Nucleic Acids Research, 1994, Vol. 22, No. 20 ***4163–4166***

Eukaryotic RNAse H shares a conserved domain with caulimovirus proteins that facilitate translation of polycistronic RNA

Arcady R.Mushegian*, Herman K.Edskes and Eugene V.Koonin[1]
Department of Plant Pathology, University of Kentucky, Lexington, KY 40546-0091 and
[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bld. 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received June 30, 1994; Revised and Accepted September 6, 1994

ABSTRACT

RNAse H (RNH1 protein) from the trypanosomatid *Crithidia fasciculata* has a functionally uncharacterized N-terminal domain dispensable for the RNAse H activity. Using computer methods for database search and multiple alignment, we show that the N-terminal domains of RNH1 and its homologue encoded by a cDNA from chicken lens are related to the conserved domain in caulimovirus ORF VI product that facilitates translation of polycistronic virus RNA in plant cells. We hypothesize that the N-terminal domain of eukaryotic RNAse H performs an as yet uncharacterized regulatory function, possibly in mRNA translation or turnover.

INTRODUCTION

Ribonuclease H (RNAse H) specifically degrades the RNA strand of an RNA:DNA heteroduplex. This enzyme has been found in both prokaryotes and eukaryotes (1–4) and is also encoded by the genomes of retroviruses, pararetroviruses and some retroelements where it is a distinct domain in a larger pol polyprotein (1, 5, 6). RNAse H-mediated degradation of RNA in RNA:DNA hybrids is an essential stage in the virus life cycle (1, 3, 6). The bacterial RNAse H has been shown to control the formation of RNA primer at the origin of DNA replication of the *Escherichia coli* plasmid ColE1 (7). In both prokaryotes and eukaryotes, RNAse H is thought to play a role in the removal of RNA primers from Okazaki fragments to enable their ligation (8).

Comparative analysis of amino acid sequences has revealed several conserved motifs in the RNAse H domains of prokaryotic, eukaryotic and viral origin (4, 7, 9–11). RNAse H (*RNH1* gene product) from the trypanosomatid *Crithidia fasciculata* is the only eukaryotic RNAse H for which complete nucleotide sequence is available (4). The 53.7 kDa RNH1 protein is much larger than prokaryotic RNAse H proteins and viral RNAse H domains, and it has been shown that only the C-terminal half of RNH1 is required for *in vitro* RNAse H activity or for rescue of *E.coli* RNAse H-deficient mutant (4). Accordingly, sequence conservation with other RNAses H is confined to the carboxyl half of the protein (4).

We show here that a domain within the N-terminal half of the RNH1 is related to a conserved domain in a virus-specific protein involved in the *trans*-activation of translation of downstream cistrons in the polycistronic mRNA of caulimoviruses, a group of plant pararetroviruses.

Amino acid and nucleotide sequences were from the SWISS-PROT, PIR and GenBank databases that are combined in the Non-Redundant sequence DataBase (NRDB) at the National Center for Biotechnology Information (NIH). Peanut chlorotic streak caulimovirus sequence was from (12). Amino acid sequences were compared with the NRDB using BLASTP and TBLASTN programs based on the BLAST algorithm (13, 14). Compositionally biased regions of query sequences were excluded from the analysis using the SEG program (14, 15). Database screening for amino acid patterns was performed using the PAST program (R.L.Tatusov, unpublished). Search for conserved motifs was also done using the MoST method which transforms ungapped alignment blocks into position-dependent weight matrices used for database scanning (16). Multiple alignments were constructed using the MACAW program (17). Protein secondary structure was predicted using the PHD program that has been reported to yield an accuracy of over 70% (18).

