1 **Novel Computational Method to Define RNA PSRs Explains Influenza A Virus**
2 **Nucleotide Conservation**

3 Andrey Chursov[1], Nathan Fridlyand[2,3], Albert A. Sufianov[4,5], Oleg I. Kiselev[6, *], Irina
4 Baranovskaya[6], Andrey Vasin[6], Jonathan W. Yewdell[3] and Alexander Shneider[5,7]

5 [1] Datomics Group, 71 Amalienstr., Munich, 80799, Germany
6 [2] Laboratory of Translational Biology, School of Biosciences and Biotechnology, University of
7 Camerino, Via Gentile da Varano III, 62032, Camerino, Italy
8 [3] Cellular Biology Section, Laboratory of Viral Diseases, NIAID, NIH
9 [4] Federal Center of Neurosurgery of the Ministry of Health of Russian Federation, 5
10 Chervishevskiy Trakt, Tyumen, 625062, Russia
11 [5] I.M. Sechenov First Moscow State Medical University, 8 Malaya Trubetskaya St., Moscow,
12 119048, Russia
13 [6] Smorodintsev Research Institute of Influenza, Prof. Popova str. 15/17, Saint-Petersburg,
14 197376, Russia
15 [*] deceased
16 [7] Department of Molecular Biology, Ariel University, Ariel, Israel
17
18 Email addresses:

19 AC: achursov@datomicsgroup.com

20 NF: snatanf@gmail.com

21 AAS: info@fcn-tmn.ru

22 OIK: oleg_kiselov@hotmail.com

23 IB: irina.baranovskaja.1992@gmail.com

24 AV: influenza.spb@gmail.com

25 JWY: jyewdell@nih.gov

26 AS: ashneider@curelab.com

27

28 **ABSTRACT**

29 RNA molecules often fold into evolutionarily selected functional structures. Yet, the literature
30 offers neither a satisfactory definition for "structured RNA regions", nor a computational
31 method to accurately identify such regions. Here, we define structured RNA regions based
32 on the premise that both stems and loops in functional RNA structures should be conserved
33 among RNA molecules sharing high sequence homology. In addition, we present a
34 computational approach to identify RNA regions possessing evolutionarily conserved
35 secondary structures, RNA ISRAEU (RNA Identification of Structured Regions As
36 Evolutionary Unchanged). Applying this method to H1N1 influenza mRNAs revealed
37 previously unknown structured RNA regions that are potentially essential for viral replication

38 and/or propagation. Evolutionary conservation of RNA structural elements may explain, in
39 part, why mutations in some nucleotide positions within influenza mRNAs occur significantly
40 more often than in others. We found that mutations occurring in conserved nucleotide
41 positions may be more disruptive for structured RNA regions than single nucleotide
42 polymorphisms in positions that are more prone to changes. Finally, we predicted
43 computationally a previously unknown stem-loop structure and demonstrated that
44 oligonucleotides complementing the stem (but not the loop or unrelated sequences) reduce
45 viral replication *in vitro*. These results contribute to understanding influenza A virus evolution
46 and can be applied to rational design of attenuated vaccines and/or drug designs based on
47 disrupting conserved RNA structural elements.

48

49 **AUTHOR SUMMARY**

50 RNA structures play key biological roles. However, the literature offers neither a satisfactory
51 definition for "structured RNA regions" nor the computational methodology to identify such
52 regions. We define structured RNA regions based on the premise that functionally relevant
53 RNA structures should be evolutionarily conserved, and devise a computational method to
54 identify RNA regions possessing evolutionarily conserved secondary structural elements.
55 Applying this method to influenza virus mRNAs of pandemic and seasonal H1N1 influenza A
56 virus generated Predicted Structured Regions (PSRs), which were previously unknown. This
57 explains the previously mysterious sequence conservation among evolving influenza strains.
58 Also, we have experimentally supported existence of a computationally predicted stem-loop
59 structure predicted computationally. Our approach may be useful in designing live
60 attenuated influenza vaccines and/or anti-viral drugs based on disrupting necessary
61 conserved RNA structures.

62

63 **INTRODUCTION**

64 The biological functions of RNA secondary structures and their evolutionary impact is a topic
65 of great interest and importance (1-11). Since 1990s, conceptually novel computational
66 approaches have been appearing to analyze RNA shapes. Multiple copies of the same RNA
67 molecule fold into different coexisting conformations constituting an ensemble of RNA
68 structures. The same nucleotide within the RNA may be coupled via W-C bonds in some
69 conformations (being a part of a stem) while remaining uncoupled in others (i.e. belonging to
70 a loop). If one analyzes the entire ensemble, they can attribute to each nucleotide within an
71 RNA sequence its base pairing probability. This probability value reflects i) the percentage of
72 RNA structures which have this particular nucleotide W-C bonded (although in different
73 structures it may be bonded to a different coupling partners), and ii) the likelihood of each
74 RNA structure within the ensemble of all possible conformations that is based on the
75 structure's free energy. Hence, if an RNA molecule contains X nucleotides, a series of X
76 numbers ranging from 0 to 1 can be estimated such that each number reflects the probability
77 of a nucleotide in this position to be coupled.

78

79 Viruses represent an excellent system for studying RNA structural biology based on several
80 factors, including the high abundance of viral RNAs, the high mutation rate of many viruses,
81 the ease of selecting conditional mutants, and the large number of closely related strains
82 available for sequence conservation analysis.

83 Mutations disrupting influenza A virus RNA secondary structures dramatically reduce the
84 levels of gene expression (12). We previously reported that influenza A virus, a negative
85 strand RNA virus with a segmented genome, possesses clusters of nucleotides that
86 significantly change their base pairing probabilities with temperature elevation (13). This
87 suggests that local structures dispersed between non-structured RNA regions (nonPSRs)
88 are evolutionarily selected.

89 Previous attempts to define "structured RNA regions" were aimed either at finding regions
90 possessing the most stable secondary structure predicted by minimum free energy (MFE) of
91 based paired sequences (14-16), or predicting a consensus secondary structure based on a
92 given multiple-sequence alignment, which can be inferred either by means of energy-
93 directed folding or using a phylogenetic stochastic context-free grammar model (17-21). The
94 latter approach works well on short non-coding RNAs, which usually have one predicted
95 stable structure, but its accuracy drops significantly with the increasing RNA length (22). In
96 addition, mRNAs are typically less structured than non-coding RNAs, since structures
97 interfere with translation by ribosomes (23). At the same time, the former approach
98 misinterprets "PSRs" as stems possessing abnormally many W-C coupled nucleotides; i.e. it
99 would view an RNA structure possessing evolutionarily conserved loops (where the
100 nucleotides are not paired) as an unstructured element. The following example
101 demonstrates how misleading interchangeable use of the words "structured" and "paired" is:
102 a cloverleaf-like secondary structure may serve an indispensable biological function and be
103 conserved in every strain of some organism, despite the fact that it may have fewer paired
104 nucleotides than a simple stem (see Figure 1). Also, such approach may introduce systemic
105 bias typically identifying mRNAs as less structured than non-coding RNAs, since excessive
106 abundance of W-C pairing may interfere with translation by ribosomes (23).

107 In addition, any particular RNA sequence beyond certain length clearly can fold into stable
108 alternative states with energies being somehow different from the global minimum (24-26);
109 and several alternative RNA conformations coexist at equilibrium (27). Some of these RNA
110 structures may be present in multiple RNAs, especially homologous ones, while others
111 would exist for the particular sequence only. One can assume that evolutionary conservation
112 of the RNA shape may be indicative of its biological function. Current approaches relied
113 upon analysis of a single sequence cannot differentiate an evolutionarily conserved
114 structural element from an RNA shape that is energetically favorable only in a particular
115 strain. Thus, future progress in the field requires a new methodology. Addressing these
116 problems and proposing a computational methodology free of these shortcuts is the main
117 aim of this paper.

118

119 **Results**

120 **A Novel Quantitative Definition of Structured RNA Regions**

3

121   Here, we present a quantitative definition of a structured RNA region (PSR) that is equally
122   useful to predicting both, stems and loops as structured regions based on their evolutionary
123   conservation, and a new computational method for identifying those regions. The method is
124   robust to random RNA shapes present in a particular sequence but not selected and
125   conserved evolutionarily. We call this method RNA ISRAEU (RNA Identification of Structured
126   Regions As Evolutionary Unchanged).

127   As the first step in this method, we created a non-redundant dataset of sequences for each
128   RNA of interest constituted of highly homologous RNAs of the same length, and built
129   multiple sequence alignments. For the second step, the probability of every nucleotide to be
130   paired was calculated for each RNA sequence in the dataset. Third, we substituted every
131   nucleotide in the multiple alignment, with the nucleotide's pairing probability, thus aligning
132   pairing probabilities by nucleotide position. We took probabilities to be paired for all the 1st
133   nucleotides in each RNA sequence and grouped them together; for all 2nd nucleotides; for all
134   N^{th} nucleotides. Thus, if we have X RNA sequences each constituted of Y nucleotides, we
135   create Y sets of numbers ranging from 0 to 1; each set contains X numbers. Standard
136   deviation was computed for each set of probability values corresponding to every position
137   within the RNA. We proposed that such standard deviations be used as a measure of
138   structural conservation in a specific position. If the standard deviation for a particular position
139   within the RNA dataset was small, the probability of a nucleotide to be in a double-stranded
140   conformation did not vary substantially across the entire dataset of aligned mRNAs. We call
141   such positions "structure-conserved". In contrast, if the standard deviation was high, the
142   probability of a nucleotide to be paired changed vastly from strain to strain, a position was
143   called "structure-variable". The mean probability at a particular position did not matter to the
144   position classification.

