Virus Evolution, 2019, 5(2): vez034

doi: 10.1093/ve/vez034
Research article

Conserved structural RNA domains in regions coding for cleavage site motifs in hemagglutinin genes of influenza viruses

Alexander P. Gultyaev,[1,2,*] Mathilde Richard,[1] Monique I. Spronken,[1] René C.L. Olsthoorn,[3] and Ron A.M. Fouchier[1]

[1]Department of Viroscience, Erasmus Medical Center, Dr. Molewaterplein 40, 3015 GD Rotterdam, The Netherlands, [2]Group Imaging and Bioinformatics, Leiden Institute of Advanced Computer Science (LIACS), Leiden University, PO Box 9512, 2300 RA Leiden, The Netherlands and [3]Leiden Institute of Chemistry, Leiden University, PO Box 9502, 2300 RA Leiden, The Netherlands

*Corresponding author: E-mail: a.goultiaev@erasmusmc.nl

Abstract

The acquisition of a multibasic cleavage site (MBCS) in the hemagglutinin (HA) glycoprotein is the main determinant of the conversion of low pathogenic avian influenza viruses into highly pathogenic strains, facilitating HA cleavage and virus replication in a broader range of host cells. In nature, substitutions or insertions in HA RNA genomic segments that code for multiple basic amino acids have been observed only in the HA genes of two out of sixteen subtypes circulating in birds, H5 and H7. Given the compatibility of MBCS motifs with HA proteins of numerous subtypes, this selectivity was hypothesized to be determined by the existence of specific motifs in HA RNA, in particular structured domains. In H5 and H7 HA RNAs, predictions of such domains have yielded alternative conserved stem-loop structures with the cleavage site codons in the hairpin loops. Here, potential RNA secondary structures were analyzed in the cleavage site regions of HA segments of influenza viruses of different types and subtypes. H5- and H7-like stem-loop structures were found in all known influenza A virus subtypes and in influenza B and C viruses with homology modeling. Nucleotide covariations supported this conservation to be determined by RNA structural constraints that are stronger in the domain-closing bottom stems as compared to apical parts. The structured character of this region in (sub-)types other than H5 and H7 indicates its functional importance beyond the ability to evolve toward an MBCS responsible for a highly pathogenic phenotype.

Key words: influenza virus; RNA structure; highly pathogenic avian influenza

1. Introduction

Wild aquatic birds are an important reservoir of influenza A viruses, from where they may spread to other species and cause influenza outbreaks in humans and domestic animals (Horimoto and Kawaoka 2005; Alexander 2007; Lupiani and Reddy 2009). Low pathogenic avian influenza (LPAI) viruses are omnipresent in wild bird populations without causing disease. Occasionally, LPAI strains evolve into highly pathogenic avian influenza (HPAI) viruses in poultry that can lead to devastating outbreaks with occasional spillovers to wild birds. Upon introduction to humans, HPAI viruses may cause high morbidity and mortality (Alexander 2007; Lupiani and Reddy 2009). Influenza A viruses have a negative-sense RNA genome consisting of eight single-stranded segments, and are classified into subtypes defined by the antigenic properties of the surface glycoproteins

hemagglutinin (HA) and neuraminidase (NA). Currently, sixteen HA subtypes (H1–H16) and nine NA subtypes (N1–N9) are known to circulate in birds, in many different H–N combinations.

The main determinant of high pathogenicity of HPAI viruses is the change of the motif in the virus hemagglutinin polyprotein HA0 amino acid sequence that determines HA0 cleavage into two chains, HA1 and HA2, to yield functional HA (Klenk and Garten 1994; Horimoto and Kawaoka 2005). In LPAI strains, HA is cleaved by trypsin-like proteases, present in a limited range of host tissues. Addition of several basic amino acids to the cleavage site (CS) motif in HPAI viruses creates a so-called multibasic cleavage site (MBCS) that can also be cleaved by furin-like proteases available in a broader range of host tissues, facilitating systemic virus spread that leads to more severe disease in poultry. Such MBCS motifs can evolve by insertions of basic amino acid codons or by substitutions in the HA RNA segment. Remarkably, in nature this has been observed only in HA genes of the H5 and H7 subtypes (Alexander 2007; Lupiani and Reddy 2009; Abdelwhab, Veits, and Mettenleiter 2013). This restriction does not seem to be determined by specific properties of H5 and H7 HA proteins, because highly pathogenic viruses of other subtypes can be produced in the laboratory by MBCS introduction using reverse genetics (Munster et al. 2010; Veits et al. 2012). This has led to the hypothesis that certain features in H5 and H7 HA RNA segments, rather than the encoded proteins, specify their ability to acquire MBCS sequences.

In particular, RNA folding in the HA RNA region coding for the cleavage site motif could be an important factor, and different structural models have been suggested (García et al. 1996; Perdue et al. 1997; Gultyaev et al. 2016; Abolnik 2017; Nao et al. 2017; Dietze et al. 2018). The most conserved structures in this region are the stem-loop domains that encompass the codons encoding the cleavage site motif (Gultyaev et al. 2016). In both H5 and H7 genes, the sites of insertions of additional basic codons are located in the loops of predicted structures, which is suggestive for their involvement in the development of an HPAI phenotype. However, the exact structural features important for this process remain undetermined. Comparative RNA structure analysis indicated the formation of different stem-loop (SL) topologies in H5 and H7 HAs (Fig. 1). On the other hand, homologous structures could be folded in HA segments of some other influenza A virus subtypes, not known to evolve toward HPAI strains (Gultyaev et al. 2016). Identification of conserved structural motifs at the HA1/HA2 boundary is important for understanding the mechanisms of HPAI emergence and potential use for outbreak warnings and as drug targets.

Here, structural comparisons were used to reveal potential H5-like and H7-like RNA structures in HA cleavage site regions of all known types and subtypes of influenza viruses with available HA sequences. The patterns of structure similarities and differences in H5 and H7 HA RNAs as compared to other subtypes were analyzed to shed light on possible contributions of identified RNA motifs to the emergence of HPAI strains and to virus fitness. In addition to 16 HA subtypes of influenza A viruses, the structures in homologous RNA sequences of other influenza viruses were modeled, including subtypes H17 and H18 of influenza A-like bat viruses (Tong et al. 2013; Ma, García-Sastre, and Schwemmle 2015), human influenza B and C viruses (Chen and Holmes 2008; Matsuzaki et al. 2016), bovine influenza D viruses (Mekata et al. 2018), and influenza viruses of eel and toad (Shi et al. 2018). All influenza virus (sub)types but influenza D and Wuhan Asiatic toad influenza viruses included a number of strains with potential H5-like or H7-like HA structured domains. The structural conservation of RNA stem-loop structures at the HA cleavage site region is indicative for functional importance beyond the ability to evolve toward an MBCS.

2. Materials and methods

2.1 Locations of H5- and H7-like structured domains

The locations of conserved domains in the cleavage site regions of H5 and H7 HA RNAs were defined by the previous comparative analysis (Gultyaev et al. 2016) (Fig. 1). The SLH5a and SLH5b structures were identified in the regions homologous to nucleotides at positions 1008–1082 and 1022–1076 respectively, in full-length H5 HA segment of an LPAI virus, e.g. A/mallard/Netherlands/3/99(H5N2), accession AY684894. Slightly shortened SLH5a domains with closing base pairs homologous to the 1013–1077 pair in the H5 structure were considered as well. The SLH7a and SLH7b domains corresponded to nucleotides 1017–1062 and 1013–1054, respectively, of full-length H7 HA, e.g. A/mallard duck/ALB/224/1977(H7N5), accession CY005973. Homologous regions in HA RNAs of influenza viruses of other subtypes were identified using alignments of encoded amino acid sequences.

2.2 RNA structure predictions

RNA structures in HA RNA regions homologous to conserved structured domains in H5 and H7 HA RNAs were predicted by Mfold program, which calculates the lowest free energy structures and suboptimal ones (Zuker 2003). For the estimates of thermodynamic stabilities of predicted structures, two folding free energy values were considered to be informative: the folding free energy of a specific structure and its difference as compared to the lowest free energy possible in the analyzed sequence region. The structures with free energies higher than −5.0 kcal/mol or with a difference more than 5 kcal/mol as compared to the optimal free energy were considered to be unlikely. Predictions were carried out for both positive-sense RNA sequences and their negative-sense counterparts.

