R. Giegerich et al. (eds.)
Computer Science and Biology. GCB'99 Hannover, Germany. Univ. Bielefeld 1999, pp. 73-79

CONSERVED SECONDARY STRUCTURES IN HEPATITIS B VIRUS RNA

Roman Stocsits[a,*], Ivo L. Hofacker[a], Peter F. Stadler[a,b]

[a]Institut für Theoretische Chemie, Universität Wien
Währingerstraße 17, A-1090 Wien, Austria
Phone: **43 1 4277 52737, 52793 (fax) E-Mail:
{roman,ivo,studla}@tbi.univie.ac.at
[*]Address for correspondence
[b]The Santa Fe Institute
1399 Hyde Park Road, Santa Fe, NM 87501, USA

1. Introduction

Almost all RNA molecules form secondary structures. The presence of secondary structure in itself therefore does not indicate any functional significance. Extensive computer simulations [4, 14] showed that a small number of point mutations is very likely to cause large changes in the secondary structures. A difference in the nucleic acid sequence of only 10% leads almost surely to unrelated structures if the mutated sequence positions are chosen randomly. Secondary structure elements that are consistently present in a group of sequences with less than, say 95%, average pairwise identity are therefore most likely the result of stabilizing selection, not a consequence of the high degree of sequence homology.

If selection acts to preserve a structural element then it must of course have some function. This observation can be used to design an algorithm that reliably detects conserved RNA secondary structure elements in a small sample of related RNA sequences [6]. Of course, we cannot tell what the function of the conserved structure elements might be. Nevertheless, knowledge about their location can be used to guide, for instance, deletion studies [11].

Recently, we have proposed a suit of methods termed `alidot/pfrali` that aims at utilizing the information contained in a multiple alignment of a small set of related sequences to extract conserved features from the pool of plausible structures generated by thermodynamic prediction for each sequence [6, 7].

Our approach is different from efforts to simultaneously compute alignment and secondary structures [2, 13, 16], because it does not assume that the sequences have a single common structure and not just a few conserved structural features. `Alidot/pfrali` combines structure prediction and *motif search* [3]. Programs such as `RNAmot` [5] or S. Eddy's `RNAbob` scan a database of sequences for RNA motifs that are specified in terms of sequence as well as secondary structure constraints. In contrast, our approach does not require any prior knowledge about the structural motifs: their structures are predicted during the search.

The performance of `alidot/pfrali` depends not only on the quality of the RNA secondary structure prediction but also on the quality of the multiple sequence alignment that is used as an input. The procedure is fairly robust w.r.t. small alignment errors, and `Clustal W` alignments of closely related nucleic acid sequences can be used "as is" in many cases. Nevertheless, it has become obvious that improved alignments may reduce the signal-to-noise ratio quite significantly.

2. Nucleid Acid Alignments Based on Amino Acid Alignments

The sequence heterogenity on the level of nucleic acids makes good alignments hard to obtain even in the case of phylogenetically closely related sequences. On the other hand, the nucleic acid sequences in our applications to viral genomes mostly code for proteins. The redundancy of the genetic code allows in the extreme case a silent mutation with all three codon positions different. Protein sequences hence may still show substantial homology when the corresponding nucleic acid sequences are already essentially randomized.

In order to alleviate the problems of nucleic acid alignments we have designed the program `ralign` which combines amino acid based alignment of the coding regions with nucleic acid based alignment of non-coding regions.

`Ralign` processes `GenBank` nucleic acid files. If the `GenBank` file contains information on ORFs, exons, introns, or the protein sequence after translation, it is used by `ralign`. Otherwise `ralign` attempts to determine ORFs that exceed a predefined minimal length. The detected coding regions are translated, and the resulting proteins are compared to the protein sequences in the `GenBank` file, if available.