145   We called regions within RNA sequences formed by consecutive structure-conserved
146   positions "Predicted Structured Regions" (PSRs), while regions predominantly formed by
147   structure-variable positions were called "non-structured" (nonPSR). Apparently, such
148   definition is stem-loop agnostic. A stretch of nucleotide positions, which demonstrate high
149   probability of being paired across the entire dataset of RNAs, may form a functionally
150   important and evolutionarily conserved stem. Similarly, a batch of nucleotide positions with
151   low base pairing probabilities, which repeats itself across all RNAs in the dataset may form a
152   functionally important and evolutionarily conserved loop. Still, further analysis is necessary to
153   confirm both the stems and the loops. In all cases, within a PSR, the probability of each
154   nucleotide to be in a double-stranded conformation does not vary significantly across the
155   entire dataset of aligned RNAs and these positions are structure-conserved positions.

156   **Identification of Structured RNA Regions in H1N1 Influenza Virus mRNAs**

157   We applied RNA ISRAEU to predict evolutionarily conserved RNA structures of influenza A
158   virus (IAV) (28, 29).

159   We selected sequences encoded by the complete genomes of 107 pre-pandemic (1999 to
160   2009) and 173 pandemic (post-2009) human H1N1 strains (Supplementary Table 2). In
161   2009, a swine IAV strain was introduced into man and rapidly replaced the circulating
162   strains.   All mRNAs selected for a given gene were of the same length. For each of the 10

4

163   major viral mRNAs, we predicted structured regions and calculated sequence variation.
164   Profiles for non-pandemic and pandemic NS2 mRNA are depicted in Figure 2 (profiles for
165   other mRNAs are presented in Supplementary Figures 2-38).

166   142 and 134 PSRs were identified in non-pandemic and pandemic influenza mRNAs
167   respectively. The length of PSRs varies from 5 to 121 nucleotides for non-pandemic and
168   from 5 to 103 nucleotides for pandemic strains (Table 1). The number of PSRs varies from 2
169   for NS2 to 31 for NP (Table 1). Only a fraction of PSRs overlap between pandemic and non-
170   pandemic strains. In some mRNAs (namely PB1, PB2, PA, HA and NA), the percentage of
171   such non-overlapping regions is higher than 70%. The location of each PSR are presented
172   in Supplementary Table 3.

173   In comparing PSR profiles with the profiles of mean pairing probabilities, we found two
174   evolutionarily conserved structural elements. One is located between positions 105 and 132
175   in non-pandemic NS2 mRNA (Figure 2), which contains a previously unknown predicted
176   conserved hairpin (Figure 3). Nucleotides 105 to 114 and 123 to 132 have a strong predicted
177   tendency to be paired while intervening nucleotides 115 to 124 have a strong tendency to be
178   unpaired. By comparing Figures 2(b) and 2(e), one can predict that this new hairpin structure
179   also exists in pandemic NS2 influenza mRNAs. The second novel PSR identified in non-
180   pandemic NS2 mRNA is between positions 24 and 89 (Figures 2 and 3). In this case
181   pandemic mRNAs contain only the PSR created by nucleotides 40 to 73.

182   **Oligonucleotides complementing the stem of the newly Predicted Structured Regions**
183   **interfere with *in vitro* viral replication**

184   To test the computationally predicted RNA structured regions, we have designed
185   oligonucleotides complementing the stem and the loop, as well as two controls of the same
186   length. The first control did not complement any sequence within the viral or human genome
187   while the other control bound a non-structured region adjacent to the PSR. The MDCK cell
188   monolayer was either transfected with one of the oligonucleotides and then infected with
189   A/California/7/09 strain or the cells were infected without prior transfection. The transfection
190   doze was not toxic for the cells as it was proven by the cell viability assay. Twenty four hours
191   post transfection and infection, the viral replication was assessed by developing the cell
192   monolayer with anti-NP ELISA. We observed that only the decamer complementing the stem
193   of the computationally predicted hairpin has significantly reduced the viral replication
194   comparing to the controls. Neither the oligonucleotide complementing the loop, nor the two
195   control oligos had a statistically significant effect on the *in vitro* viral replication (Figure 4).

196   **Location of the Most Mutable Positions**

197   We distinguish between two types of positions in the influenza genome – the mutable
198   positions which mutate quite frequently, and conserved positions. We tested if positions
199   mutating more often cluster outside of PSRs, while conserved positions are predominantly
200   located within the PSRs. The numbers for both types of positions in every mRNA of
201   pandemic and non-pandemic H1N1 strain are provided in the Supplementary Table 1.
202   Percentage of highly mutable positions in influenza mRNAs varies in a range from 7.9% in
203   M1 to 15.5% in NA and from 6.8% in M2 to 15.8% in HA for non-pandemic and pandemic
204   influenza strains respectively. Among the highly mutable positions from 56.1% for NS1 to

5

205 88.4% for NP and from 50.0% for M2 to 85.1% for M1 are third-codon positions for non-
206 pandemic and pandemic influenza mRNAs respectively. Results presented at plots (c) and
207 (f) of Figure 2 and Supplementary Figures 2-38 demonstrate that the most mutable positions
208 are randomly distributed within each mRNA and do not form clusters. Absence of
209 relationship between mutability value for every nucleotide position and corresponding value
210 of moving average of individual standard deviations of the probabilities of nucleotides to be
211 paired was confirmed by calculating Pearson correlation coefficients (Supplementary Table 4
212 and Supplementary Figures 39-58). All correlation coefficients were in a range from 0.006 to
213 0.222. This result refutes the intuitive notion that location of mutable positions would
214 correspond to the least structured RNA regions, while sequence conserved positions would
215 be collocated with the most structured RNA regions.

216 **Comparison of Mutation's Effect on RNA PSRs**

217 We generated two groups of *in silico* mutants by introducing synonymous mutations into
218 influenza mRNAs. In the first group, mutations were introduced only into positions that are
219 highly prone to being mutated; in the second one, mutations were introduced only into
220 sequence conserved positions. The number of introduced *in silico* mutations was
221 proportional to the length of every mRNA (Table 1). The effects of two groups of *in silico*
222 mutations on structured RNA regions were compared, as described in the Materials and
223 Methods section (Table 2). The results of statistical tests (Table 2) demonstrate that for
224 majority of mRNAs the mutations introduced into sequence conserved positions have a
225 greater effect on PSRs than mutations introduced into the mutable positions. This result
226 stands out the most in mRNAs of non-pandemic NP, M2, and NS1 genes and pandemic
227 NS2 gene.

228

229 **DISCUSSION**

230 Evolutionarily conserved RNA structural elements may perform important biological
231 functions. Hence, identification and/or prediction of such elements can help in the
232 understanding of the mechanism of RNA functions. This is true for identification of not only
233 paired regions (stems), but loops too. In fact, kissing loop interactions are a common type of
234 tertiary interaction motif in RNA that brings terminal loops together through Watson-Crick
235 base pairing. Also, bulged nucleotides in the loop-loop interaction can be critical for ligand-
236 dependent regulation. Yet, despite many efforts, it has still been a challenge to introduce an
237 objective, quantitative, biologically meaningful and computationally friendly definition of what
238 a "structured" RNA region is. Therefore, we had to propose a new definition and a new
239 computational methodology free of these shortcomings.

240 In analyzing an individual RNA sequence, one has little chance to distinguish a biologically
241 important structure formed by a folded molecule from simply a random shape with no
242 biological importance. However, if one observes the same RNA configuration conserved and
243 repeated across all related RNA sequences isolated from different strains and/or species,
244 this increases the likelihood of biologically significant RNA structure. Following this logic, a
245 definition of a structured RNA region should be based on a dataset of multiple aligned RNA
246 sequences. Thus, assume that some structural element in a particular location is of such

247 importance that it is present across all the strains. In this case, nucleotides in positions
248 correspondent to the stem would have very high base pairing probabilities in all aligned RNA
249 sequences of the dataset, while nucleotides in positions correspondent to the loop would
250 have very low base pairing probabilities in all the strains. At the same time, nucleotides
251 correspondent to a potential bistable structure would have their base pairing probabilities
252 neither too high nor very low across all RNA sequences. Therefore, we propose that
253 "structured RNA regions" are defined as the patterns of probability values of the nucleotides
254 to be paired, which are manifested across the spectrum of strains and/or organisms. This
255 definition equally considers the conservation of stems, loops and potential bistable structures
256 while also providing a computationally friendly quantitative definition for the degree of RNA
257 structure conservation.

258 Mathematically, the fact that nucleotides in a particular position in each RNA of a dataset are
259 likely to belong to an evolutionarily conserved structural element means that if we collect
260 values of pairing probabilities for this nucleotide from each RNA sequence in the dataset,
261 and build a sample of these values to calculate its standard deviation, this standard deviation
262 will be relatively low compared with the majority of standard deviations for other positions.
263 Indeed, if this standard deviation is low, it means that mutations occurring in the analyzed
264 RNA do not affect the base pairing probability of a nucleotide in this position across the
265 spectrum of strains. Thus, it is most likely that mutations affecting pairing probability for this
266 nucleotide are filtered out. This is a good indicator of evolutionary conservation and the
267 biological importance of the RNA structure in this position. In contrary, if the standard
268 deviation is high, it means that the correspondent nucleotide is very likely to be bonded in
269 some strains but not in others; hence, the presence of any crucial RNA structure at this
270 position is unlikely (unless there is a bistable secondary structure in this area playing roles in
271 different functions). If an RNA contained five consecutive nucleotides with low standard
272 deviations of their mean pairing probabilities, the region was considered structured.

Applications of a Newly Introduced Computational Definition

274 Introduction of a new definition adequately describing the subject matter under study and
275 development of a new technique for analysis, however, are only as good as they can be
276 applied to a multitude of biological phenomena, generate new observations and
277 experimentally testable hypotheses, explain old conundrums, and generate new questions
278 (41). The presented approach was used to examine the existence of structured RNA regions
279 in mRNAs of pandemic and non-pandemic influenza A H1N1 virus. This method revealed
280 that influenza mRNAs contain nucleotide positions highly conserved in their base pairing
281 probabilities. For every analyzed RNA type, such positions group together and constitute
282 well-defined structured RNA regions, while the rest of the RNA molecule is significantly less
283 structured. To the best of our knowledge, such mosaic structurization of RNA molecules was
284 not reported previously. *In vitro* testing has confirmed that interfering with a stem of a
285 previously unknown computationally predicted RNA structured region indeed reduces viral
286 replication. We expect that future experimental testing will reveal the functions, these
287 evolutionarily conserved RNA secondary structures, perform during the course of viral
288 infection.