Database search with the *Crithidia* RNH1 amino acid sequence revealed highly significant similarities to RNAses H from bacteria and yeast in the C-terminal part of the protein. In addition, an uncharacterized protein encoded by a cDNA from chicken lens (accession number D26340) showed high similarity to RNH1 in both the C-terminal RNAse H domain and in the N-terminal domain (probability of matching by chance, $P < 10^{-8}$). This protein is likely to be the chicken homologue of RNH1. Moderate but also statistically significant similarity ($P < 10^{-3}$) was observed between the N-terminal domains of the trypanosomatid and chicken RNH1 proteins and the ORF VI products of caulimoviruses. The most conserved sequence spanned about 40 amino acid residues within the domain that shows highest conservation among ORF VI products of different

*To whom correspondence should be addressed at: Department of Microbiology, University of Washington, Seattle, WA 98195, USA

```
                            I                     II                                  III
                     ____________        [____________________]                 _______________
consensus            ...K.....UU..       OOUU..G...GUO..W.........U.....K.O.....A....
ORF      H.i.    86  AAGKVAVFGLLKR 56    LFAVKQNHINGIENFWSQAKRILRKYNGIDRKNFPLFLKECEFR  16
RNAseH  chick(24)                        FYAVRKGRQTGVYRTWAECQQQVNRFPSASFKKFATEKEAWAFV 228
RNAseH  C.f.     89  AQVKSSVNQLAIP 53    FYVVAVGRQRGIYSTWDQCSEQVKGFSGAVYKSFRTLSEARAYL 295
                     *  * * * *          :***  *   ***  *          * *  ** * *::****
ORF VI  CaMV     71  APGKESTNPLMAS 55    YYVVYNGPHAGIYDDWGCTKAATNGVPGVAYKKFATITEARAAA 339
ORF VI  CERV     55  AIGKESSNPLMAI 65    FYVVYNGPYAGIYDHWGTAKKATNKIPGVSYKKFKDMLSARTSA 319
ORF VI  FMV      55  AQGKETPNPVKAD 87    WFAVYKGPNKEFFTEWEIVADICKK--RQKSKRFRSKEQAEVSI 315
ORF VI  PClSV    53  SIKKEQPEKLVIQ 14    YYVIYQGPGKGIYDEWGKASLFITGVKGIRHKKFLSKKEAQDSF 297
ORF VI. SoCMV    19  EQQKQTLNSLISR 73    AYVIFDGPWKGIYQDWHIVKQTVN-AQPYRYKGYNSLDEAKLAH 310
sec. struct.         lllllllllhhhh       bbbbblllllbb????lllllllllllhhhhhhhhhhhhhhh
```

Figure 1. Conserved motifs in the eukaryotic RNH1 proteins and in caulimovirus translational *trans*-activator proteins. The alignment was constructed using the MACAW program, with the boundaries of conserved blocks adjusted to achieve maximum statistical significance. The number of amino acid residues separating the blocks and the number of residues between the blocks and the protein ends is indicated; the putative chicken RNAse H sequence is apparently incomplete at the N-terminus. Asterisks show identities and colons show conservative substitutions between the *Crithidia* RNH1 sequence and the CaMV ORF VI product. The consensus shows the conserved amino acid residues, with a possible exception of one virus sequence. O designates an aromatic residue (F, Y, or W); U designates a hydrophobic residue (I, L, V, M, F, or A); $ designates serine or threonine; and dot designates any residue. Amino acid residues that conform to the consensus are shown by bold type. The predicted secondary structure is the consensus of the predictions for individual sequences (a designates α-helix; b designates β-strand; l designates loop; and ? indicates that the prediction was uncertain). The sequences were from the PIR database: S16288—putative protein encoded by a *Haemophilus influenzae* (H.i) insertion sequence; A48683—*Crithidia fasciculata* (C.f.) RNAse H; GenBank database: D26340—putative chicken RNAse H; X15828—soybean chlorotic mottle virus (SoCMV) ORF VI; SWISS-PROT database: P03558—cauliflower mosaic virus (CaMV), strain CM-1841 ORF VI; P05401—carnation etched ring virus (CERV) ORF VI; P09524—figwort mosaic virus (FMV) ORF VI; and ref. 12—peanut chlorotic streak virus (PCLSV) ORF VI. In the SoCMV sequence, an apparent frameshift error is corrected, as in (24).

caulimoviruses. Notably, between the two RNH1 proteins, this region was even more highly conserved than the RNAse H domain (data not shown). Subsequent multiple alignment analysis revealed three motifs, with the strongest conservation in motif II (Fig. 1). It has been shown that RNAses H from baker's and fission yeast also contain, in their N-terminal part, at least two of these motifs (Cerritelli, S., Shin, D. Y., and Crouch, R. J. Abstracts of the INSERM/NIH Conference on Antisense Oligonucleotides and Ribonucleases H, Arcachon, 1992; R. J. Crouch, personal communication). The fact the similarity to caulimovirus ORF VI product is shared by the RNH1 proteins from the trypanosomatid, yeast, and chicken suggests that this conserved domain is characteristic of eukaryotic RNAses H in general.

