Hypothesis

Shared motifs of the capsid proteins of hepadnaviruses and retroviruses suggest a common evolutionary origin

Adam Zlotnick[1,a,*], Steven J. Stahl[a], Paul T. Wingfield[a], James F. Conway[b], Naiqian Cheng[b], Alasdair C. Steven[b]

[a] *Protein Expression Laboratory, NIAMS-NIH, 6 Center Dr MSC 2717, Bldg 6, Rm 138, Bethesda, MD 20892-2717, USA*
[b] *Laboratory of Structural Biology Research, NIAMS-NIH, Bethesda, MD 20892-2717, USA*

Received 13 March 1998; revised version received 2 June 1998

Abstract The structure of the dimeric C-terminal domain of the HIV-1 capsid protein (CA), recently determined by X-ray crystallography (Gamble et al. (1997)), has a notable resemblance to the structure of the hepatitis B virus (HBV) capsid protein (Cp) dimer, previously determined by cryo-electron microscopy (Conway et al. (1997), Böttcher et al. (1997)). In both proteins, dimerization is effected by formation of a four-helix bundle, whereby each subunit contributes a helix-loop-helix and most of the interaction between subunits is mediated by one pair of helices. These are the first two observations of a motif that is common to the capsid proteins of two enveloped viruses and quite distinct from the eight-stranded anti-parallel β-barrel found in most other virus capsid proteins solved to date (Harrison et al. (1996)). Motivated by the structural resemblance, we have examined retroviral and HBV capsid protein sequences and found weak but significant similarities between them. These similarities further support an evolutionary relationship between these two virus families of great medical importance – the hepadnaviruses (e.g. HBV) and retroviruses (e.g. HIV).

Key words: HIV; Retrovirus; Hepatitis B virus; Hepadnavirus; Gag; Capsid; Capsid assembly

1. Hypothesis

Until recently, structural information about the capsid proteins of icosahedral viruses [5] has been dominated by crystal structures of capsids of nonenveloped viruses. The common structure among capsid proteins is the eight-stranded anti-parallel β-barrel [4], which has been observed in RNA and DNA viruses infecting vertebrates, invertebrates, plants, and bacteria. Though the structural similarity between these capsid proteins is strong and obvious, no widespread sequence similarity between capsid proteins has been reported (an evolutionary relationship between RNA virus families has been detected in virally encoded enzymes [6]). More recently, fragments of retroviral capsid proteins, determined by crystallography and NMR (for example [1,7–9]), and the HBV capsid protein, determined by cryo-electron microscopy image reconstruction [2,3], have yielded a different structural theme – helical bundles. A four-helix bundle dimerization motif has been observed in the capsid proteins of HIV-1 [1] and HBV [2,3] (Fig. 1). Although four-helix bundles are a common structural motif, dimerization by formation of a four-helix bundle is rare in general and was previously unknown in virus capsid proteins; the only other examples of this dimerization motif, with the same topology as that of HIV-1 CA and HBV Cp, are transcription activating factors (TAFs) and histones [10]. We suggest: (i) that structural and sequence similarity of hepadnavirus and retrovirus capsid proteins is further evidence of an evolutionary relationship between these two virus families and (ii) that this structural motif may be a common feature in other families of enveloped viruses.

There are striking resemblances as well as marked contrasts between retroviruses and hepadnaviruses. Both are enveloped and both employ reverse transcriptases, albeit in converse fashion: HBV is a DNA virus with an RNA intermediate and HIV is an RNA virus with a DNA intermediate. Their

