Arch Virol (1999) 144: 1139–1158

Archives of Virology

RNA binding domain of HDV antigen is homologous to the HMG box of SRY

S. Veretnik and **M. Gribskov**

San Diego Supercomputer Center, San Diego, California, U.S.A.

Accepted December 29, 1998

Summary. The delta antigen of hepatitis delta virus exhibits sequence specific binding to its own RNA and is essential for viral replication. Using statistical methods we have detected significant similarity between the RNA-binding domain of the hepatitis delta antigen and the HMG box of SRY. Our analysis suggests that the RNA-binding domain of HDV antigen evolved from the DNA-binding domain of the HMG box. SRY, or a related protein, is a probable cellular cognate of HDV.

Introduction

Hepatitis delta virus (HDV), the satellite of hepatitis B virus (HBV) is unique among animal viruses in that it contains a circular, viroid-like RNA genome. While it is fully dependent on the helper virus for its packaging, HDV has no significant sequence similarity to HBV; its replication is independent of HBV and its geographic prevalence is different [30]. HDV is a single stranded RNA virus. It replicates through a double rolling-circle mechanism, probably using RNA polymerase II and cellular transcriptional machinery [8, 22]. Two detectable HDV antigens (HDAg) are encoded by a single open reading frame on the antigenomic strand. The large antigen (L-HDAg) is a result of RNA editing which extends the protein coding sequence of the small antigen (S-HDAg) by 19 residues; L-HDAg appears later in infection, suppresses replication and initiates viral packaging [7, 25]. The small antigen (S-HDAg) is essential for viral replication and is proposed to facilitate the interactions between the cellular transcriptional machinery and the viral template. The central domain of S-HDAg, which is involved in binding to the viral genome, contains two arginine-rich motifs (ARMs) bracketing a leucine zipper. Mutational analysis indicates that both the ARMs and the leucine zipper region are essential for binding to the HDV genome [32]. Using sequence analysis methods we have discovered significant sequence similarity between this central domain of HDAg and the HMG (High Mobility Group) box of the SRY gene [40], whose product binds to DNA and is essential for sex determination. Our results indicate a possible evolution of this widely used DNA-binding domain into the RNA-binding domain of HDAg.

The organization of the HDV genome suggests two distinct sources for its genomic RNA: 25% of the genome resembles a plant viroid RNA (the smallest known self-replicating genome), while the remaining three-quarters of the genome is more like a cellular RNA. Half of this cellular domain codes for HDAg. This unique combination of domains within HDV gave rise to the hypothesis that a small viroid (300–400 bp long) captured a cellular RNA, resulting in a larger virus that retained the ribozyme activities of the viroid (such as self-cleavage and self-ligation) [3, 39]. The sequence similarity between the central domain of HDAg and the HMG (High Mobility Group) box of the SRY gene suggests that SRY may be a cellular cognate of HDAg.

Another candidate for the ‘captured’ cellular RNA was recently proposed: a 202 residue cellular protein referred to as DIPA (delta interacting protein A) was identified by co-immunoprecipitation with HDAg and by its in vivo ability to inhibit viral replication [5]. Our analysis shows that the level of similarity between HDAg and DIPA is within noise level; the similarity is likely to be due to compositional biases found in motifs consisting of arginine-rich regions and leucine zippers (bZIP motifs). This paper reports the results of the in-depth sequence analysis applied to the sequence of HDV antigen in an attempt to further understand the function and evolution of hepatitis delta virus – a unique entity in the world of animal RNA viruses.

Materials and methods

Sequence alignments

Sequence alignments were performed using programs from the GCG Wisconsin Package version 8.1-UNIX (1995). For multiple sequence alignments, the PILEUP program (based on a simplified version of the progressive alignment method of Feng and Doolittle) [20] under default conditions (GapOpen=3.0, GapExtend=0.1) was applied. The GAP program, based on the global alignment algorithm of Needleman and Wunsch [35] was used for alignment of sequence pairs. The best matched subsequences of SRY, HDAg and DIPA were detected by visual inspection; alignments of these regions are referred to as partial alignments throughout this paper. The partial alignments were constructed by aligning selected subsequences using the GAP program. The comparison matrix PAM150 [12] was used for scoring amino acid residue similarities. Gap opening and gap extension penalties were varied and are specified for each alignment. Monte Carlo evaluation of sequence alignments was done using the GAP program with the option ‘-RANDOM=10000’. This evaluation process consists of randomizing the first sequence and realigning it to the second sequence 10,000 times, each time calculating the alignment score. The average and standard deviation of 10,000 alignments of randomized sequences were then used to calculate the Z score[1] for the original alignment.

[1] $Z = \dfrac{(\text{Original Alignment Score} - \text{Average Score})}{\text{Standard Deviation}}$,

Databases and database searches

The SWISS-PROT database release 33 (2/96) was used for database searches. Evolutionary profiles [27] were constructed from groups of sequences aligned with the PILEUP program. A profile is a two-dimensional weight matrix that describes the alignment of a group of sequences. Each row of a profile describes a single position in the alignment; its 20 values estimate the likelihood that each of the 20 possible amino acid residues occurs at this position in the alignment. Two additional values in each row represent the position-specific likelihood of gap initiation and extension. Evolutionary profiles improve the description of multiple sequence alignments by modeling varying rates of sequence change at each position in the sequences: each position is modeled using the most appropriate scoring matrix to reflect its rate of change. Sequence weighting, based on a modification of Felsenstein's method [19], was performed prior to profile construction, so that the contribution of each sequence to the evolutionary profile is proportional to the amount of unique information it contributes to the multiple alignment. Structure-based profiles (to be described elsewhere) are constructed using scoring matrices based on secondary structure information. Each position in the alignment is described as a mixture of four distributions – one for each of the secondary structures (helix, sheet, turn and coil). Each distribution is weighted by the likelihood of the observed residue distribution at a given position being found in a specific secondary structure. New sequences identified during the database search with the profile of the HDAg family were then re-aligned to this profile with the PROFILEGAP program using default gap opening and gap extension penalties (GapOpen=3.0, GapExtend=0.1). The Profilegap program is based on an extension of the Smith-Waterman local similarity algorithm [26, 42].

Testing for frameshifting

The GCG Package FRAMESEARCH program was used to search for sequencing errors and potential frameshifts during translation. Framesearch is based on a local alignment algorithm that allows introduction of the gaps and shifting of the reading frame. Framesearch was used under the following conditions: GapOpen=3.0, GapExtend=0.5, scoring matrix: PAM150.

Selection of comparison matrix

PAM matrices at distances 100, 150, 200 and 250 were tested [12]. We found that the PAM150 scoring matrix consistently produced alignments with more significant Z scores. Thus PAM150 was used as a scoring matrix throughout this work.

Statistical evaluation of the alignments: P-values

P-values, the probability of achieving an equal or better score, for the best partial alignments of SRY-HDAg and DIPA-HDAg (Fig. 2A, C) were calculated by the Monte Carlo procedure described above (see *Sequence alignments*). Randomly generated sequences with the same composition as the best locally matching segments were compared to the respective query sequence 1,000,000 times. This is a very conservative test since the compositions of these segments are biased. Theory suggests that sequence alignment scores follow an extreme value distribution (EVD) similar to that of maximal segment scores [29], i.e.

$$P(score > x) = KNe^{-\lambda x + \text{constant}}$$

for large values of x (K and λ are constants). This relationship has been empirically verified by Collins et al. [11] who first described this approach to calculating P-values for alignments.

The logarithm of the frequencies of occurrence of scores greater than or equal to specific values were plotted for the top 500 (0.05%) random alignments and the P-values calculated from the resulting plot (Fig. 4).