Amino acid sequence pattern $OUUx_2Gx_4UOx_2Wx_{13-15}KxOx_5A$, where O designates an aromatic residue (F, Y, or W), U designates a hydrophobic residue (I, L, V, M, F, or A), and x designates any residue, is unique for the RNH1 proteins and caulimovirus ORF VI products. Secondary structure prediction suggested a mixed α/β structure for the conserved domain, with the most highly conserved motif II apparently forming a β-hairpin with the two glycines in the loop (Fig. 1).

Database screening with an alignment block including motif II and III using the MoST program revealed a related segment in an uncharacterized protein encoded by a *Haemophilus influenzae* insertion sequence (19), with the probability of detecting by chance of about 0.04. Further analysis using the MACAW program showed that in a three-way comparison of the sequences of *Crithidia* RNAse H, CaMV ORF VI product, and the *Haemophilus* protein, the probability of finding the alignment of motifs II and III by chance was about 10^{-5}; in addition, a counterpart to motif I was identified in the same position relative to the N-terminus and motif II as in the RNAse H (Fig. 1). Thus it is likely that this bacterial protein is a homologue of the N-terminal domains of the eukaryotic RNAse H and caulimovirus ORF VI product.

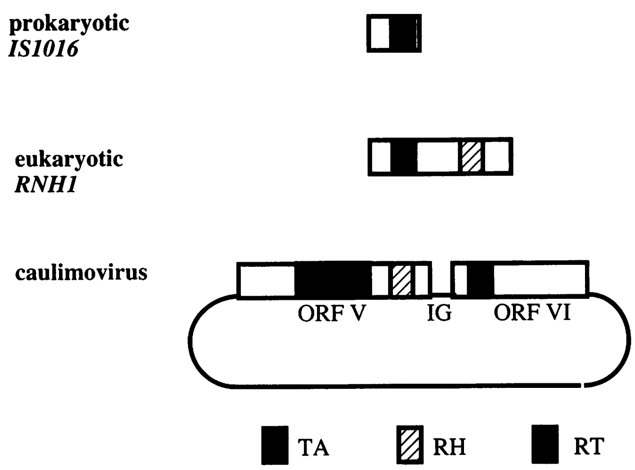

Figure 2. Relative positions of the RNAse H and *trans*-activator domains in the proteins encoded by *Haemophilus* insertion sequence, by eukaryotic *RNH1* genes, and by the caulimovirus genomes. Homologous domains are shown by identical shading. TA, *trans*-activator domain (mini-TAV); RH, RNAse H; RT, reverse transcriptase; IG, intergenic region between ORFs V and VI in CaMV. The ca. 8 kbp circular genome of CaMV is drawn not to scale.

Caulimoviruses replicate their DNA genome via reverse transcription of a genome-length RNA transcript that also serves as a polycistronic mRNA for expression of several virus genes. ORF VI is expressed from a subgenomic RNA. Its protein product has been first recognized as the major constituent of virus inclusion bodies in infected cells (20). Importantly, the ORF VI proteins from various caulimoviruses also serve as positive regulators of the expression of multiple virus genes from the genome-length polycistronic RNAs (21, 22). Moreover, bicistronic reporter constructs efficiently express both reporter genes in plant cells in the presence of cauliflower mosaic virus

ORF VI product provided *in trans* (23). Thus ORF VI protein, also called TAV, for Trans-Activator of Virus gene expression (23) ensures efficient translation of the downstream genes from almost any polycistronic mRNA. Caulimovirus ORF VI product is the only known protein with such activity.