In order to identify potential H5-like and H7-like structures in HA RNAs of other subtypes, sequences from representative strains of various lineages were used for structure predictions. The selection of representative strains was carried out maximizing the coverage of sequence diversity in each of the subtypes, assisted by published HA segment phylogenies (e.g. Chen and Holmes 2008; Dugan et al. 2008; Liu et al. 2009; Lebarbenchon and Stallknecht 2011; Dalby and Iqbal 2014; Matsuzaki et al. 2016) and alignments of HA sequences downloaded from the NCBI Influenza Virus Resource (Bao et al. 2008). The alignments of HA sequences of specified subtypes and geographic origins, carried out by the Clustal Omega program (Li et al. 2015), were also used to evaluate the conservation of structures in virus clades.

Significance of covariations in suggested base pairs was estimated as described previously, using mutual information values and the number of covariation events (Gultyaev et al. 2016). For a putative covariation of nucleotides at positions x and y, the mutual information $M(xy)$ and the normalized values with account of positional entropies $H(x)$ and $H(y)$ at two positions $R_1 = M(xy)/H(x)$ and $R_2 = M(xy)/H(y)$ were calculated. These values are in the range from 0 to 1, with 0 indicating independent variations and 1 corresponding to the perfect correlation. It should be noted, however, that calculations with permuted HA sequences have shown that a large value of (normalized) mutual information alone is not

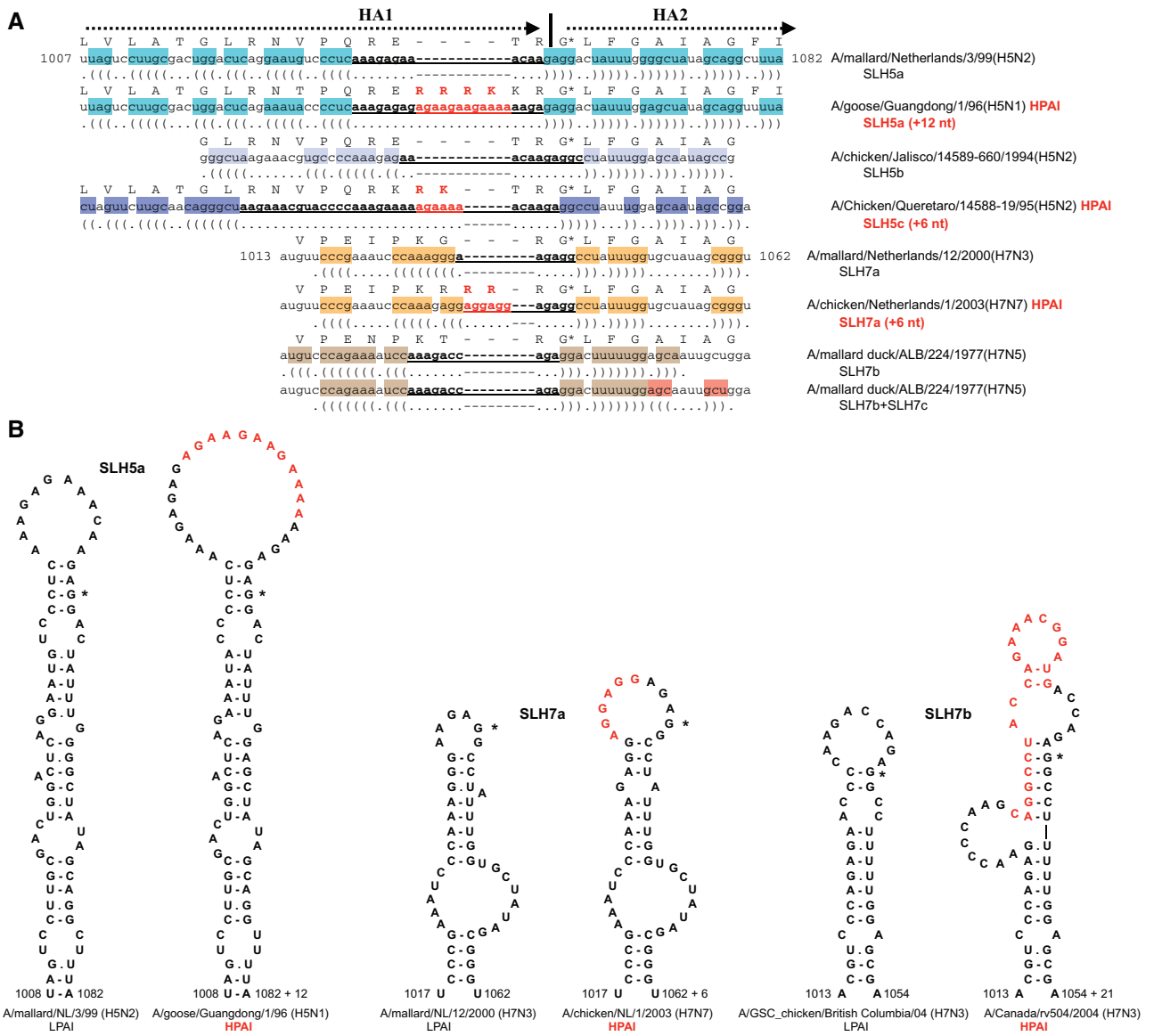


Figure 1. RNA secondary structures encompassing the cleavage site regions of H5 and H7 HA segments, derived from comparative RNA structure predictions (Gultyaev et al. 2016). (A) Alignment of structures based on encoded amino acid sequences at the boundary between HA1 and HA2 regions. The structures are shown using a bracket view with each base pair denoted with a bracket pair. Paired nucleotides are shown with background colors specific for alternative structures. The hairpin loop nucleotides are in bold type and underlined. The stem-loop (SL) structure denoted here as SLH5a is conserved in majority of H5 lineages, while the SLH5b and SLH5c are folded in the lineage of H5N2 viruses first isolated during an outbreak in Mexico. The SLH7a structures are possible in all Eurasian and American H7 strains, the alternative SLH7b folds are conserved only in H7 HA segments of American origin. A small hairpin SLH7c is conserved in H7, H10 and H15 HA RNAs (Gultyaev et al. 2016). Insertions of additional basic amino acid codons in the HPAI strains are shown in red. The N-terminal HA2 glycine amino acid residue is denoted by an asterisk. (B) Examples of conserved structures in LPAI and HPAI viruses. Adapted from Gultyaev et al. (2016), available under the Creative Commons Attribution 4.0 International License

always determined by base-pairing (Gultyaev et al. 2016), what is consistent with tests on other RNAs (Rivas, Clements, and Eddy 2017). More reliable support for predicted HA RNA structures was obtained from detection of more than one covariation in a structure or independent covariation events at the same base pair. In such cases, even two covariation events with relatively small correlation scores could be considered as a support for a structural model (Gultyaev et al. 2016).

For the segment-wide evaluation of structure conservation and stability the RNAz server (Gruber et al. 2007) was used. The RNAz algorithm scans an alignment of sequences with a running window and yields for each window a score which evaluates a probability of finding a conserved structured domain. The score is based on structural conservation and thermodynamic stability of predicted local structures. In particular, RNAz calculates the significance of free energy deviation from expected values in random sequences of the same length and base composition. The algorithm was used with default parameters, window size of 120 nucleotides, step size of 10 nucleotides and minimum value of total *P*-score of 0.85.

3. Results

3.1 Predictions of RNA structures that encompass cleavage site codons in HA segments

Using RNA folding free energy minimization in homologous HA segment regions, secondary structures similar to those folded

in H5 or H7 HA RNAs (Fig. 1) were identified in a number of other influenza viruses (Fig. 2). Strains that formed SLH5a-, SLH7a-, or SLH7b-like domains, as defined by homologous bottom domain-closing base pairs, were found not only among all influenza A virus subtypes, but also in more distantly related influenza B and C viruses. The base pairs in domain-closing stems shared more similarities than the apical parts, which contained various rod-like stem-loop structures shifted in relation to each other along the alignment or branched Y-shaped conformations. Representative structures are shown in Supplementary Figs S1–S4 and Figs 3–5. Stable homologs of SLH5b and SLH5c structures, suggested for HA segments of the Mexican H5N2 lineage (García et al. 1996; Perdue et al. 1997; Nao et al. 2017), were not found in other HA RNAs.