Results

Database searches

The similarity between the SRY and HDAg was noticed during SWISS-PROT database searches with evolutionary profiles [27] constructed from a multiple alignment of the RNA-binding domain of HDAg from nine distinct isolates of HDV. Searches with the evolutionary profile identified several target sequences that contained HMG boxes, such as HMGI and HMGY with Z score >7.0. Searches with profiles based on the predicted secondary structures identified other HMG products, such as HMG2_HUMAN and HMG2_MOUSE (Z score >7.0), and SRY_HORSE (Z score>6.0). Pairwise alignments of these target sequences with the RNA-binding domain of HDAg identified the HMG box of SRY as most closely related sequence. We then constructed profiles from the SRY-HDAg alignment, which, upon database search, identified a large family of proteins containing the HMG box: SRY genes, SRY-related genes and SOX (SRY box) genes.

Sequence alignments

HDAg shows the highest similarity to the HMG box of the human SRY and SOX3 gene products; for the sake of clarity only the SRY sequence is shown in the analysis below. The best alignment results are achieved with the first 47 residues of the HMG box (corresponding to residues 59–105 of SRY) aligned to the 42 residues of RNA-binding motif of HDAg (corresponding to the residues 86–127): 54% identity and 72% sequence similarity are observed (Fig. 2A). This region covers the first ARM and part of the leucine zipper region. This is the only alignment that produces a significant Z score (see below). A longer alignment (Fig. 2B) covers two ARMs bracketing a leucine zipper, which together comprise the RNA-binding motif in HDAg. This longer alignment between HDAg and SRY proteins extends though the C-terminal region of both proteins; it exhibits 33% identity and 59% similarity (Fig. 2B). The similarity in HDAg begins ten residues upstream of the left ARM and continues into the leucine zipper, where similarity gradually weakens but is still detectable through the right ARM of HDAg (Fig. 1, 2B). When HDAg was aligned with ARM-containing sequences, the best alignment was found with HIV *rev* gene product. This alignment shows 39% identity and 63% similarity across the entire *rev* gene (Fig. 2F).

The alignment of HDAg and DIPA, a proposed cellular homolog of HDAg [5], shows 34% identity and 57% similarity across the entire length of the protein sequences (Fig. 2D). The similarity is distributed across the entire sequence, however, the amino-terminal and central regions show a somewhat higher level of similarity (Fig. 2D). Our analysis is focused on the central region of HDAg which shows higher level of sequence similarity and also corresponds to the most similar region of HDAg and SRY. This 55-residue long subsequence corresponds

A.

```
         1                                                                                                             110
SRY      MQSYASAMLSVFNSDDYSPAVQENIPALRRSSSFLCTESCNSKYQCETGENSKGNVQDRVKRPM.NAFIVWSR.DQRRKMALENPRMRNSEISKQLGYQ.KMLTEAEKWPFFQ
         : :   ::|:  : |       |  |: :|: :         : : ::|         | |||:  :|    | |:||::|||| :       ||||   | |:  |     :
HDAg     L.GNIKGILGKKDKDG......EGAPPAKRART........DRMEVDSG........PR.KRPLRGGFTDKERQDHRRRKALENKK.......KQLGAGGKNLSREEE....E
         51      .         .         .         .         .                                                             128

         111    .         .         .         .         .                                                    204
SRY      EAQKL.QAMHREKYPNYKYRPRRKAKMLPKNCSLL....PADPASVLCSEVQ..LDNRLYRDDCTKATHSRMEHQL..GHLPPINAASSPQQ.RDRYSHWTKL
         | ::| :   |        | || |   |   : |     : |:: : :::|   :: : |     :   | :: :    | | ::: |||  | :
HDAg     ELRRLTEEDER........RERRVAGPPPGGVNPLEGGSRGAPGGGFVPNMQGVPESPFTRTG..EGLDVRGDRGFPWDILFPSDPPFSPQSCRPQ.......
         129       .         .         .         .         .                                             214

            Similarity: 59%       Identity: 30%                   Gap Open: 4.5        Gap Extend: 0.7
```

B.

```
         57                                                                                         146
            Calmodulin binding region               leucine zipper          Calmodulin binding region
         ******** ******** ********                                                   ***************
              NLS              NLS                                                      NLS
SRY      QDRVKRPM.NAFIVWSR.DQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKL.QAMHREKYPNYKYRPRRKAKMLPKNCSLL
           | |||:  :|    | |:||::|||| :       ||||   | |:  |     :| ::| :   |        | || |   |   : |
HDAg     GPR.KRPLRGGFTDKERQDHRRRKALENKK.......KQLGAGGKNLSREEE....EELRRLTEEDER........RERRVAGPPPGGVNPL
                      ARM-like domain             leucine zipper           ARM-like domain
         84                                                                                         155
```

Fig. 1. Alignment of hepatitis delta antigen (hdvna strain) and human SRY protein. **A** Alignment between entire sequence SRY and residues 51–214 of L-HDAg. Identities are marked with bars, conservative substitutions are marked with colons. Sequences were aligned using GCG GAP program using Dayhoff PAM150 as a scoring matrix. Gap penalties as well as similarity level are listed underneath the alignment. Gap penalties are selected to achieve the best possible alignment. **B** Alignment between most similar regions of human SRY and HDAg. Lines indicate identity, colons indicate conservative substitutions. An 90-residue section of SRY protein includes an entire HMG box (residues 59 through 128) and its adjacent regions. Residues of SRY protein similar to calmodulin binding site are labeled with stars above. Residues in SRY corresponding to the 2 nuclear localization signals (NLSs) are marked with bar above the sequence (left NLS is bipartite). ARM-like domains of HDAg are underlined. Leucine zippers in each protein are positioned in the center of this alignment: residues 80 through 118 in SRY and residues 108 through 133 in HDAg. Alignment was done using GAP program from GCG package under following conditions: scoring matrix=PAM150, Gap Open Penalty=4.5, Gap Extend penalty=0.7

```
             59                                               105
A.  SRY      RVKRPM.NAFIVWSR.DQRRKMALENPRMRNSEISKQLGYQWKMLT.EAE
             | |||:  :|    | |:||::|||| :       ||||   | |: | |
    HDAg     R.KRPLRGGFTDKERQDHRRRKALENKK.......KQLGAGGKNLSREEE
             86                                              127
```

Similarity: 72%   Identity: 54%   **Zscore = 6.85**   Gap Open: 4.5   Gap Extend: 0.7

```
             57                                                                         129
B.  SRY      VQDRVKRPM.NAFIVWSR.DQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKLQAMHREKYPNYKY
                | |||:  :|    | |:||::|||| :       ||||   | |:  |     :| ::| :   |:
    HDAg     ..PR.KRPLRGGFTDKERQDHRRRKALENKK.......KQLGAGGKNLSREEE....EELRRL.TEEDER......
              85                                                                   142

             130                                                                                204
    SRY      RPRRKAKMLPKNCSLL....PADPASVLCSEVQ..LDNRLYRDDCTKATHSRMEHQL..GHLPPINAASSPQQ.RDRYSHWTKL
             | || |   |   : |     : |:: : :::|   :: : |     :   | :: :    | | ::: |||| :
    HDAg     RERRVAGPPPGGVNPLEGGSRGAPGGGFVPNMQGVPESPFTRTG..EGLDVRGDRGFPWDILFPSDPPFSPQSCRPQ.......
             143                                                                         214
```

Similarity: 59%   Identity: 33%   **Zscore = 3.0**   Gap Open :4.5   Gap Extend :1.0

```
             74                                                          128
C.  DIPA     NRELRDLCCFLDSERQ..RGRRA..ARQWQL..FGTQASRAVREDLGGCWQKLAELEGRQE
             :| ||    | | ||| | |:|   :  ||   |   ||   |:|    ::|:| : |:|
    HDAg     KRPLRG..GFTDKERQDHRRRKALENKKKQLGAGGKNLSREEEEEL....RRLTEEDERRE
             87                                                         141
```