Overlapping open reading frames are quite frequent in virus genomes. If a certain part of the sequence is coding for two or three proteins, a decision has to be made

```
ADI-MAL      AU-AGUACAUGGCAGAAUAAUGGUGC----AAGA------CU-AA--GU----AAUAGCACAGAGUC---AACUGGUAGUAUCACACUCCCAUG
AE-90CF402   AU-AGUACUUGGAUA---------------AAUGGAACCAUGCAGGAGGUU--AAUGGCACAAACUC---A---GGCAAUAUCACACUUCCAUG
B-896        AU-AGUACUUGGAAU-------G-------U-UA------CUGGAGGGACA--AAUGGCACUGAAGG---AAAUGACAUAAUCACACUCCAAUG
B-ACH320A    AU-AGUACUUGG------AAUGAUACUGGGAAUGUUA---CUGAAAGGUCA--AAUAACAAUGA------AAAU------AUCACACUCCCAUG
B-D31        AU-AGUACUUGGAAU------------------GAUA---CUAAAGAGUCA--AAUAACACAAAU---------GGAACUAUCACACUCCCAUG
B-JRCSF      AU-AGUACUUGGAAU-------G-------A-UA------CUGAAAAGUCA--AGUGGCACUGAAGG---AAAUGACACCAUCAUACUCCCAUG
B-LAI        AU-AGUACUUGGUUU---AAUAGUACUUGGAGUA------CUGAAGGGUCA--AAUAACACUGAAGG---AAGUGACACAAUCACACUCCCAUG
B-MANC       AU-AGUACUUGGAAUACUGGG---------AAUGAUA---CUAGAGAGUCA--AAUGACACAAAUAA---UACUGGAAAUAUCACACUCCCAUG
B-OYI        AU-AGUACUUGGAAU------------------GAUA---CUACAAGGGCA--AAUAGCACUGAA---------GUAACUAUCACACUCCCAUG
B-YU2        -------CUUGG------AAUGAUACUAGAAA---------------GUUA--AAUAACACUGGAAG---AAAU------AUCACACUCCCAUG
D-NDK        AU-AGUACAUGGAAU----------CA--GACUAAUAG---UACAGGGUUC--AAUAAUGGCACAG---------------UCACACUCCCAUG
O-ANT70      AUUA-UACCUUU-UCA----------UGUAACGGAACCACCUGUAGUGUUAGUAAUGUUAGUCAAGG------UAACAAUGGCACUCUACCUUG
SIVCPZGAB    CUGACAACAUUA-------------------------------------CA--AAUGGCAUU------------------AUAAUACUGCCAUG


ADI-MAL      AUAGUACAUGGCAGAAUAAUGGU---GCAAGACUAAGUAAUAGCACAGAGUCAACUGGU------------------------AGUAUCACACUCCCAUG
AE-90CF402   AUAGUACUUGGAUA------------AAUGGAACCAUGCAGGAGGUUAAUGGCACAAACUCA------------------GGCAAUAUCACACUUCCAUG
B-896        AUAGUACUUGGAAU------------------GUUACUGGAGGGACAAAUGGCACUGAAGGA---------------AAUGACAUAAUCACACUCCAAUG
B-ACH320A    AUAGUACUUGGAAUGAUACUGGG------AAUGUUACUGAAAGGUCAAAUAACAAUGAA------------------------AAUAUCACACUCCCAUG
B-D31        AUAGUACUUGGAAU------------------GAUACUAAAGAGUCAAAUAACACAAAU---------------------GGAACUAUCACACUCCCAUG
B-JRCSF      AUAGUACUUGGAAU------------------GAUACUGAAAAGUCAAGUGGCACUGAAGGA---------------AAUGACACCAUCAUACUCCCAUG
B-LAI        AUAGUACUUGGUUUAAUAGUACU------UGGAGUACUGAAGGGUCAAAUAACACUGAAGGA---------------AGUGACACAAUCACACUCCCAUG
B-MANC       AUAGUACUUGGAAUACU---GGG------AAUGAUACUAGAGAGUCAAAUGACACAAAUAAU---------------ACUGGAAAUAUCACACUCCCAUG
B-OYI        AUAGUACUUGGAAU------------------GAUACUACAAGGGCAAAUAGCACUGAAGUA---------------------ACUAUCACACUCCCAUG
B-YU2        -----ACUUGGAAU------------------GAUACUAGAAAGUUAAAUAACACUGGAAGA---------------------AAUAUCACACUCCCAUG
D-NDK        AUAGUACAUGGAAUCAGACU---------AAUAGUACAGGGUUCAAUAAUGGCACA------------------------------GUCACACUCCCAUG
O-ANT70      AUUAUACCUUUUCA------UGUAACGGAACCACCUGUAGUGUUAGUAAUGUUAGUCAA------------------GGUAACAAUGGCACUCUACCUUG
SIVCPZGAB    -----------------------------------------ACUGACAACAUUACAAAUGGC---------------------AUUAUAAUACUGCCAUG
```

