VeGETA: whole viral genome multiple sequence alignments based on RNA secondary structures

Introduction

RNA secondary structures play a vital role in many different RNA viruses. Just as one example, the Human Coronavirus 229E has a local stem-loop structure in the 5' UTR of its genome that is crucial for the replication of the virus [1]. However, local structures do not capture all RNA interactions of a single RNA molecule. Intramolecular long-range interactions (LRI) also facilitate translation and replication processes in viruses, as shown in Hepatitis C virus [2].
If RNA structures and LRIs have an important role within one virus, they are commonly conserved among the same species or even genus [2]. However, multiple sequence alignments (MSA) are sequence-based, since a structure-based whole genome MSA is not feasible in terms of computation time.

Objectives

Here, we present VeGETA, a new pipeline that creates a structure-based multiple sequence alignment of representative viruses of a species, genus or even family. Thus, VeGETA gives a structure-annotated MSA, enabling RNA structure related downstream analysis.

Materials & Methods

We showcase the functionality of VeGETA with Dengueviruses, Hepatitis C virus, Filoviruses, Flaviviruses and Coronaviruses [3]. For each species (or genus) a set of representative viral sequences is created. Then, we refine a sequence-based alignment progressively using LocARNA [4], a tool that considers RNA structure for alignment creation.

Results

We are able to create structure-based whole genome MSAs of our viral input sequences. Each of the structure-based whole genome MSAs produced by VeGETA, was compared to the known structural elements from literature. VeGETA produces the previously reported 5' UTRs and 3' UTRs.
Additionally, we give insights into the secondary structures of protein-coding regions in viral genomes.

Conclusion

With our new algorithm VeGETA we create structure-based whole genome MSAs of viral species and genera. These MSAs can be used for further downstream analysis, for example, the detection of conserved long-range interactions between two structural elements.

[1]: Madhugiri, R.; Karl, N.; Petersen, D.; Lamkiewicz, K.; Fricke, M.; Wend, U.; Scheuer, R.; Marz, M.; Ziebuhr, J. Structural and functional conservation of cis-acting RNA elements in coronavirus 5'-terminal genome regions. Virology 2018, 517, 44–55. doi:10.1016/j.virol.2017.11.025.
[2]: Nicholson, BL, White, KA (2014). Functional long-range RNA-RNA interactions in positive-strand RNA viruses. Nat. Rev. Microbiol., 12, 7:493-504
[3]: B. E. Pickett, E. L. Sadat, Y. Zhang, J. M. Noronha, R. B. Squires, V. Hunt, M. Liu, S. Kumar, S. Zaremba, Z. Gu, L. Zhou, C. N. Larson, J. Dietrich, E. B. Klem, R. H. Scheuermann, ViPR: an open bioinformatics database and analysis resource for virology research., Nucleic Acids Res 40 (Database issue) (2012) D593–D598. doi:10.1093/nar/gkr859.
[4]: Will, S.; Reiche, K.; Hofacker, I.L.; Stadler, P.F.; Backofen, R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLOS Comput Biol 2007