Similarity: 63%   Identity: 45%   **Zscore = 4.68**   Gap Open:6.0   Gap Extend: 1.0

```
                  1                                                                                              81
D.  DIPA       …..MEAEAGGLEELTDEEMAALGK.EELVRRLRR.........EEATRLAALVQRGRLMQEVNRQLQGHLGEIRELKQLNRRLQAE....NRELRDLC
                       : || ||: :: :|:  | ||| | ||:         ||  :|: :  :| |    ::: :|     |   :   |:: :    :| ||
    HDAg     MSRSESKKNRGGREEVLEQWVAGRRKQEELERDLRKTKKKIKKLEEENPWLGNI..KGILGKK.DKDGEGAPPAKRARTD...RMEVDSGPRKRPLRG..
             1                                                                                                 92

             82                                                                                                175
    DIPA     CFLDSERQ..RGRRA..ARQWQL..FGTQASRAVREDLGGCWQKLAELEGRQEELLRENLALKELCLALGEEWGPRGGPSGAGGSGAGPAPELALPPCGP
              | | |||  | |:|   :  ||   |   ||   |:|    ::|:| : |:|   |   :      :|    |:||:|:|:  :     || ::
    HDAg     GFTDKERQDHRRRKALENKKKQLGAGGKNLSREEEEEL....RRLTEEDERRER..RVAGPPPGGVNPLEG..GSRGAPGGGFVPNMQGVPESPF....T
             93                                                                                                  180

             176                               202
    DIPA     RDLGDGSSSTGSVGSP.DQL........PLACSPDD
             |  |:|    |  | | | |        | :| |:
    HDAg     RT.GEGLDVRGDRGFPWDILFPSDPPFSPQSCRPQ.
             181                               214
```

Similarity: 57%   Identity: 34%   **Zscore = 2.93**   Gap Open: 5.0   Gap Extend: 1.0

```
             70
E.  DIPA     ..LQAENRELRDLCCFLD..........SERQRGRRAARQW.....QLFGTQASRAVREDLGGCWQ
               | :|:| |  |  ||:          : ||| ||| |:|     ||      ||: | :   |:
    rev      MPLGSEERRLLRLIAFLNKNNPYPPVEGTARQR.RRARRRWRQAQEQL......RALAERI...WH
             1
                                                      153
    DIPA     KLAELEGR.QEELLR..ENLALKELCLA...LGEEWGPRGGP
                   :| :|:|::  : | | :  ||   | :   | :::
    rev      ......SRVEEQLVQAIDQLVLDQQHLAIQQLPD...PPSSS
                                                      89
```

Similarity: 74%   Identity: 45%   **Zscore = 4.7**   Gap Open: 4.5   Gap Extend: 0.3

```
             1
F   rev      MAGRS..GDSDEELLKTVRLIKFLYQSNP............PPSSEGTRQARRNR..........RRR..W.....RERQ
             :|||    : : :| || : || | : ||               :||:  |:| |          |:|  :     :|||
    HDAg     VAGRRKQEELERDLRKTKKKIKKLEEENPWLGNIKGILGKKDKDGEGAPPAKRARTDRMEVDSGPRKRPLRGGFTDKERQ
             21

                                                                                        115
    rev      RHIRSISAWILSN..........YLGRPAEPVPLQLPPQRLTLDCSEDCGTSGTQGVGSPQILVESPTVLESGTKE
              | |  :   | |           |:|  |    :|  :|||    |:      :  |:|   |::   ||:|::
    HDAg     DHRRRKA...LENKKKQLGAGGKNLSREEEE...EL..RRLT....EEDERRERRVAGPPPGGVNP...LEGGSRG
                                                                                        161
```

Similarity: 63%   Identity: 39%   **Zscore = 1.95**   Gap Open: 4.7   Gap Extend: 0.7

to the first arginine rich motif (ARM) and most of the leucine zipper of HDAg (Fig. 2C). The levels of residue identity and similarity of this alignment are 45% and 63% respectively. As a control, DIPA was also aligned with HIV *rev* – an ARM containing protein unrelated to DIPA or HDAg. This alignment shows 45% identity and 74% similarity across entire length of 89-residue long *rev* protein (Fig. 2E).

The HMG box of SRY belongs to the HMG-1 subgroup. Many of the residues that are conserved between the HMG box of SRY and HMG1.2 are also conserved in the alignment between SRY and the RNA-binding domain of HDAg (data not shown). To identify the residues that are conserved among all sequences, we aligned the sequences of SRY, SOX3, HMG1.2 and the RNA-binding domain of HDAg to the profile of SRY and HDAg sequences (Fig. 3A). A total of 17 residues are identical between SOX3, SRY and HMG1.2; the alignment of the RNA-binding domain of HDAg shares 10 of these identities. The identities are concentrated in the first 2 of the 3 helices predicted for the HMG-1 subgroup (Fig. 3A, 3B).

Statistical evaluation of the sequence alignments

HDAg, SRY and DIPA contain arginine rich motifs. This compositional bias can lead to overestimation of the significance of sequence similarity in the pairwise alignments. To verify that the above HDAg alignments are not due to compositional bias of the sequences, a Monte Carlo analysis was performed for each of the pairwise alignments. One sequence was shuffled and then realigned to the other sequence under the initial conditions. In the case of partial alignments (Fig. 2A, 2C) the greatest local compositional bias is preserved, which makes it a stringent test. This process was repeated 10,000 times and the distribution of the alignment scores was examined. Matches merely due to compositional effects alone should occur at equivalent frequencies in the original and scrambled alignments. When the alignment score of the original sequence is compared to the distribution of the scores of the scrambled sequences, Z scores larger than 6.0 indicate that the original alignment is *not* simply due to compositional bias. A Z score less than 3.0 indicates that the similarities between two sequences are unlikely to be significant and are probably due to biased composition. The only potentially significant alignment is a partial alignment between the functional domains of SRY and HDAg (Fig. 2A) (Z score = 6.85). The rest of the alignments – SRY-HDAg (long alignment), DIPA-HDAg (both alignments), HDAg-*rev* and DIPA-*rev* have

◀

Fig. 2. Partial sequence alignment and Monte-Carlo process. Partial alignments were done using GAP program from GCG package with scoring matrix PAM150. Similarity level of alignment, gap penalties and Zscore based on Monte-Carlo process (see Results) are reported for each alignment. An hdvna strain of HDV is used in alignments. **A** First 47 bases of HMG box of SRY aligned with central region of HDAg. **B** An entire HMG box and remaining of SRY gene aligned with HDAg. **C** Central region of DIPA aligned with central region of HDAg. **D** An entire DIPA sequence aligned with HDAg. **E** HIV *rev* protein aligned to the central part of DIPA. **F** HIV *rev* protein (strain hv 1 mn) aligned with HDAg