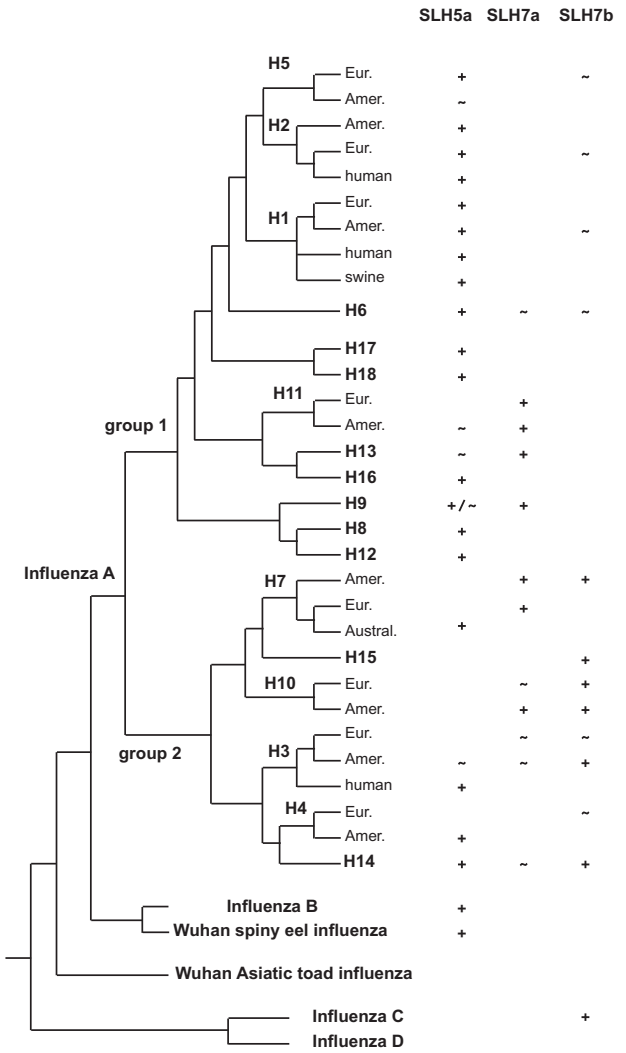


Figure 2. Identification of strains with possible H5- and H7-like domains in different clades of influenza viruses. The '–'and '+' signs define predictions of stable structures in at least some or all strains of a given clade, respectively. The prediction of SLH5a-like domains in H9 HA subtype is shown as '+/–' because this structure is not predicted only in a few outlier sequences whereas it is conserved in the overwhelming majority of H9 HA segments. Representative structures are shown in Figs 3–5 and Supplementary Figs S1–S4. The tree topology is according to phylogenetic reconstructions by Russell et al. (2004), Fouchier et al. (2005), Suzuki and Nei (2002), Tong et al. (2013), and Shi et al. (2018). The branches corresponding to geographic sublineages of avian influenza A viruses (Eur., Eurasia; Amer., America), human and swine strains are shown only in cases of different extent of conservation of structures.

Although H5-like or H7-like domains were not predicted in all strains, they were conserved in numerous lineages of influenza virus beyond the H5 and H7 subtypes of influenza A (Fig. 2). In some HA RNAs, an equilibrium between alternative structures was predicted. For instance, all H9 HA segments could form an SLH7a stem-loop structure, and most of these a stable rod-like and/or Y-shaped SLH5a domain as well (Fig. 3, Supplementary Figs S1 and S2). In the majority of HA RNA sequences, mirror configurations were predicted for the positive- and negative-sense RNAs, with only minor deviations in base-pairing due to asymmetry of stacking free energies, in particular, in case of wobble GU pairs corresponding to AC mismatches in the complement. However, some exceptions were noted. In the influenza B virus HA vRNAs, the mirror folding of the conserved positive-sense SLH5a homolog was thermodynamically less favorable than a stem-loop structure with shifted stems (Supplementary Fig. S3).

Nucleotide covariations indicated that this structure conservation, even though limited to certain groups of strains, is determined by a selective pressure to preserve RNA folding rather than by fortuitous correspondence of structure models to similar RNA sequences. Several covariations, determined by silent substitutions, occurred repeatedly in different clades. In addition to the covariation in the SLH5a-like structures reported in human H3N2 strains at positions 1040–1073 in H3 numbering (Gultyaev et al. 2016), an independent covariation in the same base pair was also detected in one of the human-origin lineages of swine H3N2 viruses (Supplementary Fig. S1). The 1008–1068 covariation in H9 HA SLH5a structures (Fig. 3) occurred at positions homologous to those involved in at least two covariation events in the monophyletic group consisting of H5, H6, H2, and H1 subtypes: once in a branch of H6 HA segments (positions 1010–1070), and a second time between H5 and H2 subtypes 1015–1075 and 1024–1084, respectively (Gultyaev et al. 2016) (Supplementary Fig. S1). Similar to these covariations, the covariation in H9 HA sublineages involved two purine/pyrimidine transversions: the GU pair, typical in the worldwide-circulating and early Chinese clades, was converted through intermediate UU mismatch into the predominant UA base pair in later Chinese clades (Fig. 3). Computed separately for H9 HA sequences, the mutual information values of this covariation are $M(xy) = 0.36$; $R_1(xy) = 0.53$ and $R_2(xy) = 0.67$, what in combination with the previously obtained similar values $M(xy) = 0.32$; $R_1(xy) = 0.67$ and $R_2(xy) = 0.61$ in the H5/H6/H2/H1 cluster (Gultyaev et al. 2016) indicates a significant structural constraint. In influenza B viruses, the homologous base pair at positions 1068–1137 was mostly GU (Fig. 4, Supplementary Fig. S3). Another covariation in the influenza B virus SLH5a-like structure at positions 1062–1143 (Fig. 4) was similar to the previously reported 1009–1081 AU->GC covariation between Eurasian and American H5 HA segments (Gultyaev et al. 2016). A unique covariation resulting in two amino acid changes occurred in the majority of H8 HA sequences where the predominant SLH5a hairpin loop-closing CG pair was converted into UA (Supplementary Fig. S1). Within the cluster of H8 and closely related H12 HA segments (Fig. 2), the mutual information values were $M(xy) = 0.37$; $R_1(xy) = 0.78$ and $R_2(xy) = 0.84$.

In the SLH7a structure, covariations of the 1020/1059 bp occurred at least three times independently: in avian H7 HA sequences (Gultyaev et al. 2016), in equine H7N7 viruses, and in HA segments from the H10N7 outbreak in seals in Europe (Supplementary Fig. S2). The mutual information values for this covariation in H7 HA sequences are $M(xy) = 0.44$; $R_1(xy) = 0.58$ and $R_2(xy) = 0.56$ (Gultyaev et al. 2016). The terminal base pair