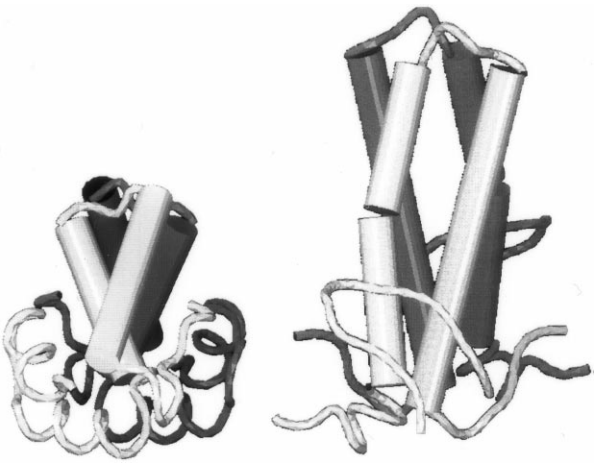

Fig. 1. Comparison of the dimerization motifs of HIV-1 (A) [1] and HBV (B) [2]. Only the helices of the 'four-helix bundles' bundles are represented as tubes. For both dimers, the monomers are shaded distinctly for ease of visualization. The HIV-1 CA monomer coordinates were accessed from the Brookhaven protein database (1AU3); the dimer was generated by application of a crystallographic two-fold. For the HBV Cp model, only the dimerization motif could be assigned with confidence using the EM density, the remainder of the model is 'domainal' not atomic [2]. The models in this figure were generated using the program O [23].

*Corresponding author.

[1]Current address: OUHSC, Biochemistry and Mol. Biol., PO Box 26901, Oklahoma City, OK 73190, USA. Fax: +1 (301) 480-7629. E-mail: adam_zlotnick@nih.gov

Abbreviations: HBV, hepatitis B virus; HIV, human immunodeficiency virus; MMLV, Moloney murine leukemia virus; CA, retrovirus capsid protein; Cp, hepadnavirus capsid protein; MHR, retrovirus capsid protein major homology region