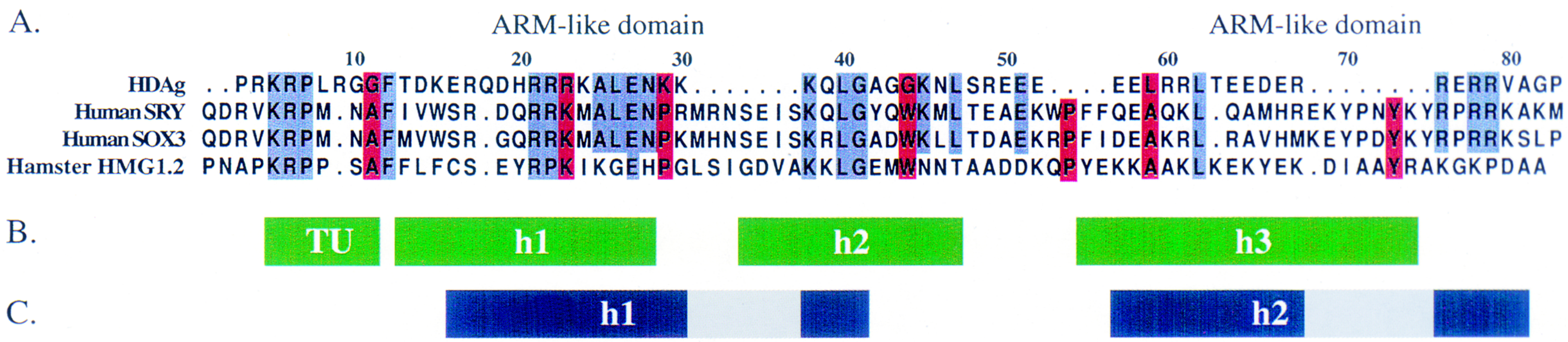
A.
ARM-like domain
ARM-like domain
10
20
30
40
50
60
70
80
HDAg ..PRKRPLRGGFTDKERQDHRRRKALENKK.......KQLGAGGKNLSREEE....EELRRLTEEDER........RERRVAGP
Human SRY QDRVKRPM.NAFIVWSR.DQRRKMALENPRMRNSEISKQLGYQWKMLTEAEKWPFFQEAQKL.QAMHREKYPNYKYRPRRKAKM
Human SOX3 QDRVKRPM.NAFMVWSR.GQRRKMALENPKMHNSEISKRLGADWKLLTDAEKRPFIDEAKRL.RAVHMKEYPDYKYRPRRKSLP
Hamster HMG1.2 PNAPKRPP.SAFFLFCS.EYRPKIKGEHPGLSIGDVAKKLGEMWNNTAADDKQPYEKKAAKLKEKYEK.DIAAYRAKGKPDAA
B.
TU
h1
h2
h3
C.
h1
h2

Fig. 3. Multiple alignment of HMG box of SRY, HMG box of SOX3, HMG1.2 and RNA-binding domain of HDAg. **A** Profilegap program was used to generated the alignment: sequences of HDAg (hdvna strain), human SOX3, human SRY and hamster HMG1 boxB were aligned to the profile of SRY and HDAg (see Results). Blue-shaded regions indicate positions with residue identity among SRY, SOX3 and HDAg or across all 4 sequences. Yellow-shaded regions indicate positions that show identity among SRY, SOX and HMG1.2, but differ from HDAg. There 17 identical residues among SOX, SRY and HMG1.2, of these 10 are also identical in HDAg. Of the seven residues in the HDAg that diverge, three show conservative substitutions (A ->.G at position 10, K ->R at position 21 and A->L at position 57). **B** Secondary structure (based on the resolved 3D structure) for HMG1.2 box in hamster. TU- terminal unit, h1–h3 are helices 1 through 3. Helix 1 is sometimes represented by two smaller helices (h1 and h1′, as seen in [38]). **C** Predicted secondary structure for the RNA-binding domain of HDAg. Helices were identified using either method of Garnier-Osgthorpe-Robson (GOR) or Chou-Fasman (CF) as described in Wang et al. [47]

Table 1. Determination of optimal gap penalty conditions for alignments

Gap penalties		Z score		Gap penalties		Z score	
Gap Open.	Gap Extend.	SRY-HDAg	DIPA-HDAg	Gap Open.	Gap Extend.	SRY-HDAg	DIPA-HDAg
3.0	0.1	6.77	3.11	5.0	0.1	6.42	3.56
	0.3	6.78	3.53		0.3	6.54	4.06
	0.5	6.73	3.94		0.5	6.45	4.35
	0.7	**6.85**	4.12		0.7	6.70	4.61
	1.0	6.6	4.44		1.0	6.23	4.63
4.0	0.1	6.6	3.42	6.0	0.1	6.20	3.63
	0.3	6.67	3.92		0.3	6.39	4.10
	0.5	6.63	3.68		0.5	6.18	4.37
	0.7	6.65	4.42		0.7	6.24	4.35
	1.0	6.48	4.61		1.0	5.89	4.50
4.5	0.1	6.26	3.57	7.0	0.1	5.97	3.78
	0.3	6.60	4.02		0.3	6.10	4.20
	0.5	6.72	4.37		0.5	6.00	4.20
	0.7	6.70	4.51		0.7	6.00	4.23
	1.0	6.49	**4.68**		1.0	5.69	4.15

Z scores as a function of gap penalties are reported here for the 47 residue long alignment of SRY and HDAg (Fig. 2A) and the 55 residue long alignment of DIPA and HDAg (Fig. 2C). Z scores were calculated using PAM 150 comparison matrix. Average and Standard Deviation values (for Z score calculation) were based on 10,000 alignments per sequence. The best Z score for each alignment is shown in boldface

Z score < 6.0 in the Monte Carlo analysis (see Fig. 2). Table 1 reports the Z scores for the optimized partial alignments of HDAg with SRY and HDAg with DIPA. In all cases the SRY-HDAg alignment is much more significant than the DIPA-HDAg alignment.

For the best partial alignments of SRY-HDAg and DIPA-HDAg we also determined the P-value of the alignment. We plotted the log of the observed frequency of occurrence for the top 0.05% of 1,000,000 random alignments as a function of their alignment score (Fig. 4). The P-value for the SRY-HDAg partial alignment (Fig. 2A) is estimated to be 7.6×10^{-7}. The P-value for the DIPA-HDAg partial alignment (Fig. 2C) is estimated to be 3.6×10^{-4}. In this experiment 326 out of 1,000,000 random alignments had better scores than the proposed DIPA-HDAg alignment; some scores were significantly better (5 alignments had scores over 50, while the actual DIPA-HDAg alignment had score 39.5). In the case of the SRY-HDAg alignment, none of the 1,000,000 random alignments exceeded the reported score of 47.8 (the highest random score was 45.9).

Testing for frameshifting during translation

SRY protein was aligned to the HDV genome in all 6 reading frames. As expected, the significant alignment is found on the antisense strand. There is no potential

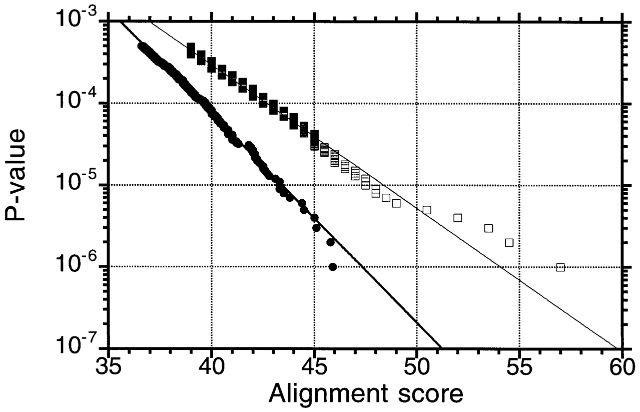

Fig. 4. P-value of partial alignments of SRY-HDAg and DIPA-HDAg. Monte Carlo analysis (as described in Materials and methods) was performed for the partial alignments displayed in Fig. 2A and 2C. The frequency of the occurrence (P) was plotted against the alignment score for the top 0.05% of the 1,000,000 randomized alignments. Open squares represent DIPA-HDAg alignment (equation for fitted line is: $y = 10^{\frac{x-19.883}{-5.703}}$), closed circles represent SRY-HDAg alignment (the equation fitted line is: $y = 10^{\frac{x-23.831}{-3.919}}$). The alignment scores were 47.8 for SRY-HDAg alignment and 39.5 for DIPA-HDAg alignment. The corresponding P values scores are: 7.6×10^{-7} for the SRY-HDAg alignments and 3.6×10^{-4} for the DIPA-HDAg alignment. The alignments were performed under the best gap penalty values (SRY-HDAg: Gap Open =3.0, Gap Extend=0.7; DIPA-HDAg: Gap Open=4.5 Gap Extend=1.0)

frameshifting detected throughout the HMG box region; all insertions/deletions are in frame.