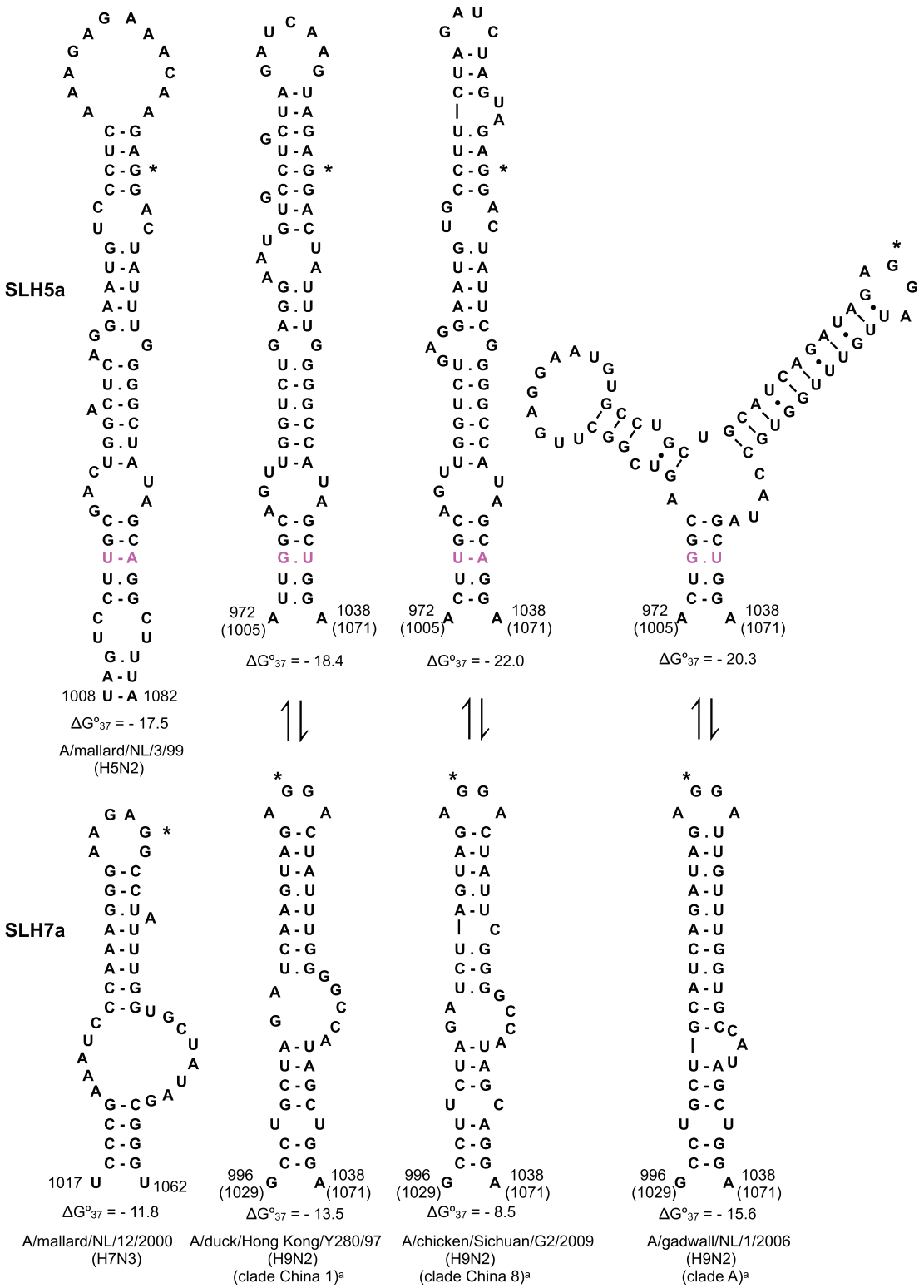
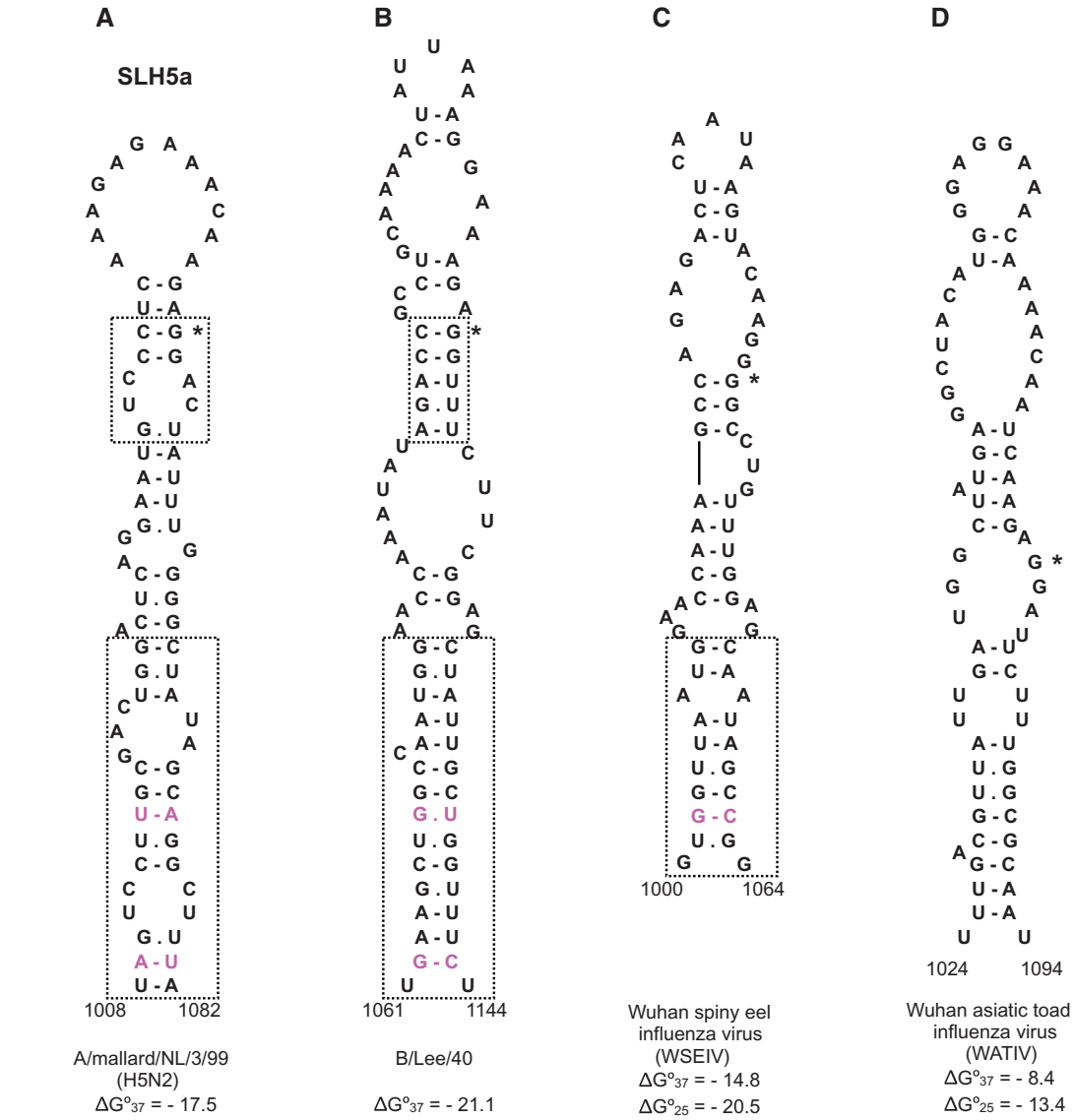


Figure 3. Alternative rod-like and Y-shaped SLH5a- and SLH7a-like structures in H9 HA RNAs. Covaried nucleotides in the SLH5a domain-closing stems are shown in magenta. The typical SLH5a and SLH7a structures (Gultyaev et al. 2016) are shown for comparison. The GGN codon coding for the N-terminal HA2 glycine amino acid residue is shown by an asterisk. Accessions: A/Duck/Hong Kong/Y280/97(H9N2), AF156376; A/chicken/Sichuan/G2/2009(H9N2), GU471800; A/gadwall/Netherlands/1/2006(H9N2), CY043864. Nucleotide positions of database sequences are given and in case of partial HA sequences the positions corresponding to a full-length segment are shown in brackets. [a]Clades of H9 HA segments are according to Dalby and Iqbal (2014).



E

```
  V  L  A  T  G  L  R  N  V  P  Q  R  E  T  -  -  -  R  G* L  F  G  A  I  A  G  F
uAguccuUgcgacuggacucaggaaugucccucaaagagaaaca---------agaggacuauuuggggcuauagcAggcuuUa A (H5)
(((..((((((...(((.((((..((((((..((((..........---------.)))))..))))))..))))))..))))))..)))
  K  L  A  N  G  T  K  Y  R  P  P  A  K  L  L  K  E  R  G* F  F  G  A  I  A  G  F
uGaagcuGgccaauggaaccaaaauauagaccgccugcaaaacuauuaaaggaaagagguuucuucggagcuauuGcUgguuuCu B/Lee/40
.(((((((((.((((((..((.....((((((..((.....((.....))...))).))))))....))..)))))))))))).
  L  V  V  N  G  T  K  A  R  D  S  I  S  T  -  -  -  R  G* L  F  G  A  I  A  G  F
uacuggUGguuaauggaaccaaagccagagacucaaauaaguaca---------agggggccuguuuuggagcaauagcCggauucc WSEIV
.....(((((((.((...(((((((((....(((.....)))).....---------...)))...)))))).))).)))))....
  F  D  V  I  D  G  Y  E  A  T  W  E  E  T  K  Q  S  R  G* F  F  G  A  I  L  G  F
 uuugacguuauuugauggcuauggaggcuacaugggaggaaacaaaacaaucaagaggauucuuuggcgcaauauuaggguuuuu WATIV
 .(((.((((((..((...((.(((........((........)).....)))))).....))..)))))))).............
```

Figure 4. SLH5a-like structures in the influenza B (B) and Wuhan spiny eel influenza (C) viruses. The typical SLH5a structure (A) (Gultyaev et al. 2016) is shown for comparison. Homologous parts of the stems are in dashed frames, covariations are shown in magenta. (D) An alternative stem-loop structure predicted in the HA RNA of Wuhan asiatic toad influenza virus. The folding free energies of structures in fish and amphibian influenza viruses (C, D) are computed at 25 °C in addition to standard values at 37 °C, in order to get approximate values consistent with the natural replication conditions. (E) Alignment of structures. Other notations are similar to Fig. 1. Accessions: B/Lee/40, K00423; Wuhan spiny eel influenza virus, MG600041; Wuhan Asiatic toad influenza virus, MG600048. Only partial sequences of the HA segments of the last two viruses are available.