7

289   We hypothesized that mosaic structurization of influenza mRNAs may explain a long-
290   standing conundrum of why different nucleotides in influenza genome mutate with such a
291   varied frequency. The enormous influence of amino acid conservation could explain only a
292   part of this phenomenon because many nucleotide substitutions are synonymous ones thus
293   cannot be explained by amino acid conservation. The first hypothesis was that if a mutation
294   happens within a structured RNA region it would disrupt the structure and be filtered out.
295   Thus, even if a mutation rate was the same for all nucleotides, the only mutations observed
296   in nature would be those happening outside of the structured RNA regions (PSR) and
297   neutral for RNA structures. If an exact picture of each RNA structure was available, it would
298   be possible to define structurally disruptive mutations visually as those that change the
299   shape(s) of the structure(s). However, modern computational methods do not make it
300   possible to predict exact RNA structures for long RNA molecules. Such predictions are
301   inaccurate and cannot be relied upon (14, 42-44). Thus, we had to define structurally
302   disruptive mutations based on the number of nucleotides in structured RNA regions, which
303   would change their W-C pairing probabilities to a level aberrant of their naturally observed
304   range. Contrary to original expectations, we showed that the nucleotide positions which are
305   the least prone to being mutated do not collocate with regions of conserved RNA structures.
306   Instead, the frequently and/or rarely mutating positions are randomly spread along the RNA
307   sequences. Although it was demonstrated that the most frequently mutating positions within
308   influenza genome are not collocated within unstructured RNA regions, this finding does not
309   refute the main hypothesis that states: "Mutations, which occur in nucleotide positions that
310   are the most prone to single nucleotide polymorphisms, have less of an effect on structured
311   RNA regions than mutations, which occur in positions that are less likely to be changed".

312   A mutation does not necessarily have to take place inside the PSR in order to be disruptive
313   for a structure. For example, prior to mutation, a particular G was paired to a particular C
314   forming a structure. If a mutation outside the structure changes some A to C, it may become
315   a new paring partner for the G, thereby leading to an energetically more favorable RNA
316   folding and disrupting the original structural element. This effect may be especially strong if
317   mutations outside of the PSRs occur in combinations. Also, mutations in certain positions
318   may have a greater effect on RNA structures than that of other positions. Thus, if RNA
319   structures should indeed remain intact for successful viral propagation, all positions, SNPs in
320   which would have a striking effect on the structures, would seem as rarely mutating
321   compared to those positions, SNPs in which would have little effect on the structures. The
322   results presented here support this hypothesis. We demonstrated for some influenza
323   mRNAs that *in silico* mutations introduced into nucleotide positions, which mutate in the wild
324   less frequently, would possess a greater disruptive effect on areas of conserved RNA
325   structures than *in silico* mutations in positions which are known to mutate more frequently.
326   As a result, mutations deleterious for vital RNA structures would be eliminated due to the
327   negative selection pressure. This demonstrates that conservation of RNA structures could
328   be a contributing mechanism defining a highly differential mutation rate for different influenza
329   nucleotide positions. Additionally, the computational conclusion stipulates a direction for
330   experimental testing. Although it is time/cost-consuming, it is possible to test RNA shapes
331   experimentally (4, 45-49). If our hypothesis is correct, then influenza mRNAs observed in
332   nature and those RNAs carrying mutations, which we predicted to be structurally non-
333   disruptive, would possess similar RNA structures. By contrast, introducing into the RNA

8

334 sequence mutations, which are predicted to disrupt structured RNA regions, would eliminate
335 at least some of the RNA structures vital for a virus.

Identification of Evolutionarily Conserved RNA Structural Elements

337 Plotting a graph with nucleotide positions on axis X and standard deviations of nucleotide
338 paring probabilities for these positions on axis Y shows stretches along the RNA sequence
339 with low standard deviations. These areas potentially have conserved RNA secondary
340 structural elements. However, these graphs alone do not demonstrate whether the
341 probability of a nucleotide to form bonds is high across different strains or low. In other
342 words, a structurization profile may help identifying localization of RNA PSRs, but it does not
343 indicate what kind of structure is there. Nevertheless, some assumptions about the RNA
344 shape can be made if we complement structurization profiles with profiles presenting mean
345 pairing probability for each nucleotide (i.e. for each nucleotide position in the RNA sequence,
346 the pairing probability values from every RNA in the dataset would be used to calculate the
347 mean for the position).

348 Extracting complex structures from comparing structurization profiles with profiles of mean
349 pairing probabilities may require special analytical tools that are not a part of this first-stage
350 study. However, discovering the simplest hairpin structure may not require additional
351 instruments. Thus, when 10 nucleotides were found to possess very high means of
352 probabilities to be bonded in the entire dataset, followed by 8 structurally conserved
353 nucleotides which were apparently uncoupled, and then another 10 nucleotides that are
354 likely to be paired and complementing the first 10 as W-C bonding partners, these findings
355 showed existence of a previously unknown evolutionarily conserved RNA hairpin structure.
356 *In vitro* testing has confirmed that interfering with a stem of a previously unknown
357 computationally predicted RNA structured region indeed reduces viral replication. We expect
358 that future experimental testing will reveal the functions, these evolutionarily conserved RNA
359 secondary structures, perform during the course of viral infection.

360 It would be important to test whether pandemic and seasonal influenza strains indeed share
361 some PSRs and whether the difference in RNA structurization may play a role in pandemic
362 vs. non-pandemic viral phenotypes. Another direction of the future research would be to
363 expand our computational definition of a structured RNA region to predict evolutionarily
364 conserved RNA tertiary structures, especially in those RNAs that are hard to study by high-
365 resolution experimental methods (50). In addition to helical segments, RNAs can fold into
366 complex three-dimensional shapes. Computational modeling of RNA tertiary structures and
367 determining of three-dimensional shapes of complex RNAs constitutes a major intellectual
368 challenge (51-55). Thus, the most practical way to expand the proposed computational
369 method to studying RNA 3D structures would be to incorporate RNA 3D structural modules
370 that define sets of non-Watson-Crick base pairs embedded in WC pairs (56, 57).

Novel Approach for the Rational Design of Live-Attenuated Vaccines and Anti-Viral Therapies

373 The method we proposed and applied to define structured RNA regions revealed several
374 areas possessing conserved secondary structures in mRNAs of H1N1 influenza virus. As a
375 next step, these structures have to be confirmed by *in vitro* analysis and their biological roles

376 have to be assessed *in vitro* and/or *in vivo*. Potentially, structurally conserved RNA regions
377 of viral RNAs may become a novel class of anti-viral drug targets. For example, anti-viral
378 agents selectively disrupting RNA structures vital for a viral life cycle may become a new
379 class of anti-viral therapies. As a preliminary proof of concept, we have demonstrated that an
380 oligonucleotide binding the computationally predicted stem of a hairpin in a PSR, indeed acts
381 as an anti-viral agent reducing *in vitro* viral replication. In contrast, statistically significant
382 effect on viral replication was not observed if the infected cells transfected with the oligos of
383 the same length, which bind outside of the predicted hairpin or do not bind to anything at all.
384 Interestingly, even an oligonucleotide complementing the loop of this hairpin was unable to
385 reduce viral replication in a statistically significant manner. Thus, the anti-viral effect was
386 specific to disrupting the hairpin's stem. RNA ISRAEU allows rapid rational design of
387 oligonucleotide cocktails interfering with multiple computationally predicted structures, so no
388 single or few mutations would result in a resistant viral strain.

389 Several approaches have been proposed for analysis of impact of SNPs on RNA structures
390 and deleterious mutation prediction (1, 58), including RNAsnp (59, 60), SNPfold (61),
391 RNAmute (62, 63), RNAmutants (64), and RDMAS (65). However, all these methods
392 compare structures of the original and mutated RNAs assessing the distance, the effect on
393 the RNA structure caused by SNPs. Although these methods are productive for the tasks
394 they were developed for, they cannot be applied to our problem. We do not compare
395 structures of an original and an altered RNA sequences. Instead, we compare structures of
396 hundreds of RNA sequences without attributing any of them the "original" status. Therefore,
397 the approach proposed here allows us to define: (i) a naturally occurring range of
398 probabilities, which represents a range of probability values that are the most likely to be
399 observed for natural RNA strains (see Quantitative Assessment of Mutation's Effect on RNA
400 PSRs in the Materials and Methods section for the specification) for every nucleotide
401 position within an RNA region possessing an evolutionarily conserved structure; (ii)
402 mutation(s) that would change base pairing probabilities within the structured RNA regions to
403 an extent that the new probabilities would not belong to a naturally occurring range for
404 corresponding positions.
405

406 Finally, we propose a new approach for the rational design of attenuated vaccines that would
407 be based on predicting mutations disruptive for conserved RNA structures and introducing
408 such mutations into viral genome. Indeed, disruption of an mRNA structure may serve as a
409 functional gene knock out reducing expression of a viral gene to a level insufficient for viral
410 cycle (12). Viral strain possessing such RNA can be administered to induce an immune
411 response with little risk for a patient. Such attenuated viral strains can be grown on
412 supporting cell lines actively expressing the limiting protein. Although LAVs are the most
413 successful achievements in the history of public health (38), we believe there were no prior
414 attempts to create LAVs based on perturbation of RNA structures.