Discussion

The discovery of sequence similarity between the RNA-binding domain of HDAg and the HMG box of SRY suggests a possible evolutionary relationship between the HMG box, a DNA-binding domain, and the RNA-binding domain of HDAg. For the sake of clarity we will discuss the relationship between HDAg and SRY, but the same argument could apply to SOX3 and, possibly, to other SRY-related genes.

HDAg-SRY sequence comparison

Statistically significant sequence similarities are limited to the functionally essential regions of both genes: in the SRY gene the similarity is confined to the HMG (High Mobility Group) box, a DNA binding motif found in the HMG protein superfamily [24, 31], in HDAg the similarity is limited to RNA-binding domain.

SRY, HDAg and DIPA are bZIP proteins - they thus possess a strong compositional bias, which influences sequence analyses. This compositional bias may reflect structural and functional constrains placed upon the RNA or DNA-binding domain, and could therefore suggest sequence convergence rather than a common evolutionary origin. Thus the high values for percent identity and similarity in compositionally biased alignments are less significant than in compositionally unbiased sequences. To address the effects of the local compositional bias we used a Monte Carlo analysis – a process that compares the alignment of interest with alignments of the randomly generated sequences of the same composition. Z scores above 6.0 indicate that the alignment of interest supports the inference of homology in spite of compositional biases. The high Z score (Z score $= 6.85$) of SRY-HDAg alignment argues that the HDAg sequence is more similar to the HMG domain of SRY than to any randomly assembled sequences of the same composition. Therefore, the compositional constraints alone upon the functional domain cannot explain the observed level of sequence similarity between HMG domain of SRY and HDAg. While it is possible that the unusual degree of sequence similarity is due to convergent evolution to an energetically favorable nucleic acid binding structure, one must keep in mind that there are a nearly unlimited number of ways to create a polypeptide that will fold into a given three-dimensional shape. Therefore there is no *a priori* reason that unrelated molecules should show significant similarity or similar spacing of conserved residues in the absence of an ancestral relationship. See Patterson [36] for a more complete development of this argument.

None of the alignments between DIPA and HDAg were found to be statistically significant based on a Monte Carlo analysis. The best Z score for an alignment is 4.68; it corresponds to the partial alignment between the 55-residue long central domain of DIPA and RNA-binding domain of HDAg (Fig. 2C). The P-value of this alignment is estimated to be 3.6×10^{-4} (Fig. 4). The E-value (expected number of comparisons achieving an equal or higher score in an equivalent database search) for this alignment is 25 for a 69,000 sequences database (size of the latest SWISS-PROT database, release 35). This E-value indicates that the database search with the RNA-binding domain of HDAg will identify approximately 25 unrelated proteins that will have alignment scores as good or better than DIPA alignment. In fact, the alignment of DIPA and the HIV *rev* protein (Fig. 2E), which is presumably unrelated to DIPA, has a similar Z score $= 4.7$. Alignments of other DIPA regions or of the entire DIPA protein with SRY produced even less significant Z scores (Fig. 2D). Statistical analysis recently performed by Long et al. considered the entire DIPA sequence; it also indicated that sequence similarity between HDAg and DIPA is not significant [34]. The P-value of the alignment of the entire DIPA and HDAg proteins is 2×10^{-3} as reported by Brazas and Ganem in their reply to Long et al. [4]. This value is less significant than the P-value for the DIPA fragment reported above. The presence of the bZIP domains, which made detection of functional interaction between DIPA and HDAg possible [5], should also serve as a warning during sequence analysis: bZIP domains are compositionally biased and this bias makes it more likely to obtain high alignment scores with unrelated proteins.

The Monte Carlo analysis of the partial alignment of the HMG domain of SRY and the RNA-binding domain of HDAg produces a Z score of 6.85 – a statistically significant value (Fig. 2A). The P-value for this alignment is 7.6×10^{-7}, indicating that an alignment this good is unlikely to arise by chance, even when analysis is limited to the compositionally biased HMG box region. E-value of this alignment is 0.05 for the latest release of SWISS-PROT (release35); thus an unrelated protein with an alignment score as good or better as that of the HMG domain of SRY and the RNA-binding domain of HDAg will be expected to be found only after 20-fold increase in the size of the current database.

A cellular gene and its RNA viral cognate diverge more rapidly than would be expected for two cellular genes due to the low fidelity of RNA replication. Because of this rapid divergence, one would expect the similarity between a DNA-encoded gene and its RNA-encoded counterpart to correspond to the regions that are absolutely essential for the function of the protein. In the case of SRY-HDAg this is indeed the case: the region of similarity is limited to the RNA-binding region of HDAg, which as been shown to be functionally important [32].

Two nuclear localization signals (NLSs) recently identified within the HMG domain of SRY are also conserved within the RNA-binding domain of HDAg [43]. These NLSs are distinct from an already identified bipartite NLS within HDAg which is located to the amino terminal side of the RNA-binding domain [49]. Our model predicts these new NLSs should be functional in HDAg. Recent experiments demonstrate that the presence of either one of the ARM sequences is sufficient for nuclear localization of the antigen, supporting our prediction [9].

HDAg-SRY structure comparison

SRY is the only gene from the Y chromosome known to be involved in testis development. It is most closely related to the SOX3 gene: these genes were proposed at one time to be alleles of a developmental gene on the undifferentiated or partially differentiated proto-X and proto-Y [21]. SRY is a member of HMG superfamily; it possesses a 79-residue long HMG box which binds to a well-defined consensus sequence: 5′-CCTTTGA and probably controls the transcription of other sex-determination genes.

The HMG superfamily is ancient – it is present in plants, yeast and animals, and is characterized by highly diverse sequences [31]. Members of the HMG family may contain single or multiple copies of the HMG box; some members bind DNA in a sequence-specific manner, others bind DNA non-specifically, and yet others recognize and bind to distorted DNA – such as 4-way junctions or cisplatin-modified DNA [38]. In all known cases, the binding of HMG box proteins causes distortion or a bend in the DNA [38]. Our model therefore predicts that HDAg will induce a similar bend when binding to the HDV genome. HDAg exhibits binding specificity which appears to be determined primarily by the structure of the HDV RNA. However some sequence specificity is required, since HDAg binds specifically to both a rod-like structure comprising the entire HDV genome, and to a multiple stem-and-loop structure formed by the fragment containing the ribozyme domain of HDV, but does not bind to the double-stranded RNA derived from an unrelated coronavirus (mouse hepatitis virus) [33].

HMG proteins can be divided into 3 major groups based on their size, sequence similarities and DNA-binding properties: HMG-1/-2, HMG-14/-17 and HMG-I/Y

[2]. The SRY gene belongs to the HMG-1/-2 family, which is represented by the HMG-1 motif and contains 2 HMG boxes referred as HMG1.1 and HMG1.2. The HMG box of SRY is more similar to HMG1.2, which is the most likely candidate for an ancestral HMG1 box [2]. Ten of 17 identities in the alignment of the HMG box of SRY with HMG1.2 are also present in its alignment with the RNA-binding domain of HDAg, and three of the remaining 7 show conservative substitutions (K -> R, A -> G, A ->L, Fig. 3). Thus the sequence constraints upon the RNA-binding domain can be traced back through the most closely related HMG box (of SRY) to the ancestral sequence of the HMG-1 family.