Figure 1. Part of two alignments of HIV sequences produced with `ClustalW` (top) and `ralign` (bottom). Note the much smaller number of gaps in the second case.

which open reading frame is used for the protein alignment. In such cases `ralign` prefers longer open reading frames over shorter ones, but also allows the user to modify the lists of ORFs that will be used for amino acid level alignments.

Starting from the ORF with the highest priority, the protein level alignments are computed using `Clustal W` [17]. End gaps in these alignments cause problems when the individual blocks are recombined at the end. Hence `ralign` removes the frail ends from the alignments and retains only well-aligned "centers" of the ORFs. The end-pieces are either attached to a coding region with lower priority or handled together with the non-coding regions at the level of nucleic acids. In the final step the aligned protein sequences are replaced by their underlying nucleic acid sequence.

`ralign` significantly improves the quality of the alignments of viral nucleic acid sequences, in particular in regions with large sequence heterogeneity. In particular, it tends to reduce the number of gaps. An example alignment can be seen in figure 1.

3. Searching Conserved Secondary Structure Elements

The algorithms `alidot` and `pfrali` for searching conserved secondary structure patterns in large RNAs are described in detail elsewhere [6, 7]. An ANSI C implementation is available from `http://www.tbi.univie.ac.at/`. Hence we restrict ourselves to a short overview.

We start from a multiple sequence alignment that is obtained without any reference to the predicted structures, and an independent structure predictions for each of the sequences. The sequence alignment determines which base pairs from the different sequences are equivalent. These data are reorganized into a list of all predicted base pairs that contains for each pair information about how often (and in the case of `pfrali` with which pairing probability) the pair was predicted, and about the sequence variation associated with each pair. This list is then rank-ordered by a hierarchical procedure such that the highest ranking pairs can be formed by all sequences in the data set and show a large number of different types of base pairs, i.e., many compensatory mutations. Then the pairs are collected into stems that inherit their ranking from the base pairs. A postprocessing step removes structural elements with low credibility from the list. The final output of the procedure is the graphic representation of candidate structural elements, see Figure 2.

This method differs from related approaches such as `Construct` [10] in that it does not *a priori* assume that there is a conserved structure. For a set of unrelated sequences `alidot` and `pfrali` return an empty prediction. We did not find it necessary to improve the alignments based on visual inspection. A modification of the alignment taking into account already predicted structures might increase the number of compensatory mutations and possibly also the number of detected structural elements. However, it would compromise the use of the sequence data for verifying the predicted structures. Hence we use the alignments produce by `ralign` "as is".

4. Application: Hepatitis B Virus RNAs

The Hepatitis B virus infects mammals and belongs to the family of the Hepadnaviridae. The genome consists of a single molecule of non-covalently closed, circular