PII: S0014-5793(98)00755-8

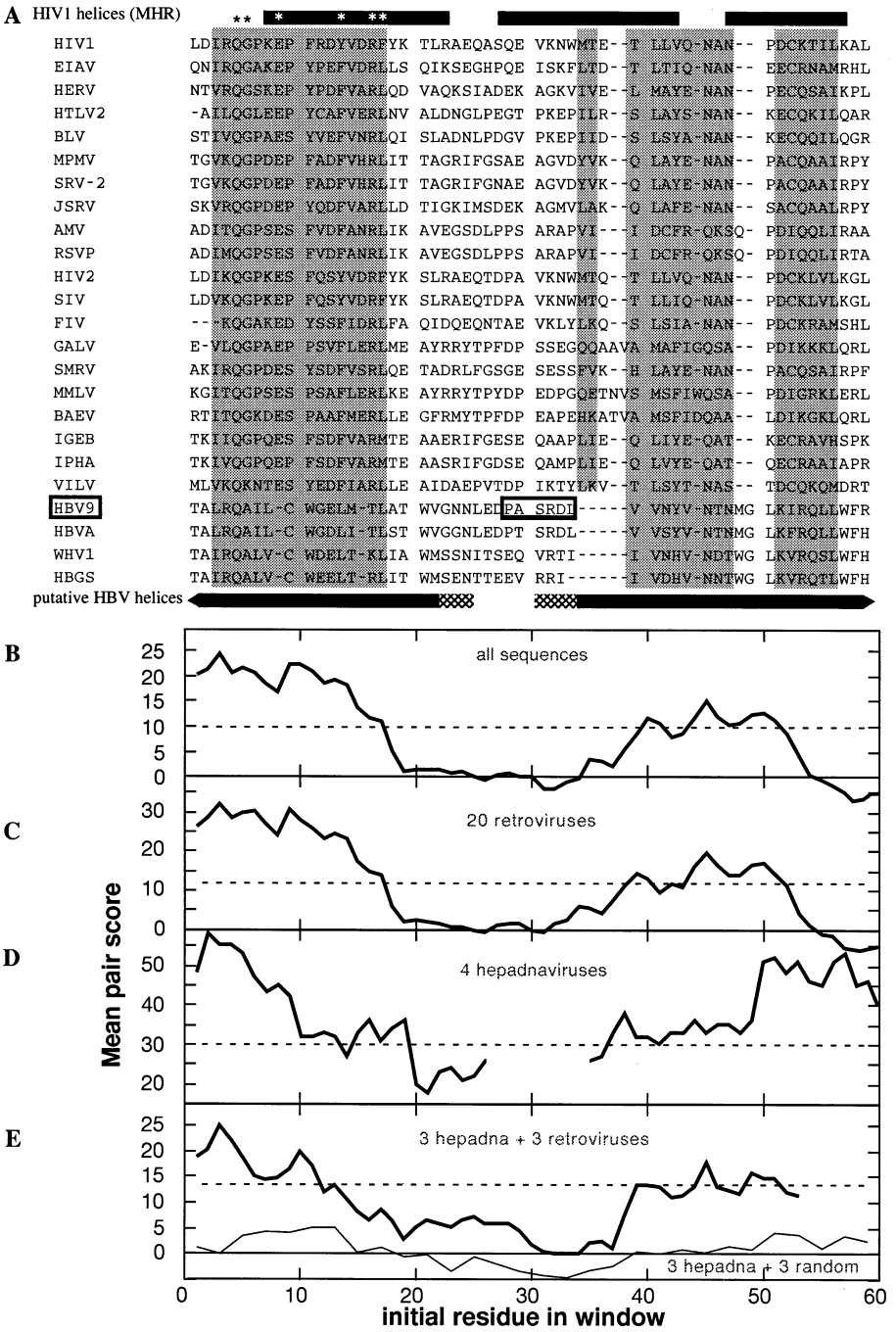

Fig. 2. Alignment of capsid protein sequences of 20 retroviruses and four hepadnaviruses. A: The solid bars represent the positions of α-helices from the crystal structure HIV-1 CA [1]; the HIV-1 four-helix bundle is formed by dimerization of helix 1-loop-helix 2. The positions of the conserved retroviral MHR residues are marked with asterisks. The most conserved regions of the alignment are shaded. The antigenic loop of HBV Cp, found at or near the top of the HBV helix-loop-helix is boxed. Predicted HBV helices are shown beneath the alignment [13,22], the positions of the end of helix 1 and the beginning of helix 2 are ambiguous (hashed area). The panel of sequences was compiled using PSI-BLAST [16] (http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/nph-psi_blast). The alignment was accomplished with CLUSTAL W [17] and MACAW [18,19] (http://www.ncbi.nlm.nih.gov/Baxevani/CSH/multaln.html). Previously noted similarity between HBV Cp and MMLV CA [15] is C-terminal of these sequences. B–D: Evaluation of the similarity of sequences over a nine-residue window, using MACAW with the BLOSUM45 amino acid comparison matrix [18]. The dashed line in each panel indicates statistical significance ($P < 0.01$) which is a function of window size and the number of different sequences to be compared. This evaluation is shown for (B) all 24 virus capsid protein sequences, (C) the 20 retroviral sequences, (D) the four hepadnaviral sequences (the last four rows in the list), (E, thick line) three hepadnavirus and three retrovirus sequences (two human HBV clones, HBV9 and A, woodchuck hepatitis virus, WHV1, human immunodeficiency virus, HIV-1, equine infectious anemia virus, EIAV, and human endogenous retrovirus K, HERV), and (E, thin line) a control for amino acid composition, the three previously noted hepdnavirus sequences and the three retroviruses with their amino acid sequences randomized. Swiss-Prot accession numbers are: HIV-1, P05887; EIAV, P03351; HERV, P10264; HTLV2, P03346; BLV, P25058; MPMV, P07567; SRV-2, P51516; JSRV, P31622; AMV, P03323; RSVP, P03322; HIV-2, P24106; SIV, P12496; FIV, P31821; GALV, P21416; SMRV, P21411; MMLV, P29168; BAEV, P03341; IGEB, P03975; IPHA, P04023; VILV, P03352; HBV9, P17099; HBVA, P24023; WHV1, P03152; HBGS, P03153.

respective reverse transcriptases share a limited but significant sequence similarity [11]. However, HBV is among the smallest human viruses whereas retroviruses are relatively large and complex. The HBV Cp is only 183 residues long and comprises an N-terminal assembly domain (149 residues) and a C-terminal protamine-like domain. The assembly domain is sufficient for assembly of both T = 3 and T = 4 icosahedral capsids. Retroviral capsid structure and assembly is much more complicated [12]. Immature retroviral capsids are assembled from >2000 copies of the ~ 500 amino acid *gag* polyprotein. After assembly this precursor is proteolyzed to release CA (amongst other proteins) which forms the surface of a truncated cone within the virus envelope. CA has two domains: an N-terminal five-helix bundle [7–9] and a C-terminal dimerization domain [1].