415

416 **MATERIALS AND METHODS**
417
418 **Data**

10

419 Selecting H1N1 influenza mRNAs for this work constituted a crucial initial step. Influenza A
420 genome consists of eight segments encoding seventeen proteins (66). Seven of those
421 proteins were excluded from the analysis due to the limited information about them. It is
422 known, however, that different influenza segments have different mutation rates (67). To
423 eliminate potential bias that can be caused by disproportional representation of similar
424 hemagglutinin (HA) and neuraminidase (NA) sequences (these two influenza genes are
425 sequenced more often than the others because they constitute major viral antigens) and to
426 compare evolutionary structure conservation between different influenza mRNAs, only
427 completely sequenced influenza genomes were utilized in the analysis. An influenza genome
428 was considered completely sequenced if it had no missing parts, no unknown nucleotides,
429 and if sequences of the ten major mRNAs (namely, PB1, PB2, PA, HA, NP, NA, M1, M2,
430 NS1, and NS2) were known. In order to further increase coherence of the dataset, only
431 human influenza strains were utilized; other hosts were excluded because they demonstrate
432 different characteristics (68). Finally, only those strains possessing the identical length of
433 each influenza mRNA were selected. The fact that every mRNA of the same type has the
434 same length in every viral genome selected eliminates potential mistakes, which could be
435 introduced by effects of deletion and insertion polymorphisms (DIPs) on RNA secondary
436 structures. Sequences of pandemic and non-pandemic complete influenza genomes
437 satisfying the above mentioned criteria were downloaded from the Influenza Virus Resource
438 (http://www.ncbi.nlm.nih.gov/genomes/FLU/FLU.html) (69).

439 **Filtering Redundant Sequences**

440 Redundancy of data may introduce significant bias. To avoid it, one must only use a
441 representative subset of sequences instead of analyzing all possible strains. Therefore,
442 strains that were too similar were eliminated from further analysis; and, a non-redundant
443 subset of strains was created. Any two strains in the non-redundant subset possess no less
444 than 50 nucleotide differences per complete genome. In short, the first strain was chosen
445 randomly from the dataset described in the previous section, then added to the non-
446 redundant subset. Then, a different strain was randomly chosen and added to the non-
447 redundant subset only if the newly chosen sequence had at least 50 nucleotide differences
448 versus all strains in the non-redundant subset. This step was repeated until no more
449 sequences could be added to the non-redundant subset. The described procedure was done
450 separately for the pandemic and non-pandemic influenza datasets described above.

451 **Structural Conservation of a Nucleotide Position**

452 As a first step, for each mRNA sequence in the datasets, the probability of every nucleotide
453 within an RNA chain to be coupled via W-C bond was calculated. For that purpose, the
454 RNAfold tool from the Vienna RNA package was used (70). RNAfold was used with the
455 command line options –p that calculates the partition function and base pairing probability
456 matrix, --noLP that disallows base pairs that can only occur as helices of length 1, and the
457 default folding temperature fixed at 37°C. As a result, if a non-redundant dataset consisted
458 of N sequences, a sample of N probability values would be created for each position (exactly
459 N for a position, in which there is no deletion/insertion polymorphisms) within analyzed RNA.
460 The standard deviation was calculated for every sample. The procedure described above
461 was conducted for each of ten mRNAs from both subsets. Thus, we have calculated

11

462   standard deviations of the probabilities of nucleotides to be paired for every nucleotide
463   position within each of ten influenza mRNAs in both pandemic and non-pandemic datasets.
464   To smooth stochastic fluctuations, moving averages of individual standard deviations with
465   the sliding window of 5-nt length were calculated (Figure 2). To determine structure
466   conserved positions, all moving average values of individual standard deviations from all
467   mRNAs were combined to one dataset of moving averages, and the mean value and
468   standard deviation of the values in that dataset were calculated. If an individual moving
469   average calculated for a particular position was smaller than the overall mean of moving
470   averages minus the overall standard deviation of moving averages, the correspondent
471   position was considered "structure-conserved".

RNA Structurization and Structured RNA Regions

473   As described above, noticeable areas of structure-conserved positions possessing low
474   standard deviation values were observed. Areas possessing at least five consequent
475   structure-conserved nucleotides were defined as "structured RNA regions". The described
476   procedure was repeated separately for pandemic and non-pandemic influenza strains.

Mutability

478   Intuitively, "mutability" demonstrates how likely it is for a nucleotide in a particular position to
479   be mutated. Mathematically, this simple notion is defined as the value of Shannon entropy
480   (35), which is calculated based on the frequencies of every ribonucleotide recorded in a
481   particular position, with a pseudocount regularizers equal to 1 being added to the frequency
482   of each of four ribonucleotides according to Laplace's rule. To identify nucleotide positions
483   that are the most/least prone to being mutated, the mutability value was computed for each
484   nucleotide position. The more variable a set of ribonucleotides observed in a particular
485   position, the higher the entropy. Then, all mutability values from all mRNAs were combined
486   into one dataset. Those positions that had their mutability values higher than the 80th
487   percentile of the dataset were considered as mutable positions. In contrast, those positions
488   that did not contain SNPs among the sequences in the dataset were considered conserved
489   positions. The described procedure was repeated separately for pandemic and non-
490   pandemic influenza strains.

Quantitative Assessment of Mutation's Effect on RNA PSRs

492   Following the analysis discussed above, a new method was proposed and implemented,
493   which defines "structurally disruptive mutations" based on their effect on structured RNA
494   regions (PSRs). As described previously, two datasets of aligned influenza sequences were
495   created. For each individual RNA sequence within the datasets, the probability of each
496   nucleotide to be paired was computed. For every nucleotide position within coding regions of
497   influenza mRNA sequences, the mean value and the standard deviation of the probabilities
498   of nucleotides to be paired were calculated. Based on these values, a naturally occurring
499   range of probabilities was calculated for every nucleotide position within a PSR. A naturally
500   occurring range of probabilities was defined as a range of probabilities from the mean value
501   decreased by two standard deviations to the mean value increased by two standard
502   deviations.

12

503   Apparently, a mutation occurring in an RNA sequence may change probability of forming
504   Watson-Crick pairs for multiple nucleotides within a particular sequence. For some of those
505   nucleotides, their new probability values would still belong to the naturally occurring range of
506   probabilities for this position. For other positions, the mutation would change their pairing
507   probabilities to an extent that the new probabilities would not belong to their naturally
508   occurring range. A quantitative effect of mutation(s) on RNA structurization is defined as a
509   number of nucleotides within structured RNA regions (PSRs), which would change their
510   probabilities to an extent that the new probabilities would not belong to a naturally occurring
511   range for corresponding positions.

512   **Statistical Analysis: Do Mutations Taking Place in the Most vs. Least Often Mutating**
513   **Positions Have a Different Effect on RNA Structurization?**

514   Some positions in influenza genome are more prone to being mutated than others. The
515   ability to define quantitatively effects of mutations on RNA structurization permitted the
516   opportunity to propose a method for assessment, if mutations taking place in the frequently
517   mutating positions have the same effect on RNA structurization as mutations occurring in the
518   conserved ones. Two sets of *in silico* mutants were generated introducing synonymous
519   mutations in nucleotide positions that are either the most or the least prone to being
520   mutated. These two sets of mutations were compared for their effect on structured RNA
521   regions.

522   In order to normalize for the length difference among influenza mRNAs, the number of
523   changed nucleotides, which were introduced into each mRNA, was in proportion to the
524   length of the mRNA (Table 1). The required number of synonymous SNPs was introduced
525   into every mRNA sequence from the original datasets. In order to generate an *in silico*
526   mutant from an original mRNA sequence, the required number of positions that are the most
527   or the least prone to being mutated were randomly selected. Every codon, which contains
528   the selected position, was changed to an alternative one encoding the same amino acid with
529   the condition that the new codon is not observed in the particular position in any mRNA
530   sequence from the datasets. Influenza mRNAs contain relatively high number of conserved
531   positions and relatively few often mutating ones. As a result, for every mRNA, the number of
532   *in silico* mutants with SNPs in conserved positions was equal to the number of wild type
533   influenza strains in the datasets. However, due to a small number of frequently mutating
534   positions, it was impossible for some mRNAs to generate the same number of unique
535   mutants by introducing SNPs only to positions prone to being mutated. In those cases, all
536   possible mutants were kept for further analysis - namely, 103 for non-pandemic M2, 95 for
537   pandemic M2, and 109 for pandemic NS2.

538   For each computer-generated mutant, the probability of every nucleotide to be in a double-
539   stranded conformation was calculated. Based on those probabilities, we calculated the
540   number of nucleotides within structured RNA regions (PSRs), which changed their
541   probability of being paired to a value outside of the naturally occurring range of probabilities
542   for this position. Such numbers were combined into two datasets: one for mutations
543   introduced into highly mutable positions and another – for mutations introduced in highly
544   conserved positions. The Mann-Whitney U test was conducted for comparing these two
545   datasets. The significance level for the test was Bonferroni-corrected by dividing the

13

546   significance level of 5% by the total number of mRNAs in influenza virus, i.e. 10. The
547   described procedure was repeated separately for pandemic and non-pandemic influenza
548   strains.

549   **Profiles of Mean Pairing Probabilities**

550   Profiles of mean pairing probabilities were created for influenza mRNAs (Figure 2 and
551   Supplementary Figures 2-38). These profiles demonstrate how likely on average each
552   nucleotide within an mRNA is to be paired based on an analysis of the entire dataset of
553   sequences. As mentioned, for every RNA sequence in the dataset, the probability of every
554   nucleotide within the RNA chain to be coupled via W-C bond was calculated. Then, for every
555   nucleotide position, we computed the mean for probability values of this nucleotide based on
556   all RNA sequences. The resulting series of means is used as a profile of mean pairing
557   probabilities for a particular mRNA. The same work was performed for every influenza
558   mRNA.