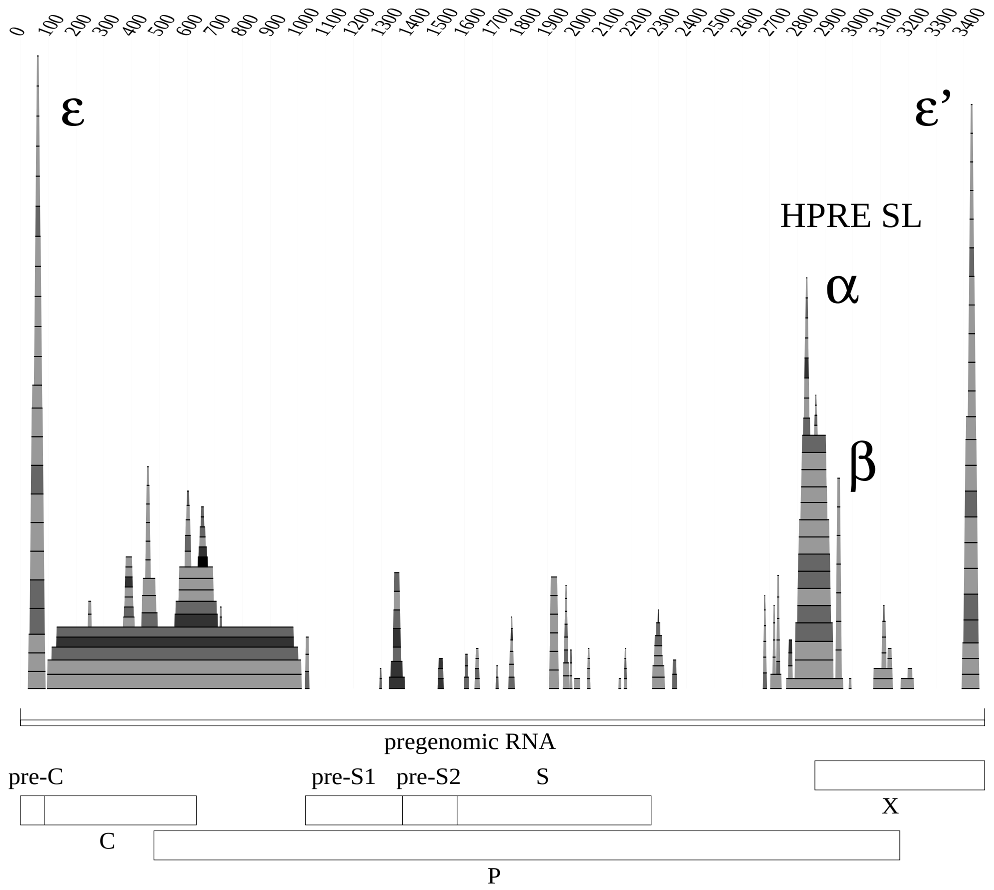

Figure 2. The four most prominent conserved elements are the ε-element and its copy ε' at the 5' and 3' end of the pregenome, and the two stem loop structures α and $\beta 1$ in the HPRE region. Colors (here replaced by grayscale from light to dark) indicate compensatory mutations from 0 to 3 different types of base pairs).

DNA that is partially double-stranded and partially single-stranded. One strand in negative sense (complementary to the viral mRNAs) is full-length (3100-3400 nt), the other is shorter and varies in size. The hepatitis B virus genome has four partially overlapping genes (S, C, P, X), all oriented in the same direction. The S-gene codes for the surface antigens. The P-gene covers 80% of the entire genome and overlaps the other three ORFs. It codes for the reverse transcriptase, with DNA polymerase and RNase H activities, and the genome-linked terminal protein. The C gene codes of the major core protein. Finally, the X-gene specifies a protein with a probable transactivation function. It varies in size among the HBV serotypes.

After insertion of the double-stranded DNA into the nucleus of infected cells, transcription yields various species of mRNAs as well as the 3.4 kb pre-genome which is longer than the genome, because of terminally redundant sequence parts. The mRNAs are unspliced and are made from distinct promotors. Two regions of the

HBV genome have transcription enhancer activities, another is similar to glucocorticoid responsive elements.

We have analysed 21 virus strains from humans and rodents. An overview of the predicted structures is shown in the mountain plot [8], figure 2. The most pronounced conserved structure motifs occur near the 3' and 5' ends of the RNA pre-genome, as well as the start of the X gene.