We observe that the HBV assembly domain and the C-terminal dimerization domain of HIV-1 CA (residues 146–231 of the *gag* polyprotein) have essentially the same fold (Fig. 1). in addition to their common fold, there is some sequence similarity between the two capsid proteins although it is subtle (see below). On the one hand, Cp is highly conserved throughout the hepadnavirus family [13]. On the other hand, there is little identity among retroviral CA proteins ($<25\%$), except for a short major homology region (MHR) with six highly conserved amino acids [14] (Fig. 2). However, a significant sequence similarity has been noted between Moloney murine leukemia virus (MMLV) CA protein and the C-terminal half of HBV Cp, including the protamine domain [15]. The MHR has been used to align CA proteins from HIV-1 and MMLV [1]. These three sequences, and a consensus based on them, were used to search for related sequences in the SWISS-PROT database with the program PSI-BLAST [16]. A panel of retrovirus and hepadnavirus sequences was compiled and aligned with the program CLUSTAL W [17]. The retroviral sequences include representatives of the retroviral genuses (B-type, C-type, D-type, avian-leukosis-sarcoma, HTLV-BLV, and lentivirus) except the spumaviruses. The alignment was refined and evaluated using an updated version of MACAW [18,19]. To minimize the effects of the redundancy of the hepadnavirus sequences [13], a small sample of these viruses was used in these alignments.

The final alignment of 20 retroviruses and four hepadnaviruses (Fig. 2) reveals two short regions of significant similarity. The similarity can be best interpreted when the alignment is correlated with the observed secondary structure of HIV-1 CA. The region of greatest similarity is in helix 1 around the MHR; the end of helix 2 extending to helix 3 also show significant similarity. The loop following helix 1 and the N-terminus of helix 2 are poorly conserved. These generalizations hold within the retrovirus family (Fig. 2C) and between retroviruses and hepadnaviruses (Fig. 2B, E).

Our alignment is consistent with experimental observations. First, the secondary structure assignment places a Cp antigenic peptide (residues 78–83) in or near the loop at the top of the four-helix bundle (Fig. 2). Böttcher et al. [3] suggested this location for the antigenic sequence. Long inserts in the corresponding region of the homologous avian hepatitis capsid protein were also localized to the spike [20]. This localization for the antigenic sequence has been confirmed for HBV by Fab labeling studies recently completed in this laboratory [21]. Second, the four-helix bundle of HBV is substantially longer than that of HIV-1 (Fig. 1). A number of helix-breaking prolines and glycines that are conserved amongst retroviruses, especially around the MHR in helix 1, are absent from the hepadnaviruses. This suggests that the HBV helix 1 is extended, compared to that of HIV-1, and that helix 2 and helix 3 are fused together. The putative helix 2-helix 3 fusion incorporates a coiled coil-forming heptad repeat [22]. Furthermore, the extension of the HBV four-helix bundle is accomplished without amino acid insertions.

We conjecture that the four-helix bundle structure is likely to be conserved in retroviruses, despite the sequence divergence. It is apparent that there is no highly conserved sequence identity specific to this structure (see Fig. 2) which is common to both the hepadna- and retroviruses. In this respect, the four-helix bundles are like the β-barrel capsid proteins, among which the lack of sequence identity belies the structural similarity [4].

While the Cp dimer is unambiguously the building-block of the HBV capsid, the status of the CA dimer, described by Gamble et al. [1], as the building-block of the HIV-1 capsid is less clear. Also, although their ultracentrifugation evidence identifies a dimer as a stable state of assembly, it remains to be determined whether this dimer is in fact the parallel dimer formed by the 4-helix bundle. There are examples of molecules that have appeared as plausible dimers in crystal structures turning out to be dimers of different kinds in solution (e.g. [24]). Notwithstanding this caveat, the sequence similarity that we have detected as well as the ability of the CA protein to dimerize by this interaction [1] supports the hypothesis of a common evolutionary origin for these two virus families.