559   **Virus and Cells**

560   Influenza virus A/California/7/09 (H1N1pdm) was provided by the Research Institute of
561   Influenza museum of viruses, Saint-Petersburg, Russian Federation. The 50% tissue culture
562   infective dose (TCID50) of this virus strain was defined by Reed–Muench method (72). The
563   aliquots of virus were stored at -80oC. According to the American Tissue Madin-Darby
564   canine kidney (MDCK) cell culture was provided from the cell collection of Research Institute
565   of Influenza, Saint-Petersburg, Russian Federation. Cells were cultivated in cultural flasks
566   using minimum essential medium Eagle alpha modification (aMEM, Biolot) with 2mM L-
567   glutamine supplemented with 10% heat-inactivated fetal bovine serum (FBS, GIBCO, USA).

568   **Design of Antisense DNA-oligonucleotides**

569   We designed antisense oligonucleotides, which may potentially disrupt the aforementioned
570   predicted RNA-structure. A random oligonucleotide, "rand10", with minimal probability of
571   having targets in the human hosts and viral genome was used as a control . Another control,
572   "off10", was an oligonucleotide with a target to the adjacent region on the NS2 gene mRNA
573   (Table 3).

574

575   **Cell Viability Assay**

576   The cell viability was determined 24 hours post infection and transfection by
577   microtetrazolium test (MTT assay). A solution of MTT [3-(4,5-dimethylthiazol-2-yl)-2,5-
578   diphenyltetrazolium bromide] (Sigma) at a concentration of 2,5 mg/ml was prepared in PBS.
579   The medium was removed, the cells were washed once with PBS, MTT solution was added
580   into the wells (100 ul/well). The cells were incubated at 37°C and 5% CO_2 for 4 hours and
581   then the solution was removed and 96% ethanol (100 ul/well) was added for formazan
582   crystals dissolving. The absorbance signal was measured using multifunctional reader
583   CLARIOstar ®(BMG LABTECH, Germany) at 535 nm.

584   **Virus Infection and Transfection**

585 The cells were detached by 0,25% trypsin/EDTA solution (Invitrogen) for 5 min and plated in
586 96-well plates (Nucl) at 104 cells per well the day before the infection experiment. Cells were
587 washed twice with Dulbecco's Phosphate-Buffered Saline (DPBS, GIBCO) and infected with
588 A/California/7/09 (H1N1pdm) viral strain in 100TCID50 dose per well. The medium for cells
589 infection was minimum essential medium Eagle alpha modification (aMEM, Biolot) with 2mM
590 L-glutamine, 2,5 ug/ml trypsin TPCK-treated from bovine pancreas (TPCK, Sigma) and
591 1:100 antibiotic-antimycotic (100X, GIBCO). Inoculation was conducted at 37°C and 5%
592 CO_2 for 60 minutes. Then the medium was removed and the cells were transfected using
593 100 µl of OptiPro SFM medium (GIBCO) contained 10 µM of DNA-oligonucleotides and 0.7
594 µl/well of Lipofectamine 2000 (Invitrogen) according to the manufacturer's protocol. In
595 addition, the transfection medium is also supplemented with 1:100 antibiotic-antimycotic
596 (100X) and 2,5 ug/ml TPCK. Viral control samples were also transfected with lipofectamin
597 2000 only, without any oligonucleotides. Four hours post-transfection, the medium was
598 replaced with fresh aMEM (Biolot) which contained 2mM L-glutamine, 2,5 ug/ml TPCK and
599 1:100 antibiotic-antimycotic (100X). Twenty four hours post- infection, cells were used for the
600 further relative ELISA analyses. Each treatment was performed in triplicates.

601 **Enzyme-Linked Immunosorbent Assays (ELISA)**

602 Twenty four hours post influenza virus infection and transfection with oligonucleotides,
603 continuity of the cell monolayer was assessed microscopically. Then, medium was removed
604 and the MDCK cells in 96-wells Nunc plates were fixed with 150 µl per well of cold 80%
605 acetone at 4°C for 30 minutes. The fixed samples were washed three times with phosphate
606 buffered saline containing 0.05% Tween (PBS-T) and blocked with 5% milk dissolved in
607 PBST (200 ul/well) for 30 minutes at 37°C. The fixed cells were incubated with 1ug/ml
608 mouse antibody against NP-protein (100 ul/well) produced in the Influenza Research
609 Institute (clone 4H1) at 37°C for 1 hour. After the next three washes the secondary goat anti-
610 mouse antibody conjugated with horseradish peroxidase (GAM-HRP, BioRad, USA) was
611 added at 1ug/ml (100 ul/well) and incubated for 1 hour at 37°C. Cells were washed three
612 times with PBS-T followed by adding TMB Peroxidase EIA Substrate Kit (Bio-Rad, USA)
613 according to manufacturer's instructions for further absorbance analysis. The absorbance
614 was measured using multifunctional reader CLARIOstar ® (BMG LABTECH) as delta optical
615 density OD 450 – OD 655. The absorbance signal from uninfected cells was taken as zero
616 and was subtracted from the obtained values of the samples. The results were presented
617 relative to infection control.

618 <u>**Statistical Analysis of Viral Replication Inhibition Assay**</u>

619 Data shown are means +/- SD as percentage of untreated "Flu" group. P-values for
620 comparing the four treatment groups with the untreated group (Flu) were calculated using
621 student T test. The significance level for the test was Bonferroni-corrected by dividing the
622 significance level of 0.05 by the total number of group comparisons, i.e. 10. Analysis was
623 performed using the R software.

624

625 **FUNDING**

15

630

631   ACKNOWLEDGEMENTS

637

638   REFERENCES
639   1.   Bevilacqua PC, Blose JM. Structures, kinetics, thermodynamics, and biological functions of
640   RNA hairpins. Annu Rev Phys Chem. 2008;59:79-103.
641   2.   Chursov A, Frishman D, Shneider A. Conservation of mRNA secondary structures may filter
642   out mutations in Escherichia coli evolution. Nucleic Acids Res. 2013;41(16):7854-60.
643   3.   Wan Y, Kertesz M, Spitale RC, Segal E, Chang HY. Understanding the transcriptome through
644   RNA structure. Nat Rev Genet. 2011;12(9):641-55.
645   4.   Wan Y, Qu K, Ouyang Z, Kertesz M, Li J, Tibshirani R, et al. Genome-wide measurement of
646   RNA folding energies. Mol Cell. 2012;48(2):169-81.
647   5.   Dethoff EA, Chugh J, Mustoe AM, Al-Hashimi HM. Functional complexity and regulation
648   through RNA dynamics. Nature. 2012;482(7385):322-30.
649   6.   Geisler S, Coller J. RNA in unexpected places: long non-coding RNA functions in diverse
650   cellular contexts. Nat Rev Mol Cell Biol. 2013;14(11):699-712.
651   7.   Rouskin S, Zubradt M, Washietl S, Kellis M, Weissman JS. Genome-wide probing of RNA
652   structure reveals active unfolding of mRNA structures in vivo. Nature. 2014;505(7485):701-5.
653   8.   Shabalina SA, Spiridonov NA, Kashina A. Sounds of silence: synonymous nucleotides as a key
654   to biological regulation and complexity. Nucleic Acids Res. 2013;41(4):2073-94.
655   9.   Mortimer SA, Kidwell MA, Doudna JA. Insights into RNA structure and function from
656   genome-wide studies. Nat Rev Genet. 2014;15(7):469-79.
657   10.   Tuplin A. Diverse roles and interactions of RNA structures during the replication of positive-
658   stranded RNA viruses of humans and animals. J Gen Virol. 2015;96(Pt 7):1497-503.
659   11.   Chursov A, Kopetzky SJ, Bocharov G, Frishman D, Shneider A. RNAtips: Analysis of
660   temperature-induced changes of RNA secondary structure. Nucleic Acids Res. 2013;41(Web Server
661   issue):W486-91.
662   12.   Ilyinskii PO, Schmidt T, Lukashev D, Meriin AB, Thoidis G, Frishman D, et al. Importance of
663   mRNA secondary structural elements for the expression of influenza virus genes. OMICS.
664   2009;13(5):421-30.
665   13.
666   Chursov A, Kopetzky SJ, Leshchiner I, Kondofersky I, Theis FJ, Frishman D, et al. Specific temperature-
667   induced perturbations of secondary mRNA structures are associated with the cold-adapted
668   temperature-sensitive phenotype of influenza A virus. RNA Biol. 2012;9(10):1266-74.
669   14.   Seetin MG, Mathews DH. RNA structure prediction: an overview of methods. Methods Mol
670   Biol. 2012;905:99-122.