The predicted secondary structure of HMG-1 and the solution structure of hamster HMG1.2 have been determined [38]. The structure consists of a stretch of basic residues followed by proline at the N-terminal end, referred by the authors as the 'terminal unit', and 3 α-helices (Fig. 3). The sequence similarity between HMG-1 and the RNA-binding domain of HDAg begins within the terminal unit and continues through first and second α-helices (h1 and h2). Together, these structures comprise the side of a flat-shaped arrowhead which is proposed to lie along the minor groove and continue into the major groove of the DNA [38]. Interactions between the terminal unit and the DNA are absolutely essential [38]. The sequence of the terminal unit is conserved between SRY and HDAg, and it is also similar to regions of HMG1/Y proteins, perhaps explaining why a database search with the RNA-binding domain of HDAg showed similarity to several HMGI/Y sequences (see 'Database searches' section in Results). Helix 3 of HMG-1 comprises the other side of the arrowhead and is proposed to lie along the phosphodiester backbone of the DNA [38]; there is little sequence similarity between this helix and HDAg – there appears to be a deletion in the corresponding region of HDAg. However, the region immediately following helix 3 has a stretch of basic residues that could be aligned with the right ARM of the RNA-binding domain of HDAg (Fig. 3A). It is notable that the residues proposed to be involved in contacting DNA (terminal unit) and those whose substitution is likely to cause gross structural perturbations (e.g. replacement of the small residue at position 40, or the presence of phenylalanine at position 11, Fig. 3) are conserved among the HMG boxes of SRY, HMG-1 and the RNA-binding domain of HDAg. Gly-40 is proposed to pack closely to Phe-10 in the first helix in the resolved structure for HMG1.2 of hamster [38]. These two residues are also preserved within the RNA-binding domain of HDAg. On the other hand, many of the residues involved in tertiary folding of the HMG domain have been replaced or deleted: in the RNA binding region of HDAg the N-terminal portion of helix 1 consists of polar and charged residues, while the same region in the HMG box is hydrophobic (Fig. 3A). Another notable feature is the lack of tryptophan residues in the RNA-binding domain (SRY has 3 tryptophan residues, SOX3 has 2 tryptophan residues and HMG-1 has 1 tryptophan residue) and an overall lower percentage of the hydrophobic residues. The structure of the RNA-binding domain of HDAg has not yet been determined, but a computer generated model [47] predicts a helix-loop-helix structure in which the first helix covers regions corresponding to helix 1 and part of helix 2 of HMG-1, while the second helix of the RNA-binding domain corresponds to helix 3 of HMG-1 (Fig. 3C).

Relationship between RNA-binding and DNA-binding domains

Based on the sequence similarities discussed above we suggest that the DNA-binding domain of SRY has evolved into the RNA-binding domain of HDAg. This particular case is hardly unique: many recent publications report proteins that bind to both DNA and RNA. Among such proteins are: DNA-binding protein SSAP, which contains an RNA Recognition Motif (RRM) and specifically binds to the enhancer region of sea urchin histone H1 [13], human transcriptional activator p54nrb which binds both DNA and pre-mRNA [1], mammalian protein H16 which binds to single stranded DNA and to the corresponding RNA sequence [23], and Y-box proteins which recognize both RNA and DNA [48]. Two RNA-binding proteins have been shown to have a similar structure to that of DNA binding domains: ribosomal protein S7 appears to have a pair of helix-turn-helix motifs similar to DNA architectural factors [28] and the ribosomal protein L11 contains 3 α-helices and is similar to homeodomain proteins [50]. On the reciprocal side, the homeodomain of the *Drosophila bicoid* protein has been shown to interact in a sequence-specific manner with 3′-untranslated region of the *caudal* gene transcript [6, 14].

Both DNA-binding and RNA-binding domains often interact with nucleic acids using α-helical structures. While both the major and minor grooves of regular A-form RNA are too narrow to permit a direct interaction with an α-helix, bulges and mismatches considerably alter the groove dimensions allowing sufficient contact surface with an α-helix [41]. The HMG box has not been reported to bind RNA, however it binds both double- and single-stranded DNA [16]. Proteins containing an HMG domain have widely diverse cellular roles, indicating that an HMG domain participates in a variety of DNA-protein interactions. It is conceivable that the wide DNA-binding range of the HMG domain could extend to RNA-binding, especially considering the high rates of mutation and selection existing in the virus.

The basic stretches of the RNA-binding domain are similar to the arginine-rich motifs (ARMs) found in the products of the HIV *tat* and *rev* genes, bacterial antiterminators and coat proteins from the RNA viruses [45]. Short arginine-rich stretches tend to form α-helical structures, which can bind in the wide groove of RNA close to loops or large bulges. Only one ARM is usually necessary for binding; an arginine-rich peptide as short as 17 residues was shown to bind specifically to RNA [44, 45]. HDAg is unique in that it requires two ARMs instead of one; moreover, the spacer between the ARMs (leucine zipper) is essential as well [30, 32]. We therefore think that each ARM constitutes only a part of the RNA-binding motif, the entire motif being defined by its alignment with the HMG box. It is not clear whether the ARM regions within the RNA-binding domain of HDAg are evolutionarily related to 'classic' ARMs or whether the sequence similarity is a consequence of the functional and structural constrains imposed on the RNA/DNA-binding domain. Alignments of HDAg and ARM-containing sequences such as *rev* and *tat* show moderately high levels of similarity, but it is interspersed across the entire sequence and no continuous stretches of similarity were found with the ARM domains of HDAg. The most extended similarity found between the HIV *rev* gene and the HDAg is outside of the RNA binding domain of HDAg (Fig. 2F).

Homology of HDAg and the DNA-binding domain of HMG box may give insight into the mechanism of neoplastic transformation and progression toward cirrhosis frequently observed in chronic hepatitis infection [18]. It is possible that the RNA-binding domain of HDAg has retained some of the DNA-binding capability of its ancestral HMG box and thus could activate cellular proto-oncogenes. It has been shown that the HMG box of the SRY protein is involved in specific binding and regulation of the expression of the cellular proto-oncogene *c-fos* – FRA1 [10]. The RNA binding domain of HDAg may be capable of similar binding. While the specificity of such DNA binding by HDAg is likely to be lower than that of a true transcription factor, the sheer abundance and persistence of the antigen during chronic hepatitis infection, as well as the evolution of the HDV genome throughout the course of infection [30, 32] might be sufficient to activate a cellular proto-oncogene. In fact, the expression of the proto-oncogene *c-myc* is a hallmark of chronic HDV infection; *c-myc* protein is observed only in cells that contain HDAg [46]. Binding of HDAg to cellular DNA has not been explored. It would be interesting to test for specific binding of the RNA-binding domain of HDAg to *c-myc* and other proto-oncogenes that could be involved in the onset of cirrhosis observed in HDV patients.