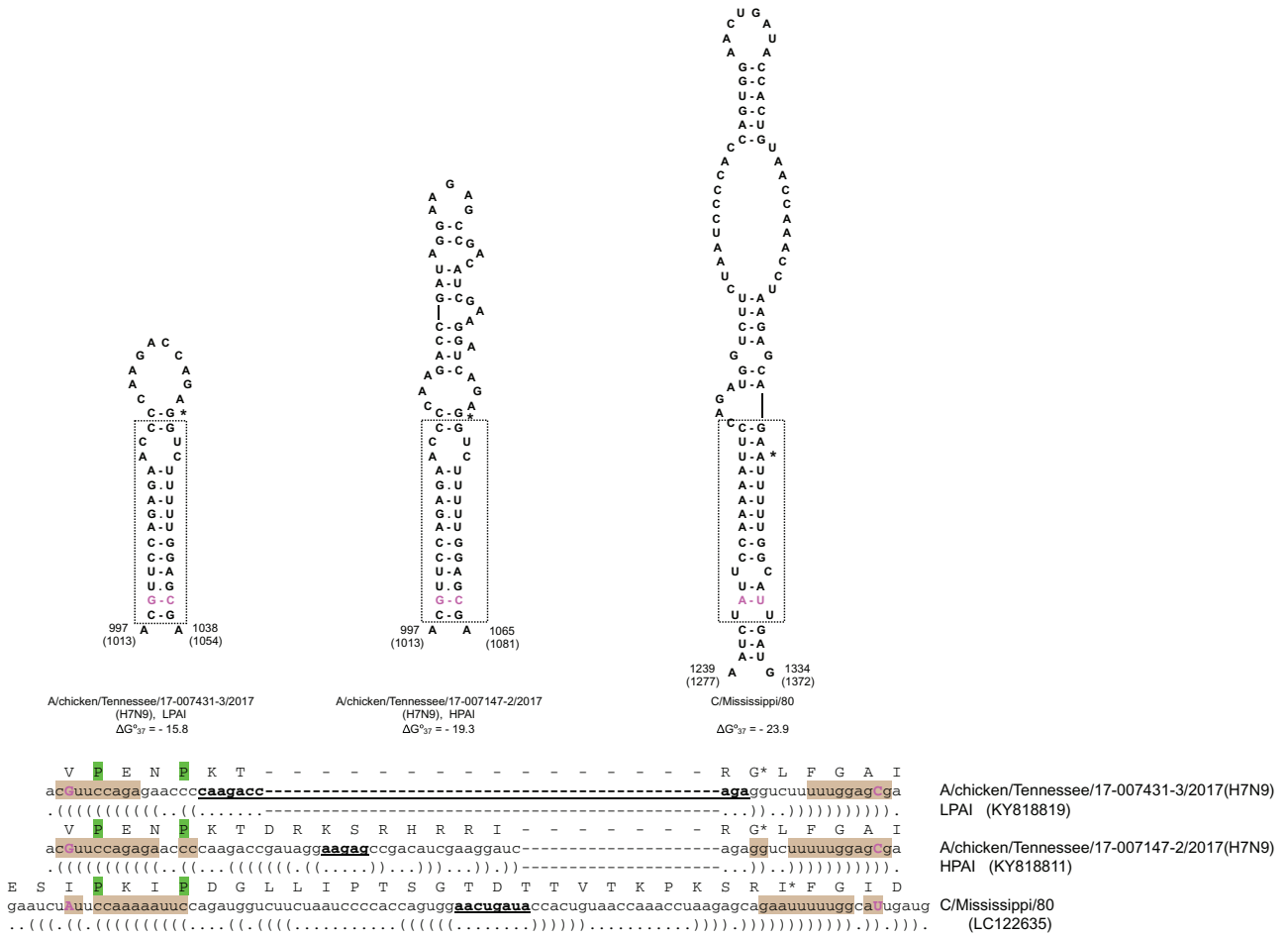


Figure 5. An example of SLH7b-like folding in the influenza C viruses, strain C/Mississippi/80. The SLH7b structures predicted in the HPAI strain A/chicken/Tennessee/17-007147-2/2017(H7N9) and its LPAI progenitor A/chicken/Tennessee/17-007431-3/2017(H7N9) (Lee et al. 2017b) are shown for comparison. Proline codons used as anchors in the HEF1/HA1 alignment are labeled in green. Other notations are similar to Fig. 4.

1014–1053 of the SLH7b structure is changed from the predominant UA combination to CG in a subgroup of American avian H7 strains corresponding to the 'M' and 'N' clusters in the phylogenetic tree of Lebarbenchon and Stallknecht (2011), e.g. in A/GSC_chicken/British Columbia/04(H7N3) (Supplementary Fig. S2). Another covariation was seen in the pair 1017–1050 of the same fold: two transversions with predominant pyrimidine–purine pairs in American avian H7 viruses versus purine–pyrimidine in H15 HA and Eurasian H10 HA sequences. The normalized mutual information value of the 1014–1053 covariation in these viruses is relatively low ($R_1(xy) = 0.19$) due to the mismatch in a number of H10 HA sequences, but in H7 HA segments this value is higher, $R_1(xy) = 0.61$. The base pair 1017–1050 is more conserved, its correlation values were $R_1(xy) = 0.49$ and $R_2(xy) = 0.76$ within the H7/H15/H10 cluster. The covariations at two positions supported the SLH7b folding.

While SLH7a structures are conserved in both Eurasian and American H7 HA segments and the SLH7b in all American ones (Gultyaev et al. 2016), the only H7 HA segments with neither stable SLH7a nor SLH7b structures, isolated in Australia, turned out to have stable H5-like SLH5a stem-loops (Supplementary Fig. S1) that apparently substituted typical H7 hairpins. The basic codon insertions in Australian HP H7N7 strains occurred in the hairpin loop of the SLH5a structure (Supplementary Fig. S1), like in HPAI H5 viruses (Fig. 1). A unique acquisition of an MBCS by a natural H4N2 virus (Wong et al. 2014) was also determined by insertion in a stable SLH5a stem-loop (Supplementary Fig. S1), which is not conserved across all H4 segments.

Thermodynamically stable stem-loop structures were also predicted in the cleavage site regions of hemagglutinin-esterase-fusion (HEF) segments of influenza C viruses. Although only the HEF2 regions of encoded proteins could be reliably aligned to homologous HA2 amino acid sequences from influenza A viruses (Suzuki and Nei 2002), a resemblance between influenza C stem-loops and SLH7b structures was noted (Fig. 5). The resemblance was based on similarity of nucleotide sequences in the regions around two proline codons present in both H7 HA1 and HEF1 mRNAs. These regions paired to homologous parts in HA2 and HEF2, respectively, resulting in similar bottom parts of predicted structures in H7 HA and HEF RNAs. As compared to the LPAI H7 HA SLH7b structure, the influenza C virus RNA domain formed an elongated stem-loop due to a large insertion into the SLH7b apical part, like in the HPAI H7 strains that evolved by recombination (Fig. 5). These elongated stem-loops were stable in all sequenced influenza C virus HEF segments, and their negative-sense vRNA counterparts folded with even lower free energies (Supplementary Fig. S4).

Stable homologous stem-loop structures were not found in HEF RNA segments of influenza D viruses, which are distantly related to influenza C viruses (Mekata et al. 2018). Using the

search for conserved local structures similar to the procedure that has been previously applied for RNA segments of influenza A viruses (Gultyaev et al. 2014, 2016), it was possible to identify a different stem-loop structure conserved in the homologous region of the influenza D virus HEF mRNA (Supplementary Fig. S5). However, in the absence of base covariations in thirty-two sequences available in GenBank at the moment of writing, the reliability of this model is relatively low. No conserved structures were found in the cleavage site region of influenza D virus negative-sense HEF vRNA.

3.2 Thermodynamic stability of structures that encompass the HA cleavage site region

Thermodynamic stabilities of the predicted structures in strains of different (sub)types did not reveal any clear correlation with evolution toward the HPAI phenotype exclusively in H5 or H7 HA genes. For instance, the most stable SLH5a-like structures were predicted in the HA segments of subtype H16, with free energies lower than typical values for LPAI H5 viruses (Supplementary Fig. S1). The elongated stem-loops in HEF vRNA of influenza C viruses had even lower free energies (Supplementary Fig. S4).

