The genome of RNA viruses is highly constrained by the convergent interplay between protein evolution and RNA structure

Nabor Lozada-Chávez*, Peter F. Stadler[1-6] and Irma Lozada-Chávez[1]*

[1]Evo-Devo & Bioinformatics Group, Department of Computer Science – IZBI, University of Leipzig, Härtelstrasse 15-18, D-04107 Leipzig, Germany
[2] Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
[3] German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Competence Center for Scalable Data Services and Solutions, and Leipzig Research Center for Civilization Diseases, University Leipzig, Germany
[4] Fraunhofer Institute for Cell Therapy and Immunology, Perlickstrasse 1, D-04103 Leipzig, Germany
[5] Center for RNA in Technology and Health, University of Copenhagen, Grønnegardsvej 3, Frederiksberg C, Denmark
[6] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe NM, 87501, USA

* ilozada@bioinf.uni-leipzig.de, nabor.lozada@gmail.com

Background: Because genome size and complexity in RNA viruses is highly restricted, a conflict might exist among the needs to encode both proteins and structured RNAs. As a working hypothesis, we believe that the main features promoting genome conservation and novelty [1,2], such as codon selection, overlapping genes, intrinsic disordered regions and structured RNAs, are likely forced to constraint each other rather than to evolve independently. Here, we assessed how extended and strong is the co-evolution between the RNA structure and protein-coding codes along the RNA viral genomes.

Results: We systematically analyzed 10,000 complete genomes from 15 different RNA viruses to estimate four major features along each genome: selection (at the codon, domain and gene levels), amino acid conservation, IPDRs, and RNA secondary structures. We found that RNA viral genomes are mainly driven by purifying selection at the codon level, and that most viral proteins exhibit several short IPDRs. However, structural genes are characterized by an enrichment of disordered motifs and protein regions under positive selection; particularly, those accessory and regulatory proteins involved in interactions with other viral and host factors. By contrast, non-structural genes are usually dominated by more protein regions under negative selection and less disordered motifs. As in [3] and [4], we found a negative correlation between the extent of base-pairing in RNA structures and amino acid variability located at the same genomic regions. Likewise, amino acid sites encoded by structured RNAs exhibited stronger purifying selection than those not involved in structure-forming RNA. Finally, we found that coupled genes with the relaxed action of selection within overlapping regions are characterized by short sequences, more amino acid variability and IPDRs regions, as well as “flexible” RNA structures.

Conclusions: Our findings suggest that RNA structures and protein sequences in RNA viruses are indeed likely forced to constraint to each other rather than to evolve independently. This convergent interplay seems to promote the genome architecture into associated region-specific divisions of labor.

References:

[1] Brandes N, Linial M (2016) Gene overlapping and size constraints in the viral world. *Biol Direct* 11:26

[2] Xue B, et al. (2014) Structural disorder in viral proteins. *Chem Rev.* 114:6880

[3] Sanjuán R1, Bordería AV (2011) Interplay between RNA structure and protein evolution in HIV-1. *Mol Biol Evol.* 28:1333

[4] Watts JM, et al (2009) Architecture and secondary structure of an entire HIV-1 RNA genome. Nature 460:711