671   15.   Gruber AR, Findeiss S, Washietl S, Hofacker IL, Stadler PF. RNAz 2.0: improved noncoding
672   RNA detection. Pac Symp Biocomput. 2010:69-79.
673   16.   Soldatov RA, Vinogradova SV, Mironov AA. RNASurface: fast and accurate detection of
674   locally optimal potentially structured RNA segments. Bioinformatics. 2014;30(4):457-63.
675   17.   Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermuller J, et al. Structured
676   RNAs in the ENCODE selected regions of the human genome. Genome Res. 2007;17(6):852-64.
677   18.   Hofacker IL, Fekete M, Stadler PF. Secondary structure prediction for aligned RNA
678   sequences. J Mol Biol. 2002;319(5):1059-66.
679   19.   Knudsen B, Hein J. RNA secondary structure prediction using stochastic context-free
680   grammars and evolutionary history. Bioinformatics. 1999;15(6):446-54.
681   20.   Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free
682   grammars. Nucleic Acids Res. 2003;31(13):3423-8.
683   21.   Washietl S, Hofacker IL. Consensus folding of aligned sequences as a new measure for the
684   detection of functional RNAs by comparative genomics. J Mol Biol. 2004;342(1):19-30.
685   22.   Gardner PP, Giegerich R. A comprehensive comparison of comparative RNA structure
686   prediction approaches. BMC Bioinformatics. 2004;5:140.
687   23.   Chen C, Zhang H, Broitman SL, Reiche M, Farrell I, Cooperman BS, et al. Dynamics of
688   translation by single ribosomes through mRNA secondary structures. Nat Struct Mol Biol.
689   2013;20(5):582-8.
690   24.   Schultes EA, Bartel DP. One sequence, two ribozymes: implications for the emergence of
691   new ribozyme folds. Science. 2000;289(5478):448-52.
692   25.   Hobartner C, Micura R. Bistable secondary structures of small RNAs and their structural
693   probing by comparative imino proton NMR spectroscopy. J Mol Biol. 2003;325(3):421-31.
694   26.   Weeks KM. Advances in RNA structure analysis by chemical probing. Curr Opin Struct Biol.
695   2010;20(3):295-304.
696   27.   Thirumalai D, Hyeon C. RNA and protein folding: common themes and variations.
697   Biochemistry. 2005;44(13):4957-70.
698   28.   Fitch WM, Leiter JM, Li XQ, Palese P. Positive Darwinian evolution in human influenza A
699   viruses. Proc Natl Acad Sci U S A. 1991;88(10):4270-4.
700   29.   Nelson MI, Simonsen L, Viboud C, Miller MA, Taylor J, George KS, et al. Stochastic processes
701   are key determinants of short-term evolution in influenza a virus. PLoS Pathog. 2006;2(12):e125.
702   30.   Furuse Y, Suzuki A, Kamigaki T, Oshitani H. Evolution of the M gene of the influenza A virus in
703   different host species: large-scale sequence analysis. Virol J. 2009;6:67.
704   31.   Gultyaev AP, Fouchier RA, Olsthoorn RC. Influenza virus RNA structure: unique and common
705   features. Int Rev Immunol. 2010;29(6):533-56.
706   32.   Carrillo-Santisteve P, Ciancio BC, Nicoll A, Lopalco PL. The importance of influenza
707   prevention for public health. Hum Vaccin Immunother. 2012;8(1):89-95.
708   33.   de Jong JC, Claas EC, Osterhaus AD, Webster RG, Lim WL. A pandemic warning? Nature.
709   1997;389(6651):554.
710   34.   Garten RJ, Davis CT, Russell CA, Shu B, Lindstrom S, Balish A, et al. Antigenic and genetic
711   characteristics of swine-origin 2009 A(H1N1) influenza viruses circulating in humans. Science.
712   2009;325(5937):197-201.
713   35.   Shannon CE, Weaver W. The mathematical theory of communication. Urbana,: University of
714   Illinois Press; 1949. v (i.e. vii), 117 p. p.
715   36.   Cox NJ, Kitame F, Kendal AP, Maassab HF, Naeve C. Identification of sequence changes in the
716   cold-adapted, live attenuated influenza vaccine strain, A/Ann Arbor/6/60 (H2N2). Virology.
717   1988;167(2):554-67.
718   37.   Klimov AI, Cox NJ, Yotov WV, Rocha E, Alexandrova GI, Kendal AP. Sequence changes in the
719   live attenuated, cold-adapted variants of influenza A/Leningrad/134/57 (H2N2) virus. Virology.
720   1992;186(2):795-7.

17

721   38.   Ilyinskii PO, Thoidis G, Shneider AM. Development of a vaccine against pandemic influenza
722   viruses: current status and perspectives. Int Rev Immunol. 2008;27(6):392-426.
723   39.   Simmonds P, Tuplin A, Evans DJ. Detection of genome-scale ordered RNA structure (GORS) in
724   genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence.
725   RNA. 2004;10(9):1337-51.
726   40.   Aouacheria A, Navratil V, Lopez-Perez R, Gutierrez NC, Churkin A, Barash D, et al. In silico
727   whole-genome screening for cancer-related single-nucleotide polymorphisms located in human
728   mRNA untranslated regions. BMC Genomics. 2007;8:2.
729   41.   Shneider AM. Four stages of a scientific discipline; four types of scientist. Trends Biochem
730   Sci. 2009;34(5):217-23.
731   42.   Shapiro BA, Yingling YG, Kasprzak W, Bindewald E. Bridging the gap in RNA structure
732   prediction. Curr Opin Struct Biol. 2007;17(2):157-65.
733   43.   Marti-Renom MA, Capriotti E. Computational RNA structure prediction. Current
734   Bioinformatics. 2008;3(1):32-45.
735   44.   Chursov A, Walter MC, Schmidt T, Mironov A, Shneider A, Frishman D. Sequence-structure
736   relationships in yeast mRNAs. Nucleic Acids Res. 2012;40(3):956-62.
737   45.   Deigan KE, Li TW, Mathews DH, Weeks KM. Accurate SHAPE-directed RNA structure
738   determination. Proc Natl Acad Sci U S A. 2009;106(1):97-102.
739   46.   Mortimer SA, Trapnell C, Aviran S, Pachter L, Lucks JB. SHAPE-Seq: High-Throughput RNA
740   Structure Analysis. Curr Protoc Chem Biol. 2012;4(4):275-97.
741   47.   Loughrey D, Watters KE, Settle AH, Lucks JB. SHAPE-Seq 2.0: systematic optimization and
742   extension of high-throughput chemical probing of RNA secondary structure with next generation
743   sequencing. Nucleic Acids Res. 2014;42(21).
744   48.   Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, et al. Genome-wide measurement
745   of RNA secondary structure in yeast. Nature. 2010;467(7311):103-7.
746   49.   Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. In vivo genome-wide profiling
747   of RNA secondary structure reveals novel regulatory features. Nature. 2014;505(7485):696-700.
748   50.   Ding F, Lavender CA, Weeks KM, Dokholyan NV. Three-dimensional RNA structure
749   refinement by hydroxyl radical probing. Nat Methods. 2012;9(6):603-8.
750   51.   Cruz JA, Blanchet MF, Boniecki M, Bujnicki JM, Chen SJ, Cao S, et al. RNA-Puzzles: a CASP-like
751   evaluation of RNA three-dimensional structure prediction. RNA. 2012;18(4):610-25.
752   52.   Miao Z, Adamiak RW, Blanchet MF, Boniecki M, Bujnicki JM, Chen SJ, et al. RNA-Puzzles
753   Round II: assessment of RNA structure prediction programs applied to three large RNA structures.
754   RNA. 2015;21(6):1066-84.
755   53.   Tinoco I, Jr., Bustamante C. How RNA folds. J Mol Biol. 1999;293(2):271-81.
756   54.   Laing C, Schlick T. Computational approaches to 3D modeling of RNA. J Phys Condens
757   Matter. 2010;22(28):283101.
758   55.   Laing C, Schlick T. Computational approaches to RNA structure prediction, analysis, and
759   design. Curr Opin Struct Biol. 2011;21(3):306-18.
760   56.   Leontis NB, Westhof E. Geometric nomenclature and classification of RNA base pairs. RNA.
761   2001;7(4):499-512.
762   57.   Theis C, Zirbel CL, Zu Siederdissen CH, Anthon C, Hofacker IL, Nielsen H, et al. RNA 3D
763   Modules in Genome-Wide Predictions of RNA 2D Structure. PLoS One. 2015;10(10):e0139900.
764   58.   Barash D, Churkin A. Mutational analysis in RNAs: comparing programs for RNA deleterious
765   mutation prediction. Brief Bioinform. 2011;12(2):104-14.
766   59.   Sabarinathan R, Tafer H, Seemann SE, Hofacker IL, Stadler PF, Gorodkin J. RNAsnp: efficient
767   detection of local RNA secondary structure changes induced by SNPs. Hum Mutat. 2013;34(4):546-
768   56.
769   60.   Sabarinathan R, Tafer H, Seemann SE, Hofacker IL, Stadler PF, Gorodkin J. The RNAsnp web
770   server: predicting SNP effects on local RNA secondary structure. Nucleic Acids Res. 2013;41(Web
771   Server issue):W475-9.

18

772   61.   Halvorsen M, Martin JS, Broadaway S, Laederach A. Disease-associated mutations that alter
773   the RNA structural ensemble. PLoS Genet. 2010;6(8):e1001074.
774   62.   Churkin A, Barash D. RNAmute: RNA secondary structure mutation analysis tool. BMC
775   Bioinformatics. 2006;7:221.
776   63.   Churkin A, Gabdank I, Barash D. The RNAmute web server for the mutational analysis of RNA
777   secondary structures. Nucleic Acids Res. 2011;39(Web Server issue):W92-9.
778   64.   Waldispuhl J, Devadas S, Berger B, Clote P. RNAmutants: a web server to explore the
779   mutational landscape of RNA secondary structures. Nucleic Acids Res. 2009;37(Web Server
780   issue):W281-6.
781   65.   Shu W, Bo X, Liu R, Zhao D, Zheng Z, Wang S. RDMAS: a web server for RNA deleterious
782   mutation analysis. BMC Bioinformatics. 2006;7:404.
783   66.   Vasin AV, Temkina OA, Egorov VV, Klotchenko SA, Plotnikova MA, Kiselev OI. Molecular
784   mechanisms enhancing the proteome of influenza A viruses: an overview of recently discovered
785   proteins. Virus Res. 2014;185:53-63.
786   67.   Webster RG, Bean WJ, Gorman OT, Chambers TM, Kawaoka Y. Evolution and ecology of
787   influenza A viruses. Microbiol Rev. 1992;56(1):152-79.
788   68.   Brower-Sinning R, Carter DM, Crevar CJ, Ghedin E, Ross TM, Benos PV. The role of RNA
789   folding free energy in the evolution of the polymerase genes of the influenza A virus. Genome Biol.
790   2009;10(2):R18.
791   69.   Bao Y, Bolotov P, Dernovoy D, Kiryutin B, Zaslavsky L, Tatusova T, et al. The influenza virus
792   resource at the National Center for Biotechnology Information. J Virol. 2008;82(2):596-601.
793   70.   Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al.
794   ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6:26.
795   71.   Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic
796   Acids Res. 2003;31(13):3406-15.
797   72.   Reed LJ, Muench H. A simple method of estimating fifty per cent endpoints. Am J Hyg.
798   1938;27:493-497

19

TABLE AND FIGURE LEGENDS

Table 1: Numbers of identified structured RNA regions within influenza mRNAs.