HDV genome structure and the ‘capture’ hypothesis

Analysis of the HDV genome suggests that a large part of the viral genome, including the sequence coding for HDAg, is of cellular origin, while the remaining part resembles a plant viroid sequence [17, 30]. The ribozyme activities of HDV, such as self-cleavage and self-ligation, further resemble the properties of the viroid. A possible mechanism explaining the origin of HDV – the only known animal virus with a circular RNA genome – involves the capture of a cellular transcript by a viroid. The elegant capture mechanism proposed by Brazas and Ganem [5] involves the fusion of the cellular and viral transcript, and duplication of the cellular transcript during the synthesis of the antisense strand, followed by the recircularization of the entire transcript (Fig. 5). The size of the SRY protein matches the HDV size requirements: SRY is 204 residues long; duplication of the SRY RNA would result in a cellular part of HDV over 1200 bases long (the exact length will depend on the length of the non-coding regions of the SRY transcript). This agrees with the approximately 1300 bases of highly self-complementary RNA of cellular origin in HDV. The replication of HDV is conducted by cellular machinery, with the addition of HDAg. HDAg is absolutely required for in vivo HDV replication [32] and serves as a bridging factor between the viral template and cellular transcriptional complex. Coopting a single component of the cellular

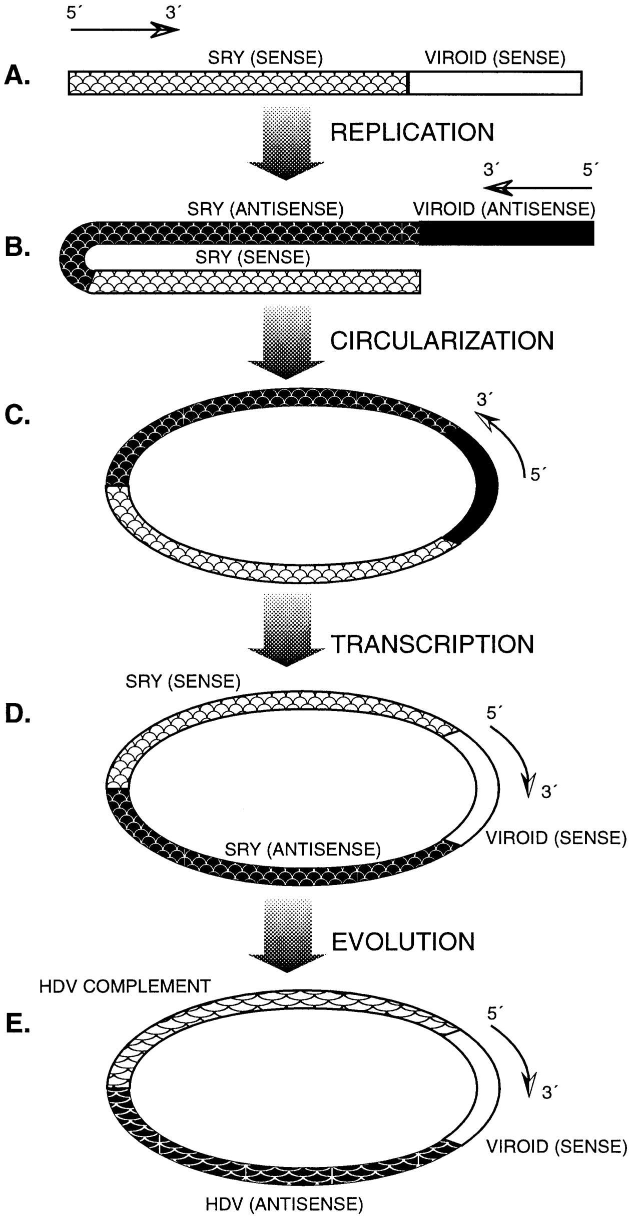

Fig. 5. Possible mechanism of HDV formation. **A** Viroid sense strand joins at the 3′-end of SRY transcript. **B** Replication begins at the 3′-end of the viroid and proceeds through SRY, synthesizing an antisense of the viroid genome and of the SRY transcript. A hairpin structure at the 5′-end of SRY template allows transcription to continue synthesizing a complement of the SRY- that is a sense version of SRY transcript. The resulting construct has an antisense strand of the viroid, followed by the antisense copy of SRY followed by the sense copy of SRY. **C** Circularization of the linear transcript. **D** Additional round of replication results in the reversal order of the regions: viroid region (sense) is followed by antisense strand of SRY, which in turn is followed by the sense strand of SRY. **E** Subsequent evolution of the genome results in the organization of domains seen in the HDV today: a viroid region is followed by an antisense of HDV antigen followed by a non-coding region which is highly complementary to the region coding for the HDV antigen

transcriptional machinery and adapting its specificity to prefer the viral template is an elegant way to ensure viral replication and survival.

It is notable that the tissue-specificity of SRY does not match that of HDV: SRY is expressed primarily in germ cells and in somatic cells of the testis [51], while HDV is restricted to liver tissue. It has been shown experimentally, however, that the replication of the HDV can proceed in the skeletal muscle [37]. A low level of SRY could be expressed in other tissues, including the liver, allowing for transcript capture, or alternatively, HDV could have evolved from an SRY-related gene specifically expressed in liver, such as SOX-18 [15]. Finally, it is possible that the expression of original viroid was not confined to liver cells and the initial capture event of cellular RNA could have occurred in a stem or germ cell where SRY is abundantly expressed [51]. This would introduce HDV into many cell types; its eventual dependence for replication on hepatitis B virus would limit its range to the liver cells.

It may never be possible to conclusively determine the origin of HDAg based on sequence similarity alone, since RNA-encoded genes evolve rapidly. However several arguments can be made in support of SRY as a cellular counterpart of HDAg. 1. The sequence similarity between SRY and HDAg corresponds to the functional domains experimentally identified. 2. Sequence similarity is restricted to functional domains – as expected for rapidly evolving RNA-encoded genes. 3. Both proteins – SRY and HDAg – bind nucleic acids and are involved in transcription. 4. Sequence and structure similarity between DNA-binding to RNA-binding motifs is supported by a number of recent publications. The above arguments make the SRY protein a strong candidate for a cellular counterpart of HDAg.

Our hypothesis predicts presence of additional nuclear localization signals within ARMs which was recently confirmed experimentally. It also predicts structural perturbation of the HDV RNA upon HDAg binding, and possible specific binding between HDAg and the regulatory regions of cellular proto-oncogenes. Testing these predictions and determining the structure of HDAg may shed further light on possible evolutionary relationship between the HMG domain of SRY and the RNA binding domain of HDAg.

Acknowledgements

We are grateful to D. W. Smith and T. L. Bailey for suggestions and critical reading of the manuscript. This work was supported by NIH/NCRR grant RR08605 and by NSF grant ASC-890285.

References

1. Basu A, Dong B, Krainer AR, Howe CC (1997) The intracisternal A-particle proximal enhancer-binding protein activates transcription and is identical to the RNA- and DNA-binding protein p54nrb/NonO. Mol Cell Biol 17: 677–686
2. Baxevanis AD, Landsman D (1995) The HMG-1 box protein family: classification and functional relationship. Nucleic Acids Res 23: 1 604–1 614