Deletion analysis of CaMV ORF VI has shown that an N-terminal fragment of ca. 120 amino acids ('mini-TAV') retains substantial translation activation capacity (24). It has been noticed that this segment contains the most conserved block in the otherwise divergent sequences of ORF VI proteins (24). As shown in Fig. 1, it is this block that is also conserved in RNH1 proteins.

The mechanism of the ORF VI-mediated translation of downstream cistrons from the polycistronic mRNAs is unknown. Binding of ORF VI products of CaMV and figwort mosaic virus to RNA *in vitro* is weak, whereas mini-TAV does not bind to RNA at all (24, and H.K.E, unpublished observations). On the other hand, TAV has been found in complexes with polysomes (6). Specific protein−protein interactions may be involved in the ORF VI protein activity as has been demonstrated, for example, for another unusual mechanism of viral RNA translation, the internal initiation of translation in picornaviruses (25). It remains to be determined whether the conserved motifs identified here belong to a protein-binding domain or, for example, to an RNA-binding domain whose function may be facilitated by other domains.

In eukaryotes and eukaryotic viruses, RNAse H typically is a C-terminal portion of a larger protein. In retroviruses, pararetroviruses, and retroelements, it is fused to the C-terminus of the reverse transcriptase (RT; 2, 5, 11). In caulimoviruses, RT and RNAse H domains are domains of the polyprotein encoded by the ORFV, which is located upstream if the ORFVI (Fig. 2). The retrovirus RT domain is required for the full activity of RNAse H , apparently through mediating interactions of RNAse H with its substrate (26, 27). By analogy, it may be speculated that the N-terminal, TAV-related domain of the eukaryotic RNAse H targets it to a specific substrate, distinct from the RNA primers in Okazaki fragments. It is even possible that RNH1 has a cytoplasmic function, given the high levels of RNAse H activity observed in some eukaryotic cell- and nuclear-free RNA translation systems (28, 29). The similarity to the caulimovirus *trans*-activator of gene expression suggests that eukaryotic RNAse H may perform yet unidentified, regulatory role in mRNA translation or turnover. Interestingly, the identification of the putative mini-TAV homologue in *Haemophilus* suggests that this domain may exist also as a stand-alone protein (Fig. 2).

The *RNH1* gene and the caulimovirus genome each encode both the RNAse H domain and the mini-TAV domain. The arrangement of the two domains is, however, different. While in RNH1 they are expressed as a single protein, in caulimoviruses they belong to two different proteins. Moreover, the coding sequences for the two domains are swapped in caulimoviruses as compared to the RNH1 genes (Fig. 2).

The combination of the RNAse H and the mini-TAV domains in the caulimovirus genome could result from the capture of a cellular RNAse H gene that already contained both domains, followed by a rearrangement. Alternatively, the two domains could be captured by an ancestor virus independently. Comparative analysis of genome organization and amino acid sequences favors the latter possibility. Two groups of plant pararetroviruses, caulimoviruses and badnaviruses, are currently known. Badnaviruses lack a recognizable equivalent of ORF VI and appear to use distinct strategies for translation of their mRNA (6, 30). In contrast, the essential RNAse H domain is shared by all retroviruses, pararetroviruses and many retroelements. The overall conservation among the amino acid sequences of RNAses H is limited to only a few functionally important motifs (4, 11), precluding construction of a reliable phylogenetic tree. Nevertheless, we observed that the caulimovirus RNAse H domain showed significant similarity to the homologous domains of badnaviruses and animal pararetroviruses from the hepadnavirus group, but not to non-viral RNAses H (data not shown). Therefore it is likely that the common ancestor of pararetroviruses encoded RNAse H domain but not the mini-TAV domain which was acquired by the caulimovirus lineage after caulimovirus−badnavirus divergence. Thus, a cellular gene and a viral genome might have evolved independently towards possessing the same pair of related domains.

ACKNOWLEDGEMENTS

We thank Dr R.J.Crouch for communicating his results prior to publication, Drs D.Ray and A.Telesnitsky for helpful discussions, and Dr R.L.Tatusov for the PAST program. A.R.M and H.K.E. are grateful to Dr R.J.Shepherd for constant support and encouragement.