None of the structures exhibited an extraordinary thermodynamic stability. For instance, in a simple estimate with free energy minimization of permuted sequences of the SLH5a region in H5 HA RNA four out of ten permutations were folded with free energies lower than that of predicted structure. This was also consistent with the results of RNAz algorithm analysis: although it suggested a structured character of this region (Supplementary Fig. S6), the reported z-score value of −2.28 that measures how many standard deviations the free energy is lower than the expected value for random sequences of the same length and base composition, is not sufficient to conclude about significantly low folding free energy. It has been shown that z-scores higher than −4 cannot be trusted (Rivas and Eddy 2000). In several subtypes, e.g. H5 and H8, the RNAz algorithm suggested that the domain surrounding the cleavage site region is one of the most structured ones in the HA segment along with the terminal packaging signal regions. This is consistent with the genome-wide analysis of RNA structure conservation in influenza A viruses of different subtypes (Moss, Priore, and Turner 2011). However, in many (sub)types the scores in the cleavage site region were not better than in other HA parts (Supplementary Fig. S6). In all cases, including the structures with relatively low free energies, like in H16 or influenza C HA RNAs, z-scores were negative, but not sufficient to indicate significantly low free energies. This is similar to patterns observed in a number of structured non-coding RNAs (Rivas and Eddy 2000).

Similar to other viruses, structural constraints in the influenza virus RNA coding regions can affect synonymous substitution rates (Tuplin et al. 2002; Gog et al. 2007; Moss, Priore, and Turner 2011). Indeed, a lower than average value of suppression of synonymous codon usage (SSCU) has been reported for the HA region 961–1080, which includes the structures studied here (Moss, Priore, and Turner 2011). We also used the data on normalized mean pairwise distance (MPD) for H1 HA codons (Gog et al. 2007) to calculate the average MPD in the predicted RNA structure around cleavage site (Supplementary Fig. S1), which corresponds to codons 329–353. The normalized MPD values show the deviation of codon usage from the expected one, and the majority of codons in the influenza genome have the scores of <1, presumably due to their role in RNA packaging in virions (Gog et al. 2007). Yet the codons within the structure turned out to have even slightly lower average MPD (0.64) than the average value calculated for all HA codons (0.67), including the main packaging signals. Of course, this difference is rather small, and therefore only a weak, if any, effect of specific structure on average codon usage was detected in the studied domain as compared to the generally lowered variation in HA codons (Gog et al. 2007). On the other hand, this does not exclude the constraints on wobble positions in codons that exhibit pairwise correlations in the detected covariations.

3.3 Potential folding of small hairpin adjacent to the HA cleavage site region

The folding potential of the small hairpin SLH7c with just three base pairs, which partially overlaps with stem-loop domains containing the cleavage site region (Fig. 1), was difficult to evaluate. Its formation can lead to shortened stem-loops and was supported by a covariation in H7 and H10 HA RNAs (Gultyaev et al. 2016). The covarying base pair was not conserved in all HA RNAs of other influenza A virus subtypes. However, even in these RNAs the SLH7c formation cannot be excluded because two other base pairs are strongly conserved in HA segments of influenza A and B viruses, being formed by two GCN codons (where N is any nucleotide) encoding conserved alanine residues (Fig. 1, Supplementary Figs S1–S3). These two base pairs are sufficient for the hairpin to have a slightly negative value of free energy in the range between −0.3 and −1.2 kcal/mol, depending on adjacent mismatched nucleotides. The third base pair, present in many strains of influenza A and B viruses, further stabilized this structure.

4. Discussion

The presented comparison of RNA structures predicted at the HA1/HA2 boundaries of HA segments from different influenza virus lineages showed the presence of structured domains encompassing the cleavage site region. The conserved base pairs in the domain-closing stems defined similar locations of stable H5- and H7-like domains in HA segments from a number of other clades. Thermodynamic stabilities of these structures in H5 and H7 HA segments were not unique either. However, the apical parts exhibited considerable structural diversity in both inter- and intra-subtype comparisons.

The hairpins corresponding to the apical parts of the models for SLH5b and SLH7a (Fig. 1) have been suggested to facilitate MBCS insertions in a strain of the Mexican H5N2 lineage (Nao et al. 2017) and German H7N7 HPAI strains of 2015 (Dietze et al. 2018), respectively, due to polymerase stuttering on homopolymer-like stretches of nucleotides in the hairpin loops. However, if the stem-loop RNA structures containing the insertion sites in their loops were of importance for insertions by polymerase stuttering and/or strand slippage, the structures should be formed by template RNAs with their apical parts within polymerase elongation complexes. Modeling of the template entrance and exit channels in the influenza virus polymerase suggests a template path of about twenty nucleotides through the polymerase (Gerlach et al. 2015; te Velthuis and Fodor 2016). Thus the loop-closing stems in the suggested H5 and H7 hairpins (Fig. 1) are unlikely to be accommodated when the loop sequence is inside the polymerase active center.

On the other hand, the template entrance and exit are close to each other on the influenza A virus polymerase surface, minimizing vRNP structure permutation during template copying

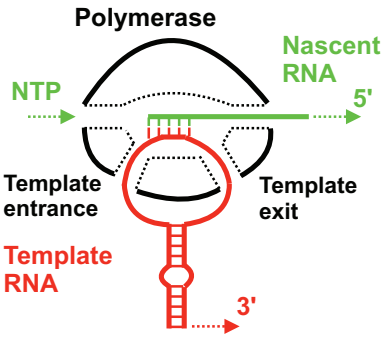


Figure 6. A model of influenza virus polymerase slowdown induced by domain-closing stems in template RNA (vRNA or cRNA). The schematic configuration of entry and exit channels of template RNA, NTP entry, and nascent RNA exit channels is according to modeling by te Velthuis and Fodor (2016).

(te Velthuis and Fodor 2016). Therefore, the RNA stem-loop extensions outside the polymerase might still interact to form a domain with a larger hairpin loop threaded through the polymerase. An important role of base-pairing in these domain-closing parts is suggested by the patterns of conservation of both H5- and H7-like structures. It is tempting to speculate that the conserved stems could slow down the polymerase locked between the paired regions, thereby increasing the probability of slippage (Fig. 6). However, such a configuration alone is not sufficient to explain why MBCS insertions occur only in H5 an H7 HA segments while homologous stems of comparable stability can form in segments of other subtypes. The differences may be determined by yet unidentified subtype-specific features of the apical parts of the domains. It should be noted that addition of non-templated nucleotides within a large RNA hairpin loop has also been suggested to occur in paramyxovirus SV5 (Thomas, Lamb, and Paterson 1988).

Insertions by RNA recombination, documented in a number of American HP H7 strains (reviewed by Abdelwhab, Veits, and Mettenleiter 2013), seem to correlate with the presence of the SLH7b structure, which was not conserved in H7 HA segments of Eurasian origin (Fig. 2). The location of crossover points within the loop of a hairpin is suggestive for a hairpin-induced template switching like in recombination of retroviral genomes (Moumen et al. 2003) and discontinuous RNA transcription in arteriviruses (van den Born et al. 2005). However, it was not possible to assign features that distinguished SLH7b from other predicted structures in this respect, and stable SLH7b structures were also folded in HA segments of other subtypes (Figs 2 and 3, Supplementary Fig. S2).

The slippage of the influenza virus polymerase during mRNA poly(A) tail synthesis by reiterative U-stretch copying is dependent on a vRNA 'hook' hairpin located seven nucleotides upstream of the U-stretch (Pritlove et al. 1999; Pflug et al. 2017). The size of the small conserved SLH7c hairpin, folded upstream of the HA1/HA2 boundary in HA vRNA (Fig. 1) resembles the hook hairpin, but its distance to the insertion site in the HPAI HA genes is twice as long, making its role in polymerase stuttering less likely. Hairpins that stimulate the slippage of reverse transcriptase on slippage-prone sequences by a roadblock-like effect in RNA templates are also located closer to indel sites (Penno et al. 2017). Here, conserved hairpins in H5 or H7 HA RNAs at a comparable position near the insertion sites were not identified.