The ε-structure is a 5'-proximal encapsidation signal which mediates the specific packaging of the transcript into viral capsids after reverse transcription of the terminally redundant RNA pregenome by interaction with the reverse transcriptase (P-protein) [9, 12]. Because of the redundancy of the RNA pre-genome a second copy of the ε-structure occurs at the 3'-end.

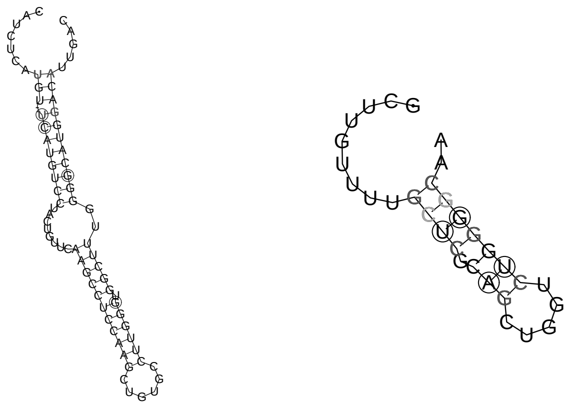

Figure 3. Predicted secondary structure of ε, α, and β1 element. Circles indicate compensatory mutations.

The post-transcriptional regulatory element (HPRE) facilitates the cytoplasmic localization of intronless transcripts and is composed of at least two independent sub-elements (α and β) which are necessary for full HPRE function and conserved throughout the mammalian Hepadnaviruses. Two regions of these sub-elements (SLα and SLβ1) form stem-loop structures [15], Figure 3.

Neither `alidot` nor `pfrali` has unambiguously detected novel conserved secondary structure elements in the pregenomic RNA. We have therefore also investigated all known sub-genomic RNAs of HBV. In the C-transcript we found one stem-loop structure, Figure 4 that does not appear to be conserved in the pregenomic RNA because interactions with a region outside the C-gene are thermodynamically favorable. The fact that we find a conserved structure in C and distinctly different folding in other context makes this feature a candidate for a regulatory function.

Avihepadnavirus has a similar replication strategy and similar genomic organization (they lack the X-gene) compared with mammalian HBV. In fact, the two genera are grouped together in the family Hepadnaviridae. We have not been able to produce useful alignments of sequences from the two genera, however. The available avihepadnavirus sequences are too similar for `alidot` or `pfrali`.

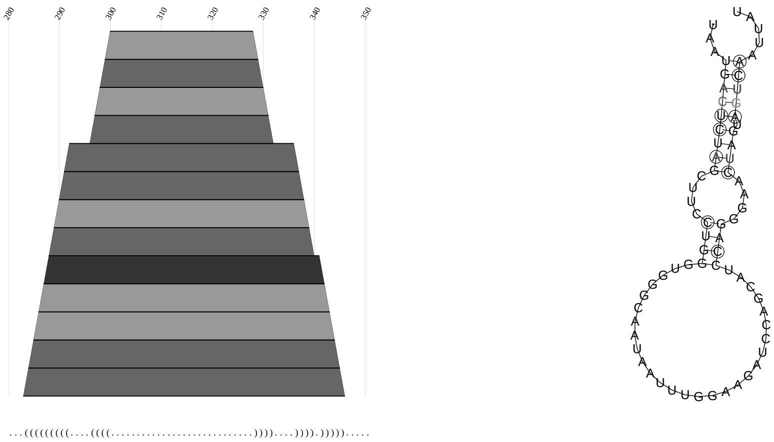

Figure 4. A secondary structure element (as mountain plot and in traditional reprensentation) with unknown function that appears to be conserved in the C-mRNA but does not appear as conserved structure in complete pregenomic RNA.

An ε-like stem-loop element, analogous to the ε-element in orthohepadnavirus has been postulated in the literature [1]. While the ε-element is well supported by sequence variation in mammalian sequences, the sequence variations among isolates from duck and heron shows that there is no single conserved conformation in avihepadnavirus. If this region is indeed vital, then different structures function in different isolates.