References

[1] Gamble, T.R., Yoo, S., Vajdoos, F.F., von Schwedler, U.K., Worthylake, D.K., Wang, H., McCutcheon, J.P., Sundquist, W.I. and Hill, C.P. (1997) Science 278, 849–853.
[2] Conway, J.F., Cheng, N., Zlotnick, A., Wingfield, P.T., Stahl, S.J. and Steven, A.C. (1997) Nature 386, 91–94.
[3] Böttcher, B., Wynne, S.A. and Crowther, R.A. (1997) Nature 386, 88–91.
[4] Harrison, S., Wiley, D.C. and Skehel, J.J. (1996) in: B.N. Fields, D.M. Knipe, P.M. Howley, R.M. Chanock, J.L. Melnick, T.P. Monath, B. Roizman and S.E. Straus (Eds.), Fields Virology, 3rd Edn., Lippincott - Raven Publishers, Philadelphia, PA, pp. 59–100.
[5] Caspar, D.L.D. and Klug, A. (1962) Cold Spring Harbor Symp. Quant. Biol. 27, 1–24.
[6] Dolja, V.V. and Koonin, E.V. (1991) J. Gen. Virol. 72, 1481–1486.
[7] Gitti, R.K., Lee, B.M., Walker, J., Summers, M.F., Yoo, S. and Sundquist, W.I. (1996) Science 273, 231–235.
[8] Momany, C., Kovari, L.C., Prongay, A.J., Keller, W., Gitti, R.K., Lee, B.M., Gorbalenya, A.E., Tong, L., McClure, J., Ehrlich, L.S., Summers, M.F., Carter, C. and Rossmann, M.G. (1996) Nat. Struct. Biol. 3, 763–770.
[9] Gamble, T.R., Vajdos, F.F., Yoo, S., Worthylake, D.K., Houseweart, M., Sundquist, W.I. and Hill, C.P. (1996) Cell 87, 1285–1294.
[10] Xie, X., Kokubo, T., Cohen, S.L., Mirza, U.A., Hoffmann, A., Chait, B.T., Roeder, R.G., Nakatani, Y. and Burley, S.K. (1996) Nature 380, 316–322.
[11] Toh, H., Hayashida, H. and Miyata, T. (1983) Nature 305, 827–829.
[12] Coffin, J.M. (1996) in: B.N. Fields, D.M. Knipe, P.M. Howley, R.M. Chanock, J.L. Melnick, T.P. Monath, B. Roizman and S.E. Straus (Eds.), Fields Virology, 3rd Edn., Lippincott - Raven Publishers, Philadelphia, PA, pp. 1767–1847.
[13] Bringas, R. (1997) J. Struct. Biol. 118, 189–196.

[14] Patarca, R. and Haseltine, W.A. (1985) Nature 318, 390.
[15] Miller, H. and Robinson, W.S. (1986) Proc. Natl. Acad. Sci. USA 83, 2531–2535.
[16] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Nucleic Acids Res. 25, 3389–3402.
[17] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Nucleic Acids Res. 22, 4673–4680.
[18] Schuler, G.D., Altschul, S.F. and Lipman, D.J. (1991) Proteins 9, 180–190.
[19] Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F. and Wootton, J.C. (1993) Science 262, 208–214.
[20] Kenney, J.M., von Bonsdorff, C.H., Nassal, M. and Fuller, S.D. (1995) Structure 3, 1009–1019.
[21] Conway, J.F., Cheng, N., Zlotnick, A., Stahl, S.J., Wingfield, P.T., Belnap, D.M., Kanngiesser, U., Noah, M. and Steven, A.C. (1998) J. Mol. Biol., in press.
[22] Yu, M., Miller, R.H., Emerson, S. and Purcell, R.H. (1996) J. Virol. 70, 7085–7091.
[23] Jones, T.A., Zou, J.Y., Cowan, S.W. and Kjeldgaard, M. (1991) Acta Crystallogr. A 47, 110–119.
[24] Ferentz, A.E., Opperman, T., Walker, G.C. and Wagner, G.T. (1997) Nat. Struct. Biol. 4, 979–983.