Gene name	Non-pandemic						Pandemic						Number of *in silico* SNPs
	Total number	Length			Do not overlap with PSRs in pandemic strains		Total number	Length			Do not overlap with PSRs in non-pandemic strains		
		Min	Max	Median	Number	Percentage of total, %		Min	Max	Median	Number	Percentage of total, %	
PB2	23	5	31	7	19	82.6	18	5	34	8.5	13	72.2	7
PB1	15	5	32	7	13	86.7	15	5	103	9	13	86.7	7
PA	20	5	54	8	17	85.0	13	5	27	9	10	76.9	7
HA	11	5	10	9	9	81.8	21	5	24	8	19	90.5	5
NP	28	5	24	9	15	53.6	31	5	67	7	17	54.8	5
NA	7	5	18	10	6	85.7	6	7	24	10.5	5	83.3	5
M1	22	5	57	10	10	45.5	15	5	30	8	2	13.3	2
M2	5	7	34	14	2	40.0	5	5	43	9	3	60.0	1
NS1	8	5	46	22.5	6	75.0	2	5	26	15.5	0	00.0	2
NS2	3	7	121	14	0	00.0	8	5	48	11	2	25.0	1

20

| Total | 142 | 5 | 121 | 9 | 97 | 68.3 | 134 | 5 | 103 | 8 | 84 | 62.7 | 42 |

802

Type	Gene name	Total length of PSRs	Percentage of mRNA covered by PSRs	Number of nucleotides within structured RNA regions, which changed their probability of being paired to a value outside of the naturally occurring range of probabilities for this position				P-value
				Mutations in positions that are prone to be mutated		Mutations in conserved positions		
				Mean	Standard deviation	Mean	Standard deviation	
Non-pandemic	PB2	257	0.113	18.6	13.5	20.8	14.9	0.1395
	PB1	156	0.069	11.4	10.5	11.8	10.4	0.3264
	PA	262	0.122	21.8	20.4	24.0	18.8	0.064
	HA	85	0.050	5.9	5.4	5.3	5.6	0.0813
	NP	300	0.200	24.6	17.2	31.3	17.7	0.001
	NA	75	0.053	4.0	5.1	4.4	6.2	0.2283
	M1	334	0.440	24.3	23.3	21.1	21.2	0.1068
	M2	77	0.262	4.1	4.7	7.0	7.5	0.0022
	NS1	183	0.264	9.6	11.0	15.2	15.7	0.0008
	NS2	142	0.388	11.4	19.2	12.0	22.8	0.0311
ic	PB2	232	0.102	16.7	14.0	16.0	13.0	0.2883

PB1	251	0.110	15.8	14.3	17.0	14.6	0.1722
PA	138	0.064	12.0	10.7	12.5	10.8	0.3317
HA	205	0.121	17.5	18.0	15.9	18.5	0.0273
NP	351	0.234	41.2	27.5	44.8	27.8	0.0955
NA	84	0.060	4.7	6.8	5.5	7.7	0.3006
M1	174	0.229	11.1	11.8	12.0	11.1	0.0917
M2	77	0.262	10.5	8.7	9.8	11.2	0.0395
NS1	31	0.047	2.8	5.7	3.3	5.7	0.1341
NS2	146	0.399	10.4	14.7	21.5	28.0	0.0007

Table 2: Comparative analysis of the effects on RNA PSRs elicited by *in silico* mutations in frequently vs. rarely mutating positions of H1N1 influenza mRNAs.

807 **Table 3:** List of antisense oligonucleotides used to examine influenza viral replication inhibition.

Name	Sequence	Nucleotide position in NS2 mRNA
stem	CAGAGACTCG	105 -114
loop	TATATTTT	115-122
off10	CTTATTTCCT	221-230
rand10	TATCCCACAC	NA

808
809

810 ***Figure 1****: Computationally predicted by mfold 3.6 (71) (a) optimal and (b) one of many suboptimal secondary structures of tRNA. Mfold was*
811 *used with the default energy parameters including the folding temperature fixed at 37°C. Despite the fact that the left structure contains more*
812 *base pairs, the right structure is functional and evolutionarily conserved.*

813

814 ***Figure 2****: Structure variability and mutability profiles for non-pandemic ((a), (b), and (c)) and pandemic ((d), (e), and (f)) NS2 influenza mRNAs.*
815 *Plots (a) and (d) demonstrate structure conservation profiles; namely, they show the moving average that was calculated by applying a sliding*
816 *window approach to smooth individual fluctuations of standard deviations of nucleotide base pairing probabilities. The blue solid line*
817 *demonstrates the mean level of all moving average values, and the blue dashed line demonstrates the level equal to the mean of all moving*
818 *average values decreased by the standard deviation of all moving average values. In this case, the mean and the standard deviation were*
819 *computed based on all moving average values from all mRNAs of a particular type (pandemic or non-pandemic) of influenza strains. According*
820 *to our definition, when the moving average goes below the blue dashed line, it is a structured RNA region. Such regions are colored with either*
821 *yellow or green across the plots. Plots (b) and (e) demonstrate profiles of the mean values of probabilities of nucleotide positions to be in a*
822 *double-stranded conformation. If this value is close to 1, it means that in most strains in the dataset the correspondent nucleotide has a very*
823 *high probability to be paired; and, if this value is close to 0, the correspondent nucleotide is very likely to be unpaired in most strains in the*
824 *dataset. Plots (c) and (f) demonstrate mutability profiles for NS2 mRNAs. Mutability of every nucleotide position is computed as a value of*
825 *Shannon entropy which is calculated based on frequency of every ribonucleotide in a particular position. Areas within RNA colored with yellow*
826 *or green demonstrate identified structured RNA regions. Meanwhile, areas colored with green show regions in which particular secondary*
827 *structure was determined.*

23

828

Figure 3: *Secondary structural elements identified in the NS2 mRNA of H1N1 influenza A virus. These structural elements are evolutionarily conserved among analyzed strains. Hairpin at plot (b) was identified in both non-pandemic and pandemic H1N1 influenza. Structure shown at plot (a) exists in non-pandemic influenza virus, while pandemic mRNAs contain only part of that structure covered by nucleotide positions 40 to 73.*
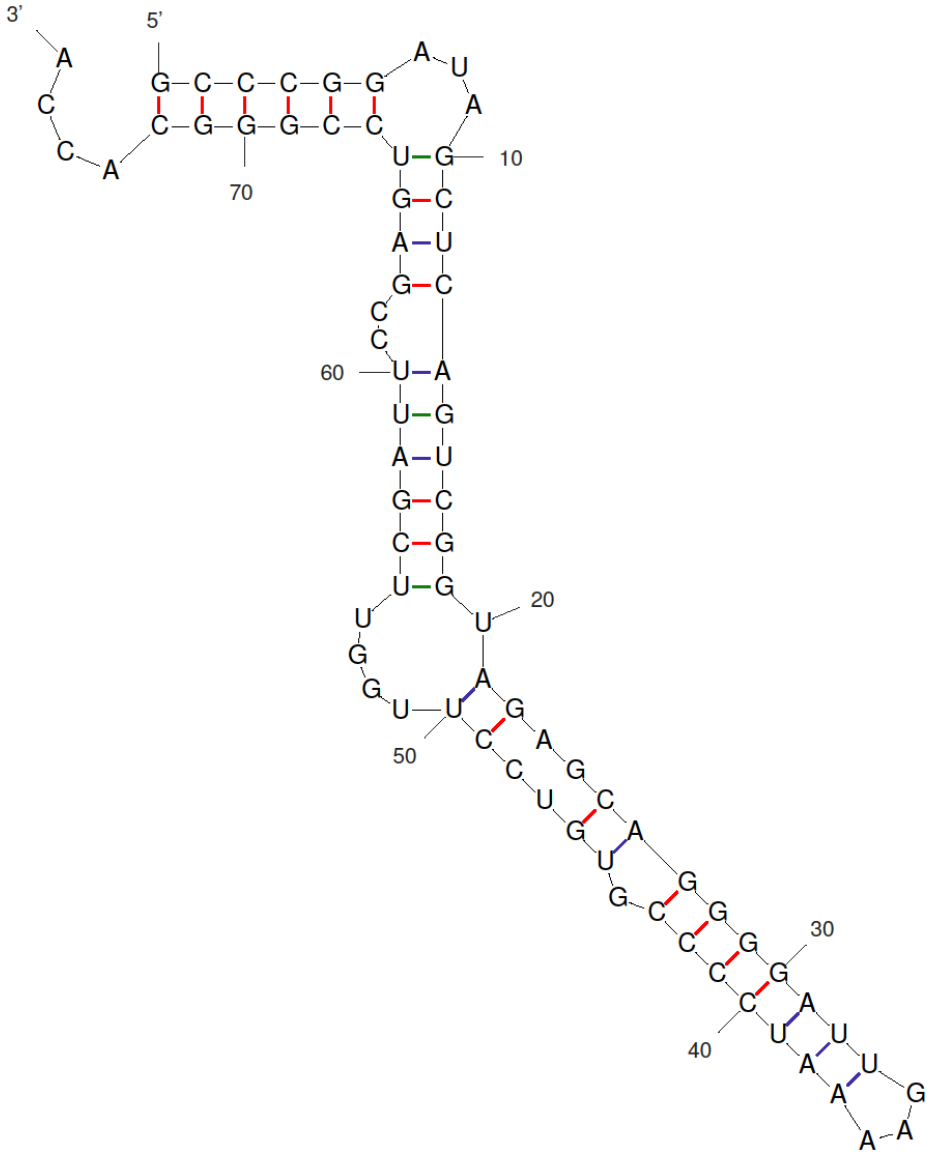
833