Structural constraints in domains encompassing cleavage site regions in HA segments of different subtypes suggest that these structures could have some function(s) in virus replication cycle. For instance, vRNA structural motifs that are not disrupted by nucleoprotein binding in vRNP complexes could determine intersegmental interactions responsible for accurate packaging of all segments in the virions (Noda et al. 2012; Gavazzi et al. 2013a,b; Ferhadian et al. 2018; Shafiuddin and Boon 2019). According to the data on non-uniform NP binding to vRNA segments (Lee et al. 2017a; Williams et al. 2018), the HA vRNA region corresponding to the structures suggested in this work is characterized by relatively low NP binding. A functional pseudoknot structure in the NP segment (Gultyaev et al. 2014) has also been shown to be NP-free (Williams et al. 2018). Interestingly, the region around HA1/HA2 boundary is the only structural region indicated in HA segments of influenza A viruses with statistics of local folding free energies across different subtypes (Moss, Priore, and Turner 2011). Probably, conserved stable structures in the HA cleavage site regions prevent NP binding, while their subtype-specific apical parts are exposed for specific interactions like those regulating gene reassortment (Essere et al. 2013; Gavazzi et al. 2013a,b; Gerber et al. 2014; White, Steel, and Lowen 2017). In general, virus RNA folding may be involved in diverse functions such as maintenance of RNA stability, interference with host innate immune response (McFadden et al. 2013; Witteveldt et al. 2014), modulation of co-translational folding of encoded proteins (Mortimer, Kidwell, and Doudna 2014), or compaction of the virus RNA genome (Yoffe et al. 2008). It should be noted that our structure predictions do not give a clue whether positive- or negative-sense structures, or both, are important for the function(s) of the predicted structures.

The observed conformational diversity of structured domains at the HA1/HA2 boundary could be facilitated by a non-vital character of their function. Disruption of the SLH5a structure in mutant viruses generated by reverse genetics did not reveal any effect on virus replication in cell culture (Gultyaev et al. 2016). Presumably, beneficial properties of the structures lead to their persistence during evolution, but viruses with a destabilized domain could still replicate with opportunities to evolve an alternative topology that could serve the same purpose. On the other hand, similar to other viruses (Knoepfel and Berkhout 2013; McFadden et al. 2013; Kutchko et al. 2018), there is no direct correlation between conservation of influenza virus RNA structures and their functional importance. The structures demonstrated to be functional in packaging signals of NP, M, and PB2 segments of influenza A viruses (Gultyaev et al. 2014; Kobayashi et al. 2016, 2018; Spronken et al. 2017; Williams et al. 2018; Takizawa et al. 2019) are conserved in all strains indeed, but defects in vRNA packaging can be also observed upon disruption of the interaction between hairpins in PB1 and NS segments that is only partially conserved in avian H5N2 viruses (Gavazzi et al. 2013b). In HA packaging signals, disruption of a hairpin conserved in segments of several subtypes did not change the virus phenotype (Gultyaev et al. 2016), while such effects were observed for another hairpin present only in swine H1N1 viruses (Canale et al. 2018).

Functional properties of structures at the HA1/HA2 boundary may modulate both evolution of HA central regions in diverse influenza viruses and subtype-specific emergence toward an HPAI phenotype. Structural features of this region may determine non-random character of repeated insertions at the same location of H5 and H7 HA segments. The survey of potential structures, presented here, could help future experimental studies that should identify the structure–function relationships in these domains.

Acknowledgement

The authors thank Aartjan te Velthuis for useful discussions.

Funding

This work was supported by NIAID/NIH contract (HHSN272201400008C).

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Conflict of interest: None declared.

References

Abdelwhab, E. M., Veits, J., and Mettenleiter, T. (2013) 'Genetic Changes That Accompanied Shifts of Low Pathogenic Avian Influenza Viruses toward Higher Pathogenicity in Poultry', *Virulence*, 4: 441–52.

Abolnik, C. (2017) 'Evolution of H5 Highly Pathogenic Avian Influenza: Sequence Data Indicate Stepwise Changes in the Cleavage Site', *Archives of Virology*, 162: 2219–30.

Alexander, D. J. (2007) 'An Overview of the Epidemiology of Avian Influenza', *Vaccine*, 25: 5637–44.

Bao, Y. et al. (2008) 'The Influenza Virus Resource at the National Center for Biotechnology Information', *Journal of Virology*, 82: 596–601.

Canale, A. S. et al. (2018) 'Synonymous Mutations at the Beginning of the Influenza A Virus Hemagglutinin Gene Impact Experimental Fitness', *Journal of Molecular Biology*, 430: 1098–115.

Chen, R., and Holmes, E. C. (2008) 'The Evolutionary Dynamics of Human Influenza B Virus', *Journal of Molecular Evolution*, 66: 655–63.

Dalby, A. R., and Iqbal, M. (2014) 'A Global Phylogenetic Analysis in Order to Determine the Host Species and Geography Dependent Features Present in the Evolution of Avian H9N2 Influenza Hemagglutinin', *PeerJ*, 2: e655.

Dietze, K. et al. (2018) 'From Low to High Pathogenicity—Characterization of H7N7 Avian Influenza Viruses in Two Epidemiologically Linked Outbreaks', *Transboundary and Emerging Diseases*, 65: 1576–87.

Dugan, V. G. et al. (2008) 'The Evolutionary Genetics and Emergence of Avian Influenza Viruses in Wild Birds', *PLoS Pathogens*, 4: e1000076.

Essere, B. et al. (2013) 'Critical Role of Segment-Specific Packaging Signals in Genetic Reassortment of Influenza A Viruses', *Proceedings of the National Academy of Sciences of the United States of America*, 110: E3840–8.

Ferhadian, D. et al. (2018) 'Structural and Functional Motifs in Influenza Virus RNAs', *Frontiers in Microbiology*, 9: 559.

Fouchier, R. A. et al. (2005) 'Characterization of a Novel Influenza A Virus Hemagglutinin Subtype (H16) Obtained from Black-Headed Gulls', *Journal of Virology*, 79: 2814–22.

García, M. et al. (1996) 'Heterogeneity in the Haemagglutinin Gene and Emergence of the Highly Pathogenic Phenotype among Recent H5N2 Avian Influenza Viruses from Mexico', *Journal of General Virology*, 77: 1493–504.

Gavazzi, C. et al. (2013a) 'An In Vitro Network of Intermolecular Interactions between Viral RNA Segments of an Avian H5N2 Influenza A Virus: Comparison with a Human H3N2 Virus', *Nucleic Acids Research*, 41: 1241–54.

—— et al. (2013b) 'A Functional Sequence-Specific Interaction between Influenza A Virus Genomic RNA Segments', *Proceedings of the National Academy of Sciences of the United States of America*, 110: 16604–9.

Gerber, M. et al. (2014) 'Selective Packaging of the Influenza A Genome and Consequences for Genetic Reassortment', *Trends in Microbiology*, 22: 446–55.

Gerlach, P. et al. (2015) 'Structural Insights into Bunyavirus Replication and Its Regulation by the vRNA Promoter', *Cell*, 161: 1267–79.

Gog, J. R. et al. (2007) 'Codon Conservation in the Influenza A Virus Genome Defines RNA Packaging Signals', *Nucleic Acids Research*, 35: 1897–907.

Gruber, A. et al. (2007) 'The RNAz Web Server: Prediction of Thermodynamically Stable and Evolutionarily Conserved RNA Structures', *Nucleic Acids Research*, 35: W335–8.

Gultyaev, A. P. et al. (2014) 'RNA Structural Constraints in the Evolution of the Influenza A Virus Genome NP Segment', *RNA Biology*, 11: 942–52.

—— et al. (2016) 'Subtype-Specific Structural Constraints in the Evolution of Influenza A Virus Hemagglutinin Gene', *Scientific Reports*, 6: 38892.

Horimoto, T., and Kawaoka, Y. (2005) 'Influenza: Lessons from Past Pandemics, Warnings from Current Incidents', *Nature Reviews Microbiology*, 3: 591–600.

Klenk, H. D., and Garten, W. (1994) 'Host Cell Proteases Controlling Virus Pathogenicity', *Trends in Microbiology*, 2: 39–43.

Knoepfel, S. A., and Berkhout, B. (2013) 'On the Role of Four Small Hairpins in the HIV-1 RNA Genome', *RNA Biology*, 10: 540–52.

Kobayashi, Y. et al. (2016) 'Computational and Molecular Analysis of Conserved Influenza A Virus RNA Secondary Structures Involved in Infectious Virion Production', *RNA Biology*, 13: 883–94.

—— et al. (2018) 'Conserved Secondary Structures Predicted within the 5' Packaging Signal Region of Influenza A Virus PB2 Segment', *Meta Gene*, 15: 75–9.