3. Branch AD, Benenfeld BJ, Baroudy BM, Wells FV, Gerin JL, Robertson HD (1989) An ultra-sensitive RNA structural element in viroid-like domain of hepatitis delta virus. Science 243: 649–652
4. Brazas R, Ganem, D. (1997) Delta-interacting protein A and the origin of the hepatitis delta virus. Science 276: 824–825
5. Brazas R, Garnem D (1996) A cellular homolog of hepatitis delta antigen: implications for viral replication and evolution. Science 274: 90–94
6. Chan C-K, Struhl G (1997) Sequence-specific RNA binding by bicoid. Nature 388: 634
7. Chao M, Hsieh SY, Taylor J (1990) Role of two forms of hepatitis delta virus antigen: evidence for a mechanism of self-limiting genome replication. J Virol 64: 5 066–5 069
8. Chen PJ, Kalpana G, Goldberg J, Mason W, Werner B, Gerin J, Taylor J (1986) Structure and the replication of the genome of hepatitis delta virus. Proc Natl Acad Sci USA 83: 8 774–8 778
9. Chou HC, Hsieh TY, Sheu GT, Lai MM (1998) Hepatitis delta antigen mediates the nuclear import of hepatitis delta virus RNA. J Virol 72: 3 684–3 690
10. Cohen DR, Sinclair AH, McGovern JD (1994) SRY protein enhances transcription of Fos-related antigen 1 promoter constructs. Proc Natl Acad Sci USA 91: 4 372–4 376
11. Collins JF, Coulson AFW, Lyall A (1988) The significance of protein sequence similarities. CABIOS 4: 67-71
12. Dayhoff MO, Schwartz RM, Ocrutt BC (1978) A model of evolutionary changes in protein. In: Dayhoff MO (ed) Atlas of protein sequence and structure. National Biomedical Research Foundation, Washington, vol 5 Suppl 3, pp 345–358
13. DeAngelo DJ, DeFalco J, Rybacki L, Childs G (1995) The embryonic enhancer-binding protein SSAP contains a novel DNA-binding domain which has homology to several RNA-binding proteins. Mol Cell Biol 15: 1 254–1 264
14. Dunbau J, Struhl G (1996) RNA recognition and translational regulation by a homeodomain protein. Nature 379: 694–699
15. Dunn TL, Mynett-Johnson L, Hoskin BM, Koopman PA, Muscat GEO (1995) Sequence and expression of Sox-18, a new HMG-box transcription factor. Gene 161: 223–225
16. Einck L, Bustin M (1985) The intracellular distribution and function group of the high mobility chromosomal proteins. Exp Cell Res 156: 295–310
17. Elena SF, Dopazo J, Flores R, Diener TO, Moya A (1991) Phylogeny of viroids, viroidlike satellite RNAs, and the viroidlike domain of hepatitis delta virus RNA. Proc Natl Acad Sci USA 88: 5 631–5 634
18. Fattovich G, Boscaro S, Noventa F, Pornaro E, Stenico D, Alberti A, Ruol A, Realdi G (1987) Influence of hepatitis delta virus infection on progression to cirrhosis in chronic hepatitis type B. J Infect Dis 155: 931–935
19. Felsenstein J (1973) Maximum-likelihood estimation of evolutionary trees from continuous characters. Am J Hum Genet 25: 471–492
20. Feng DF, Doolittle RF (1987) Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J Mol Evol 25: 351–360
21. Foster JW, Graves JA (1994) An SRY-related sequence of the marsupial X chromosome: implications for the evolution of the mammalian testis-determining gene. Proc Natl Acad Sci USA 91: 1 927–1 931
22. Fu TB, Taylor J (1993) The RNAs of hepatitis delta virus are copied by RNA polymerase II in nuclear homogenates. J Virol 67: 6 965–6 972
23. Gaillard C, Cabannes E, Strauss F (1994) Identity of the RNA-binding protein K of hnRNP particles with protein H16, a sequence-specific single strand DNA-binding protein. Nucleic Acids Res 22: 4 183–4 186

24. Giese K, Cox J, Grosschedl R (1992) The HMG domain of lymphoid enhancer factor 1 bends DNA and facilitates assembly of functional nucleoprotein structures. Cell 69: 185–195
25. Glen JS, White JM (1991) Trans-dominant inhibition of human hepatitis delta virus genome replication. J Virol 65: 2 357–2 361
26. Gribskov M, McLachlan AD, Eisenberg D (1987) Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci USA 84: 4 355–4 358
27. Gribskov M, Veretnik S (1996) Identification of sequence pattern with profile analysis. Methods Enzymol 266: 198–212
28. Hosaka H, Nakagawa A, Tanaka I, Harada N, Sano K, Kimura M, Yao M, Wakatsuki S (1997) Ribosomal protein S7: a new RNA-binding motif with structural similarities to a DNA architectural factor. Structure 5: 1 199–1 208
29. Karlin S, Altschul SF (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring scheme. Proc Natl Acad Sci USA 87: 2 264–2 268
30. Lai MM (1995) The molecular biology of˙hepatitis delta virus. Ann Rev Biochem 64: 259–268
31. Laudet V, Stehelin D, Clevers H (1993) Ancestry and diversity of the HMG box superfamily. Nucleic Acids Res 21: 2 493–2 501
32. Lee CZ, Lin JH, Chao M, McKnight K, Lai MM (1993) RNA-binding activity of hepatitis delta antigen involves two arginine-rich motifs and is required for hepatitis delta virus RNA replication. J Virol 67: 2 221–2 227
33. Lin JH, Chang MF, Baker SC, Govindarajan G, Lai MMC (1990) Characterization of hepatitis delta antigen: specific binding to hepatitis delta virus RNA. J Virol 64: 4 051–4 058
34. Long M, de Sourza SJ, Gilbert W (1997) Delta -interacting protein A and the origin of hepatitis delta virus. Science 276: 824–825
35. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48: 443–453
36. Patterson C (1988) Homology in classical and molecular biology. Mol Biol Evol 5: 603–625
37. Polo JM, Govindarajan S, Lai MM (1995) Replication of the hepatitis delta virus RNA in mice after intramuscular injection of plasmid DNA. J Virol 69: 5 203–07
38. Read CM, Cary PD, Crane-Robinson C, Driscoll PC, Norman DG (1993) Solution structure of a DNA-binding domain from HMG1. Nucleic Acids Res 21: 3 427–3 436
39. Robertson HD (1992) Replication and evolution of viroid-like pathogens. Curr Topics Microbiol Immunol 176: 213–219
40. Sinclair AH, Berta P, Palmer MS, Hawkins JR, Griffiths BL, Smith MJ, Foster JW, Frischauf AM, Lovell-Badge R, Goodfellow PN (1990) A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. Nature 346: 240–244
41. Siomi H, Dreyfuss G (1997) RNA-binding proteins as regulators of gene expression. Curr Opin Genet Devel 7: 345–353
42. Smith TF, Waterman MS (1981) Comparison in biosequences. Adv Appl Math 2: 482–489
43. Sudbeck P, Scherer G (1997) Two independent nuclear localization signals are present in the DNA-binding high-mobility group domains of SRY and SOX9. J Biol Chem 272: 27 848–27 852
44. Tan R, Chen L, Buettner JA, Hudson D, Frankel AD (1993) RNA recognition by an isolated a helix. Cell 73: 1 031–1 040

45. Tan R, Frankel AD (1995) Structural variety of arginine-rich RNA-binding peptides. Proc Natl Acad Sci USA 92: 5 282–5 286
46. Tappero G, Natoli G, Anfossi G, Rosina F, Negro F, Smedile A, Bonino F, Angeli A, Purcell RH, Rizzetto M, Levrero M (1994) Expression of the c-myc protooncogene product in cells infected with hepatitis delta virus. Hepatology 20: 1 109–1 114
47. Wang JG, Jansen RW, Brown EA, Lemon SM (1990) Immunogenic domains of hepatitis delta virus antigen: peptide mapping of epitopes recognized by human and woodchuck antibodies. J Virol 64: 1 108–1 116
48. Wolffe AP (1994) Structural and functional properties of the evolutionary ancient Y-box family of nucleic acid binding proteins. Bioassays 16: 245–251
49. Xia YP, Yeh CT, Ou JH, Lai MMC (1992) Characterization of nuclear targeting signal of hepatitis delta antigen: nuclear transport as a protein complex. J Virol 66: 914–921
50. Xing Y, GuhaThakurta D, Draper DE (1997) The RNA binding domain of ribosomal protein L11 is structurally similar to homeodomains. Nat Struct Biol 4: 24–27
51. Zwingman T, Fujimoto H, Lai LW, Boyer T, Ao A, Stalvey JR, Blecher SR, Erickson RP (1994) Transcription of circular and noncircular forms of SRY in mouse testes. Mol Reprod Dev 37: 370–381

Authors' address: Dr. S. Veretnik, San Diego Supercomputer Center, P. O. Box 85608, San Diego, CA 92186–5608, U.S.A.

Received June 16, 1998