequences representing putative terminal protein domain. Identical positions in these sequences were marked by asterisks. The arrow indicates Tyr in the consensus Gly-X-Tyr. Conserved α-helix region is underlined. The region of DHBV polymerase from position 100 to 118 is substituted by #. For description of other conventional symbols see the legend to fig.1.

```
          * **   *  *   * *   *
HRV-14    GPYSG NPPHN KLKAPTLRPVVVQ
HBV ayw   GLYSSTVPVFNPHWKTPS FPNIHL
          + +++  +  + + ++++  + +
```

Fig.3. Comparison of sequences of human rhinovirus 14 VPg (HRV-14) and protein P region 61–84 of hepatitis B virus subtype ayw (HBV ayw). Identical positions were marked by asterisks. Symbol + indicates amino acids which belong to the same groups in both sequences (see legend to fig.1).

a negatively charged α-helix located immediately upstream of the Ser residue which is linked to the 5′-end of genomic DNA [9]. An analogous structure is also revealed at a short distance from the Gly-X-Tyr sequence of P proteins (fig.2). In hepadnaviruses the region of P gene coding for this α-helix overlaps with C gene [1]. It has been proposed that the hepadnavirus polymerase is synthesized as a nucleocapsid-polymerase fusion protein formed by means of a change of translation phase in the gene overlapping region [14]. Our findings provide evidence on the functional importance of the region of P protein coded by the overlapping part of these genes. Therefore, it may be suggested that the change of translation phase occurs somewhere near the first methionine codon of the P gene. In favour of this assumption, antibodies have been detected in the serum of hepatitis B patients which interacted with a synthetic peptide corresponding to region 29–38 of the P protein [15].

In conclusion, if the hypothesis on the mechanism of hepadnavirus DNA polymerase synthesis [14] is correct, then the terminal protein might contain at least sequences of the nucleocapsid protein and the N-terminal part of P polypeptide. Our understanding of the functional organization of the hepadnavirus reverse transcriptase 'precursor' is summarized in fig.4. Region 200–400 is highly hydrophilic and the most variable part of P protein. We assume that this region functions as tether between terminal protein and reverse transcriptase domains. When this article was being submitted experimental evidence for the model of the retrovirus pol proteins' functional organization [3] was published [16]. Indirectly these results confirm our prediction of relative localization of polymerase and ribonuclease H domains in the hepadnavirus P proteins.

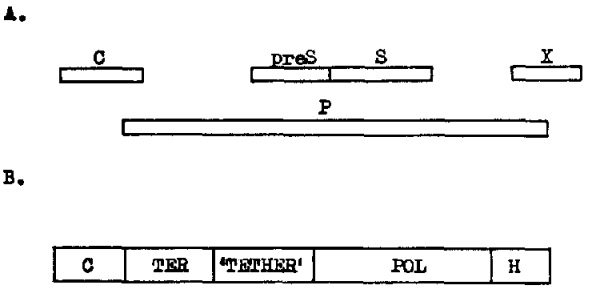

Fig.4. (A) Relative arrangement of four protein coding zones in hepatitis B virus genome. C, gene of nucleocapsid protein; preS and S, gene of HBsAg with preS-region; X, gene of X protein; P, gene P. (B) Putative domain structure of probable precursor polyprotein in hepadnaviruses. C, address functional domain represented by part of nucleocapsid protein [14]; TER, terminal protein domain; 'TETHER', highly hydrophilic and the most variable part of P protein; POL and H, polymerase and ribonuclease H domains, respectively.

REFERENCES

[1] Masson, W.S., Taylor, J.M. and Hull, R. (1987) in: Retroid Virus Genome Replication (Maramorosch, K. et al. eds) Adv. Virus Res. vol.32, pp.35–96, Academic Press, New York.
[2] Toh, H., Hayashida, H. and Miyata, T. (1983) Nature 305, 827–829.
[3] Johnson, M.S., McClure, M.A., Feng, D.-A., Gray, J. and Doolittle, R.F. (1986) Proc. Natl. Acad. Sci. USA 83, 7648–7652.
[4] Feng, D.F., Johnson, M.S. and Doolittle, R.F. (1985) J. Mol. Evol. 21, 112–125.
[5] Kyte, J. and Doolittle, R.F. (1982) J. Mol. Biol. 157, 105–132.
[6] Sweet, R.M. and Eisenberg, D. (1983) J. Mol. Biol. 171, 479–488.
[7] Chow, P. and Fasman, G. (1978) Adv. Enzymol. 47, 45–148.
[8] Miller, R.H. (1987) Science 236, 722–725.
[9] Vartapetian, A.B. and Bogdanov, A.A. (1987) Prog. Nucleic Acids Res. Mol. Biol. 34, 209–251.
[10] Kanaya, S. and Crouch, R.J. (1983) J. Biol. Chem. 258, 1276–1281.
[11] Hsieh, J.-C., Jung, G., Leavitt, M.C. and Ito, J. (1987) Nucleic Acids Res. 15, 8999–9009.
[12] Wellink, J. and Van Kammen, A. (1988) Arch. Virol. 98, 1–26.
[13] Molnar-Kimber, K.L., Summers, J., Taylor, J.M. and Mason, W.S. (1983) J. Virol. 45, 165–172.
[14] Will, H., Salfeld, J., Pfaff, E., Manso, C., Theilmann, L. and Schaller, H. (1986) Science 231, 594–596.
[15] Feitelson, M.A., Millman, I., Duncan, G.D. and Blumberg, B.S. (1988) J. Med. Virol. 24, 121–136.
[16] Tanese, N. and Goff, S.P. (1988) Proc. Natl. Acad. Sci. USA 85, 1777–1781.