REFERENCES

1. Temin, H.M. (1985) *Mol. Biol. Evol.*, **2**, 455−468.
2. Temin, H.M. (1989) *Nature*, **339**, 252−255.
3. Crouch, R.J. (1990) *New Biol.*, **2**, 771−777.
4. Campbell, A.G., and Ray, D.S. (1993) *Proc. Natl. Acad. Sci. USA*, **90**, 9350−9354.
5. Xiong, Y, and Eickbush, T.H. (1990) *EMBO J.*, **9**, 3353−3362.
6. Rothnie, H.M., Chapdelaine, Y., and Hohn, T. (1994) *Adv. in Virus Res.* **44**, in press.
7. Itoh, T., and Tomizawa, J. (1980) *Proc. Natl. Acad. Sci. USA*, **77**, 2450−2454.
8. Goulian, M., Richards, S.H., Heard, C.J., and Bigsby, B.M. (1990) *J. Biol. Chem.*, **261**, 18461−18471.
9. Johnson, M.S., McClure, M.A., Feng, D.-F., Gray, J., and Doolittle, R.F. (1986) *Proc. Natl. Acad. Sci. USA*, **83**, 7648−7652.
10. Doolittle, R.F., Feng, D.-F., Johnson, M.S., and McClure, M.A. (1989) *Q. Rev. Biol.*, **64**, 1−30.
11. McClure, M.A. (1991) *Mol.Biol.Evol.*, **8**, 835−856.
12. Richins, R.D. (1993) Ph.D. Thesis. University of Kentucky.
13. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) *J. Mol. Biol.*, **215**, 403−410.
14. Altschul, S.F., Boguski, M.S., Gish, W., and Wooton, J.C. (1994) *Nature Genetics*, **6**, 119−129.
15. Wooton, J.C., and Fedrehen, S. (1993) *Comput.Chem.*, **17**, 149−163.
16. Tatusov, R. L., Altschul, S. F., and Koonin, E. V. (1994) *Proc. Natl. Acad. Sci. USA*, in press.
17. Schuler, G.D., Altschul, S.F., and Lipman, D.J. (1991) *Proteins: Struct. Funct. Genet.*, **9**, 180−190.
18. Rost, B., and Sander, C. (1993) *J. Mol. Biol.*, **232**, 584−599.
19. Kroll, J. S., Loynds, B. M., and Moxon, E. R. (1991) *Mol. Microbiol.* **5**, 1549−1560.
20. Odell, J.T., and Howell, S.H. (1980) *Virology* **102**, 349−359.
21. Bonneville, J.M., Sanfacon, H., Futterer, J., and Hohn, T. (1989) *Cell*, **59**, 1135−1143.
22. Scholthof, H.B., Gowda, S., Wu, F.C., and Shepherd, R.J. (1992) *J. Virol.*, **65**, 5190−5195.
23. Futterer, J., and Hohn, T. (1991) *EMBO J.*, **10**, 3887−3996.
24. De Tapia, M., Himmelbach, A., and Hohn, T. (1993) *EMBO J.*, **12**, 3305−3314.

25. Svitkin, Y.V., Meerovitch, K., Lee, H.S., Dholakia, J.N., Kenan, D.J., Agol, V.I., and Sonenberg, N. (1994) *J. Virol.*, **68**, 1544–1550.
26. Yang, W., Hendrikson, W.A., Crouch, R.J., and Satow, Y. (1990) *Science*, **249**, 1398–1405.
27. Kohlstaedt, L.A., Wang, J., Rice, P.A., Friedman, J.M., and Steitz, T.A. (1993) In Skalka, A.M., and Goff, S.P. (eds.). Reverse Transcriptase. Cold Spring Harbor Laboratory Press. pp. 223–249.
28. Walder, R.Y., and Walder, J.A. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 5011–5015.
29. Cazenave, C., Frank, P., and Busen, W. (1993) *Biochimie*, **75**, 113–122.
30. Futterer, J., Potrykis, I, Valles Brau, M.P., Dasgupta, I., Hull, R., and Hohn, T. (1994) *Virology*, **198**, 663–670