Figure 4: *Influenza viral replication inhibition effect of antisense oligonucleotides, 24 hours post infection. P-value for the comparison between "Stem" and "Flu" is 0.0043 and is the only statistically significant difference.*
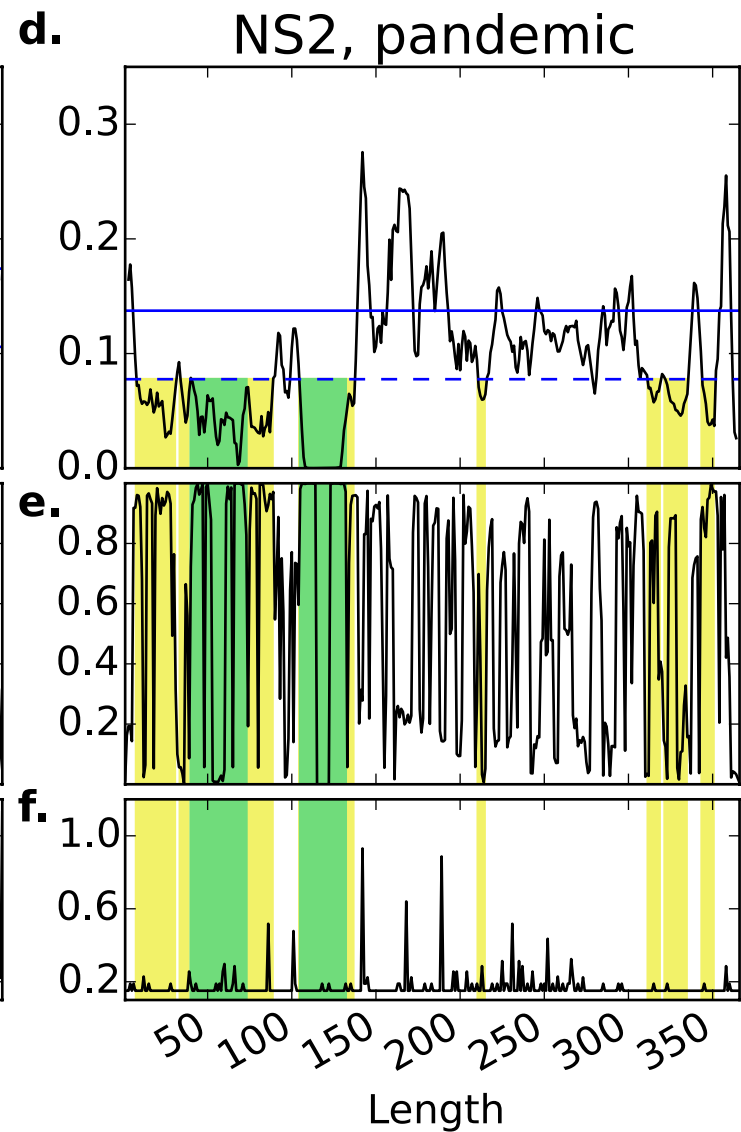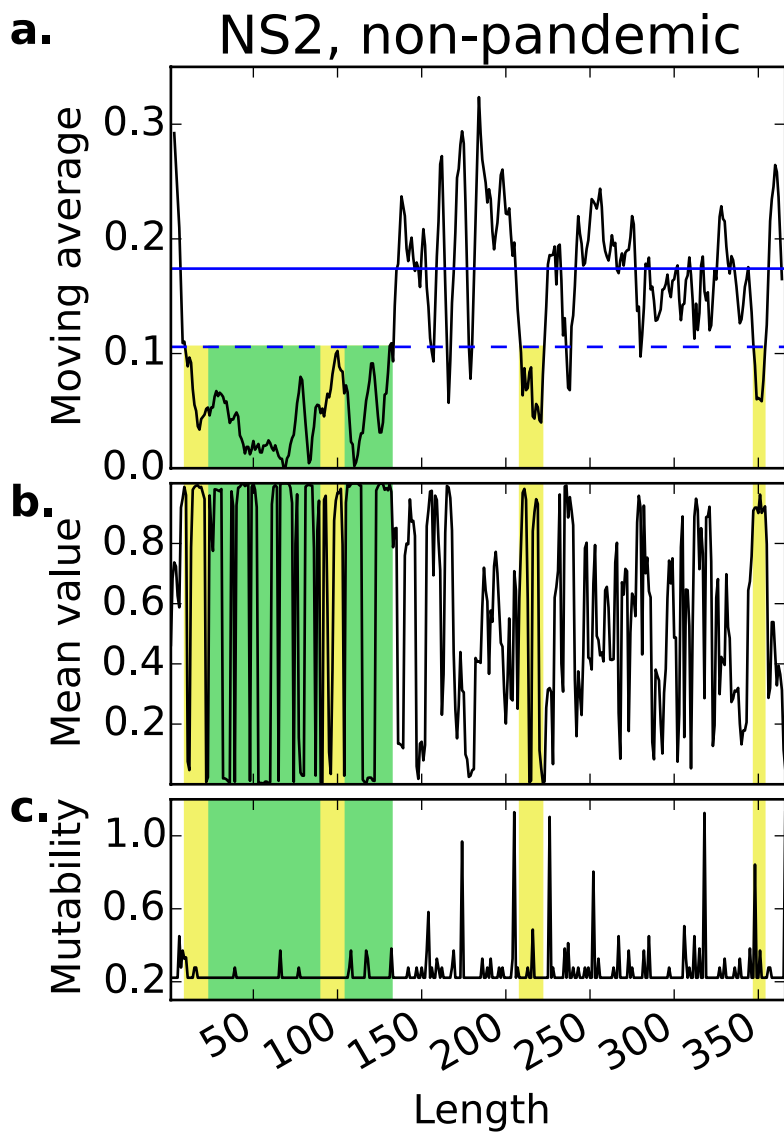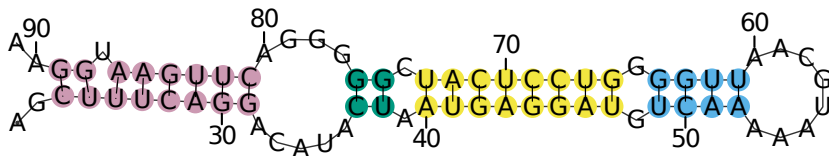
836

a.

b.

dG = -33.60

dG = -32.18

a. NS2, non-pandemic

d. NS2, pandemic

b.

e.

c.

f.

Moving average

Mean value

Mutability

Length
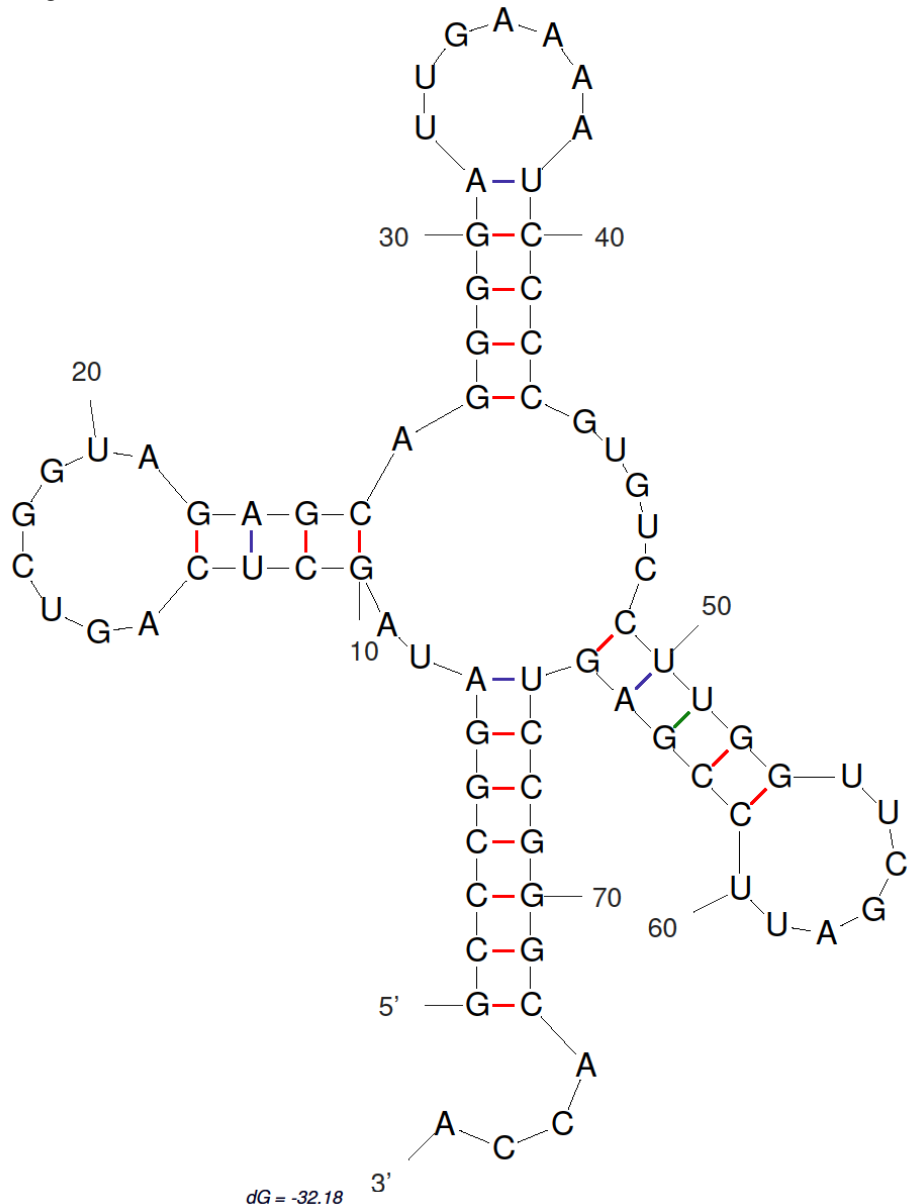
a.

... AG CUUUCAGG ACAUA CU A AUGAGGAU G UCAA AAAUGCAA UUGG G GUCCUCAU C GG GGGA CUUGAA U GG AA ...

b.

... UU CGAGUCUCUG AAAAUAUA CAGAGAUUCG CU ...