Kutchko, K. M. et al. (2018) 'Structural Divergence Creates New Functional Features in Alphavirus Genomes', *Nucleic Acids Research*, 46: 3657–70.

Lebarbenchon, C., and Stallknecht, D. E. (2011) 'Host Shifts and Molecular Evolution of H7 Avian Influenza Virus Hemagglutinin', *Virology Journal*, 8: 328.

Lee, D. H. et al. (2017) 'Highly Pathogenic Avian Influenza A(H7N9) Virus, Tennessee, USA, March 2017', *Emerging Infectious Diseases*, 23: 1860–3.

Lee, N. et al. (2017) 'Genome-Wide Analysis of Influenza Viral RNA and Nucleoprotein Association', *Nucleic Acids Research*, 45: 8968–77.

Li, W. et al. (2015) 'The EMBL-EBI Bioinformatics Web and Programmatic Tools Framework', *Nucleic Acids Research*, 43: W580–4.

Liu, S. et al. (2009) 'Panorama Phylogenetic Diversity and Distribution of Type A Influenza Virus', *PLoS One*, 4: e5022.

Lupiani, B., and Reddy, S. M. (2009) 'The History of Avian Influenza', *Comparative Immunology, Microbiology and Infectious Diseases*, 32: 311–23.

Ma, W., García-Sastre, A., and Schwemmle, M. (2015) 'Expected and Unexpected Features of the Newly Discovered Bat Influenza A-like Viruses', *PLoS Pathogens*, 11: e1004819.

McFadden, N. et al. (2013) 'Influence of Genome-Scale RNA Structure Disruption on the Replication of Murine Norovirus—Similar Replication Kinetics in Cell Culture but

Attenuation of Viral Fitness *In Vivo*', *Nucleic Acids Research*, 41: 6316–31.

Matsuzaki, Y. et al. (2016) 'Genetic Lineage and Reassortment of Influenza C Viruses Circulating between 1947 and 2014', *Journal of Virology*, 90: 8251–65.

Mekata, H. et al. (2018) 'Molecular Epidemiological Survey and Phylogenetic Analysis of Bovine Influenza D Virus in Japan', *Transboundary and Emerging Diseases*, 65: e355–60.

Mortimer, S. A., Kidwell, M. A., and Doudna, J. A. (2014) 'Insights into RNA Structure and Function from Genome-Wide Studies', *Nature Reviews Genetics*, 15: 469–79.

Moss, W. N., Priore, S. F., and Turner, D. H. (2011) 'Identification of Potential Conserved RNA Secondary Structure throughout Influenza A Coding Regions', *RNA*, 17: 991–1011.

Moumen, A. et al. (2003) 'Evidence for a Mechanism of Recombination during Reverse Transcription Dependent on the Structure of the Acceptor RNA', *The Journal of Biological Chemistry*, 278: 15973–82.

Munster, V. J. et al. (2010) 'Insertion of a Multibasic Cleavage Motif into the Hemagglutinin of a Low-Pathogenic Avian Influenza H6N1 Virus Induces a Highly Pathogenic Phenotype', *Journal of Virology*, 84: 7953–60.

Nao, N. et al. (2017) 'Genetic Predisposition to Acquire a Polybasic Cleavage Site for Highly Pathogenic Avian Influenza Virus Hemagglutinin', *mBio*, 8: e02298–16.

Noda, T. et al. (2012) 'Three-Dimensional Analysis of Ribonucleoprotein Complexes in Influenza A Virus', *Nature Communications*, 3: 639.

Penno, C. et al. (2017) 'Stimulation of Reverse Transcriptase Generated cDNAs with Specific Indels by Template RNA Structure: Retrotransposon, dNTP Balance, RT-Reagent Usage', *Nucleic Acids Research*, 45: 10143–55.

Perdue, M. L. et al. (1997) 'Virulence-Associated Sequence Duplication at the Hemagglutinin Cleavage Site of Avian Influenza Viruses', *Virus Research*, 49: 173–86.

Pflug, A. et al. (2017) 'Structural Insights into RNA Synthesis by the Influenza Virus Transcription-Replication Machine', *Virus Research*, 234: 103–17.

Pritlove, D. C. et al. (1999) 'A Hairpin Loop at the 5' End of Influenza A Virus Virion RNA Is Required for Synthesis of Poly(a)+ mRNA *In Vitro*', *Journal of Virology*, 73: 2109–14.

Rivas, E., Clements, J., and Eddy, S. R. (2017) 'A Statistical Test for Conserved RNA Structure Shows Lack of Evidence for Structure in lncRNAs', *Nature Methods*, 14: 45–8.

——, and Eddy, S. R. (2000) 'Secondary Structure Alone Is Generally Not Statistically Significant for the Detection of Noncoding RNAs', *Bioinformatics*, 16: 583–605.

Russell, R. J. et al. (2004) 'H1 and H7 Influenza Haemagglutinin Structures Extend a Structural Classification of Haemagglutinin Subtypes', *Virology*, 325: 287–96.

Shafiuddin, M., and Boon, A. C. M. (2019) 'RNA Sequence Features Are at the Core of Influenza A Virus Genome Packaging', *Journal of Molecular Biology*, doi: 10.1016/j.jmb.2019.03.018.

Shi, M. et al. (2018) 'The Evolutionary History of Vertebrate RNA Viruses', *Nature*, 556: 197–202.

Spronken, M. I. et al. (2017) 'A Compensatory Mutagenesis Study of a Conserved Hairpin in the M Gene Segment of Influenza A Virus Shows Its Role in Virus Replication', *RNA Biology*, 14: 1606–16.

Suzuki, Y., and Nei, M. (2002) 'Origin and Evolution of Influenza Virus Hemagglutinin Genes', *Molecular Biology and Evolution*, 19: 501–9.

Takizawa, N. et al. (2019) 'Local Structural Changes of the Influenza A Virus Ribonucleoprotein Complex by Single Mutations in the Specific Residues Involved in Efficient Genome Packaging', *Virology*, 531: 126–40.

te Velthuis, A. J. W., and Fodor, E. (2016) 'Influenza Virus RNA Polymerase; Insights into the Mechanisms of Viral RNA Synthesis', *Nature Reviews Microbiology*, 14: 479–93.

Thomas, S. M., Lamb, R. A., and Paterson, R. G. (1988) 'Two mRNAs That Differ by Two Nontemplated Nucleotides Encode the Amino Coterminal Proteins P and V of the Paramyxovirus SV5', *Cell*, 54: 891–902.

Tong, S. et al. (2013) 'New World Bats Harbor Diverse Influenza A Viruses', *PLoS Pathogens*, 9: e1003657.

Tuplin, A. et al. (2002) 'Thermodynamic and Phylogenetic Prediction of RNA Secondary Structures in the Coding Region of Hepatitis C Virus', *RNA (New York, N.Y.)*, 8: 824–41.

van den Born, E. et al. (2005) 'Discontinuous Subgenomic RNA Synthesis in Arteriviruses Is Guided by an RNA Hairpin Structure Located in the Genomic Leader Region', *Journal of Virology*, 79: 6312–24.

Veits, J. et al. (2012) 'Avian Influenza Virus Hemagglutinins H2, H4, H8, and H14 Support a Highly Pathogenic Phenotype', *Proceedings of the National Academy of Sciences of the United States of America*, 109: 2579–84.

White, M. C., Steel, J., and Lowen, A. C. (2017) 'Heterologous Packaging Signals on Segment 4, but Not Segment 6 or Segment 8, Limit Influenza A Virus Reassortment', *Journal of Virology*, 91: e00195–17.

Williams, G. D. et al. (2018) 'Nucleotide Resolution Mapping of Influenza A Virus nucleoprotein-RNA Interactions Reveals RNA Features Required for Replication', *Nature Communications*, 9: 465.

Witteveldt, J. et al. (2014) 'The Influence of Viral RNA Secondary Structure on Interactions with Innate Host Cell Defences', *Nucleic Acids Research*, 42: 3314–29.

Wong, S. S. et al. (2014) 'Characterization of an H4N2 Influenza Virus from Quails with a Multibasic Motif in the Hemagglutinin Cleavage Site', *Virology*, 468–470: 72–80.

Yoffe, A. M. et al. (2008) 'Predicting the Sizes of Large RNA Molecules', *Proceedings of the National Academy of Sciences of the United States of America*, 105: 16153–8.

Zuker, M. (2003) 'Mfold Web Server for Nucleic Acid Folding and Hybridization Prediction', *Nucleic Acids Research*, 31: 3406–15.