Acknowledgements. A part of the analysis of Hepatitis B Virus RNA reported here was performed during the *Bioinformatik I* computer lab course in the fall semester 1998/99 by Ingrid Abfalter, Phillip Kobel, and Clemens Uanschou. Thanks a lot for your help! This work was supported in part by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung* Proj. P13545-MAT.

References

[1] J. Beck, H. Bartos, and M. Nassal. Experimental confirmation of a hepatitis B virus (HBV) ε-like bulge-and-loop structure in avian HBV RNA encapsidation signals. *Virology*, 227:500–504, 1997.

[2] J. Corodkin, L. J. Heyer, and G. D. Stormo. Finding common sequences and structure motifs in a set of RNA molecules. In T. Gaasterland, P. Karp, K. Karplus, C. Ouzounis, C. Sander, and A. Valencia, editors, *Proceedings of the ISMB-97*, pages 120–123, Menlo Park, CA, 1997. AAAI Press.

[3] T. Dandekar and M. W. Hentze. Finding the hairpin in the haystack: searching for RNA motifs. *Trends. Genet.*, 11:45–50, 1995.

[4] W. Fontana, D. A. M. Konings, P. F. Stadler, and P. Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33:1389–1404, 1993.

[5] D. Gautheret, F. Major, and R. Cedergren. Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA. *Comput. Appl. Biosci.*, 6:325–331, 1990.

[6] I. L. Hofacker, M. Fekete, C. Flamm, M. A. Huynen, S. Rauscher, P. E. Stolorz, and P. F. Stadler. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucl. Acids Res.*, 26:3825–3836, 1998.

[7] I. L. Hofacker and P. F. Stadler. Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comp. & Chem.*, 23:401–414, 1999. Santa Fe Institute preprint 98-06-058.

[8] P. Hogeweg and B. Hesper. Energy directed folding of RNA sequences. *Nucl. Acids Res.*, 12:67–74, 1984.

[9] A. H. Kidd and K. Kidd-Ljunggren. A revised secondary structure model for the 3'-end of hepatitis B virus pregenomic RNA. *Nucl. Acids Res.*, 24:3295–3301, 1996.

[10] R. Lück, G. Steger, and D. Riesner. Thermodynamic prediction of conserved secondary structure: Application to the RRE element of HIV, the tRNA-like element of CMV, and the mRNA of prion protein. *J. Mol. Biol.*, 258:813–826, 1996.

[11] C. W. Mandl, H. Holzmann, T. Meixner, S. Rauscher, P. F. Stadler, S. L. Allison, and F. X. Heinz. Spontaneous and engineered deletions in the 3'-noncoding region of tick-borne encephalitis virus: Construction of highly attenuated mutants of flavivirus. *J. Virology*, 72:2132–2140, 1998.

[12] A. Rieger and M. Nassal. Distinct requirements for primary sequence in the 5'-and 3'-part of a bulge in the hepatitis b virus rna encapsidation signal revealed by a combined in vivo selection/in vitro amplification system. *Nucl. Acids Res.*, 23:3909–3915, 1995.

[13] D. Sankoff. Simultaneous solution of the RNA folding, alignment, and proto-sequence problems. *SIAM J. Appl. Math.*, 45:810–825, 1985.

[14] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proc. Royal Society London B*, 255:279–284, 1994.

[15] G. J. Smith, J. E. Donello, R. Lueck, G. Steger, and T. J. Hope. The hepatitis b virus post-transcriptional regulatory element contains two conserved RNA stem-loops which are required for function. *Nucl. Acids Res.*, 26,:4818–4827, 1998.

[16] J. E. Tabaska and G. D. Stormo. Automated alignment of RNA sequences to pseudoknotted structures. In T. Gaasterland, P. Karp, K. Karplus, C. Ouzounis, C. Sander, and A. Valencia, editors, *Proceedings of the ISMB-97*, pages 311–318, Menlo Park, CA, 1997. AAAI Press.

[17] J. D. Thompson, D. G. Higgs, and T. J. Gibson. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucl. Acids Res.*, 22:4673–4680, 1994.