Furin Cleavage Sites in Sarbecoviruses and Merbecoviruses

MouthOfMadness

Abstract

This is a working document for now, to gather ideas and experiments (mostly computationals) around the natural or engineered emergence of Furin Cleavage Sites (FCS) in Sarbecoviruses and Merbecoviruses. We will rely on phylogenetic trees, propabilistic modeling etc.

Keywords: Sarbecoviruses, Merbecoviruses, FCS

1 Introduction

We will try to explore the characteristics in terms of propabilistic models the appearance or presence of furin-like sites.

We will focus on the following specific polybasic patterns:

- RXXR
- RRXR or RXRR

These are seen inside the Merbecovirus clade, it was recently demonstrated by Coutard et al (2020) that an FCS now also exist inside the Sarbecovirus clade due to the somewhat unique appearance of the Sars-Cov-2 and all its subsequent variants.

We will also look at quasi-patterns that we will note:

- $\overline{\text{RXXR}}$
- $\overline{\text{RRXR}}$ or $\overline{\text{RXRR}}$

These patterns represent sequences of nucleotides that are one mutation away from their representative pattern. It is important to note that these quasi-patterns exclude their representative version.

1.1 Data and Alignements

We used two datasets, one for the Merbecoviruses Ω_{Merb} and one for the Sarbecoviruses Ω_{Sarb}. Each dataset was independently aligned using the MAFFT program. Note that the original Wuhan-Hu-1 is included in Ω_{Sarb}.

1.2 RNA Stucture

Following Wu and Zhao (2021) we will segment the RNA virus strands in multiple regions. As a generic template we can use the structure of the Wuhan-Hu-1 genome (see Fig.1).

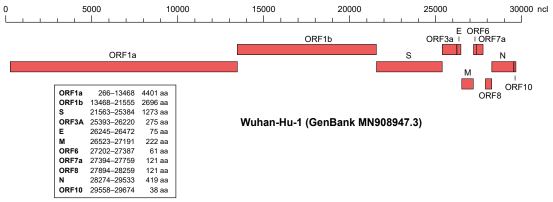

Fig. 1 Wuhan-Hu-1 Genome Structure

Here ORFx are Open Reading Frames that do not encode for structural proteins, M (resp. E) stands for the membrane (resp. envelope) protein, N for the nucleocapside (that is responsible for RNA strand packaging), and finally S the now infamous spike protein.

We also subdivide the S region in the three following subregions: S_1, S_{link} and S_2:

- S_1 represents the first S protein subunit wich contains the Receptor Binding Domain (RBD)
- S_{link} the amino-based link between S_1 and S_2
- S_2 represents the second S protein subunit.

1.3 Propabilistic Models

In the following Sections we will explore different types of propabilistic models. We will start with simple presence models in Sec.2.

2 Structural Uniform Model

We first focus on a very simple presence model for the FCS patterns. It is based on a piecewise uniform Probability Density Function (pdf).

$$f(x) = \begin{cases} p_{[b]} & x \in [0, S_1^b[\\ p_{[S_1]} & x \in [S_1^b, S_1^e] \\ p_{[S_{link}]} & x \in]S_1^e, S_2^b[\\ p_{[S_2]} & x \in [S_2^b, S_2^e] \\ p_{[e]} & x \in]S_2^e, N] \end{cases}$$

We use b (res. e) as short mnemonic for *begin* (resp. *end*). The size of the full RNA strand is N.

2.1 Merbecoviruses Ω_{Merb}

I will simply report (for now) the results.

$$f_{\mathrm{RXXR}}(x) = 1e^{-3} \times \begin{cases} p_{[b]} & = 0.525189 \;\; x \in [0, S_1^b[\\ p_{[S_1]} & = 0.843278 \;\; x \in [S_1^b, S_1^e] \\ p_{[S_{link}]} & = 6.61268 \;\; x \in]S_1^e, S_2^b[\\ p_{[S_2]} & = 0.491756 \;\; x \in [S_2^b, S_2^e] \\ p_{[e]} & = 1.02473 \;\; x \in]S_2^e, N] \end{cases}$$

$$f_{\overline{\mathrm{RXXR}}}(x) = 1e^{-2} \times \begin{cases} p_{[b]} & = 1.11984 \;\; x \in [0, S_1^b[\\ p_{[S_1]} & = 0.757632 \;\; x \in [S_1^b, S_1^e] \\ p_{[S_{link}]} & = 1.93922 \;\; x \in]S_1^e, S_2^b[\\ p_{[S_2]} & = 0.591803 \;\; x \in [S_2^b, S_2^e] \\ p_{[e]} & = 1.29879 \;\; x \in]S_2^e, N] \end{cases}$$

2.2 Sarbecoviruses Ω_{Sarb}

I will simply report (for now) the results.

$$f_{\mathrm{RXXR}}(x) = 1e^{-3} \times \begin{cases} p_{[b]} & = 0.327155 \;\; x \in [0, S_1^b[\\ p_{[S_1]} & = 0.29627 \;\; x \in [S_1^b, S_1^e] \\ p_{[S_{link}]} & = 0.0470411 \;\; x \in]S_1^e, S_2^b[\\ p_{[S_2]} & = 0.283475 \;\; x \in [S_2^b, S_2^e] \\ p_{[e]} & = 1.97162 \;\; x \in]S_2^e, N] \end{cases}$$

$$f_{\overline{\mathrm{RXXR}}}(x) = 1e^{-2} \times \begin{cases} p_{[b]} & = 1.06255 \;\; x \in [0, S_1^b[\\ p_{[S_1]} & = 1.0844 \;\; x \in [S_1^b, S_1^e] \\ p_{[S_{link}]} & = 0.23991 \;\; x \in]S_1^e, S_2^b[\\ p_{[S_2]} & = 1.3954 \;\; x \in [S_2^b, S_2^e] \\ p_{[e]} & = 1.82017 \;\; x \in]S_2^e, N] \end{cases}$$

References

Coutard B, Valle C, de Lamballerie X, et al (2020) The spike glycoprotein of the new coronavirus 2019-ncov contains a furin-like cleavage site absent in cov of the same clade. Antiviral Research 176:104,742. https://doi.org/https://doi.org/10.1016/j.antiviral.2020.104742, URL https://www.sciencedirect.com/science/article/pii/S0166354220300528

Wu Y, Zhao S (2021) Furin cleavage sites naturally occur in coronaviruses. Stem Cell Research 50:102,115. https://doi.org/https://doi.org/10.1016/j.scr.2020.102115, URL https://www.sciencedirect.com/science/article/pii/S1873506120304165