Information Theoretic Study of Covid 19 Genome

Philippe Jacquet

Information Theoretic Study of Covid 19 Genome

Philippe Jacquet IEEE Fellow
Inria
Paris, France
philippe.jacquet@inria.fr

Abstract

In this paper we analyse the genome sequence of covid 19 on a information point of view and we compare with past and present genomes. We use the powerful tool of joint complexity in order to quantify the various potential parent genomes.

Index Terms

Genome, Covid, Joint Complexity, pattern matching.

I. Introduction

The outbreak of the pandemic SARS-2 Covid 19 disease has been the major event of these two last years. There have been many speculations on the origin of the virus and about its future and pasts mutations. Therefore the genome of the SARS-2 has draw many attention. The foundation of Information Theory is to extract patterns and similarities between structures without necessarily relying on the functional significance of the common fragments, *e.g.* the meaning of words in texts, or the proteins translated in genomes. Nevertheless we will show that the information theoretic tools, such as the joint complexity, are powerful enough to draw some conclusions about the recent speculations on the origin of the virus even in absence of medical specific background.

The paper is organized as follows: first we will shortly introduce the concept of joint complexity and reminds the background results on random strings about "weak" pattern matching and establish some new results about "strong" pattern matching. Second we will present our result about weak pattern matching which establish that the covid 19 is a descendant of bat coronaviruses. We also establish that the virus of the HIV should not be an ancestor. Third we address the area of strong pattern matching in analysing the similarities with recent bat coronaviruses.

II. The joint complexity tool and performance

Let's take a finite alphabel $\mathcal{A}$ and a finite sequence X over $\mathcal{A}$. A factor is sequence $\mathbf{v}$ such that there exists two other sequences u and w such that $X = u\mathbf{v}w$. We call "string complexity" of X the number of different factors of X [4], [5]. Let Y be another sequence, we call "joint complexity" the number of different factors common to X and Y.

These quantities are easy to compute. Indeed the string complexity is simply the number of internal nodes in the extended suffix tree [3] (also called the spaghetti suffix tree). It can also be computed via the compressed suffix tree but with leave extensions pointing in the string. Since the later can be built in a time proportional to the length X, denoted $|X|$, thus computing $C(X)$ is proportional to $|X|$.

The genomes are written in the alphabet $\mathcal{A} = \{A, C, G, T\}$, made of the four nucleo-bases. Although the genomes sequences are not purely random, we will use for comparison, randomly generated sequence over $\mathcal{A}$. Since the bases mostly appear uniformly in each genome, most of the time we will rely on memoryless uniform generation on $\mathcal{A}$, but all the results which will be stated below have been obtained under more general sequence generation models, such that biaised memoryless, Markov with finite memory, mixing models [**?**].

Theorem 1 ([2]): The average complexity of a string X built on a memoryless or a Markov source satisfies:

$$\begin{aligned} E[C(X)] &= \tfrac{(|X|+1)|X|}{2} + |X| - \tfrac{|X|\log|X|}{h} - \left(\tfrac{1}{2} + \tfrac{\gamma}{h}\right)|X| \\ &\quad + O(\log|X|), \end{aligned} \tag{1}$$

where h is the per symbol entropy rate of the source model and γ is the Euler-Mascheroni constant.
When the source model is uniform memoryless on the four bases we have $h = \log 4$. We notice that the string complexity in our models is mostly quadratic, indicating that almost the factor comprised between every pair of position in X is unique.

A. Weak and accidental pattern matching

By weak and accidental pattern matching we mean the joint complexity between two random sequence X and Y independently generated.

Theorem 2 ([3], [6] chapter 10): When X and Y are of same length but generated on two different source models (*e.g.* a Markov transition matrix with different parameters): when $|X| \to \infty$

$$E[J(X,Y)] \sim \frac{|X|^\kappa}{\sqrt{a \log |X| + b}} \tag{2}$$

with $\kappa < 1$, and some parameter a and $b > 0$. When X and Y are of different length but on the source model then when both $|X|$ $|Y|$ tend to infinity

$$E[J(X,Y)] \quad \sim \quad \frac{(|X|+|Y|)\log(|X|+|Y|) - |X|\log|X| - |Y|\log|Y|}{h}. \tag{3}$$

Proof: All the proof are [6] chapter 10 the only new result is the refinement of the result about $E[J(X,Y)$ when X and Y are on the same source model. To simplify we only hint the proof on a memoryless source. We know from [6] that $J(X,Y) \sim C(|X|,|Y|)$ where $C(z_1,z_2)$ is function which satisfies:

$$C(z_1,z_2) = (1-e^{-z_1})(1-e^{-z_2}) + \sum_{a\in\mathcal{A}} C(p_a z_1, p_a z_2) \tag{4}$$

with p_a the probability of the occurrence of symbol a in a random sequence. If we denote $f_\lambda(z) = C(z,\lambda z)$ we get the functional equation

$$f_\lambda(z) = (1-e^{-z})(1-e^{-\lambda z}) + \sum_{a\in\mathcal{A}} f_\lambda(p_a z) \tag{5}$$

whose asymptotic is obtained via the Mellin transform as described in [1]. ■

B. Strong pattern matching

We call strong pattern matching when the sequences X and Y are so close that they are just a light alteration of each other.

Theorem 3: Let $k \geq 1$ be a fixed integer, assume X is generated by a memoryless or a Markov source and Y differs via k symbol substitution. We have the estimate when $|X| \to \infty$:

$$E[J(X,Y)] \sim \frac{(|X|+1)(|X|+2-k)}{(k+2)(k+1)}. \tag{6}$$

Notice that when $k = 0$ we don't find back the estimate $E[C(X)]$ since $C(X) = J(X,X)$. In strong pattern matching mode the joint complexity remains quadratic.

Proof: To compute the leading term we look at the factors which don't overlap the positions where the k substitution occurs between X and Y. These factors are common to both X and Y and we know that almost surely they are unique. Thus our analysis rigorously is a lower bound, since we have no room to develop the upper bound proof. Let J_n^k be the cumulated number of such factors considering all the $\binom{n}{k}$ combination of substitued positions between nX and Y. Therefore $E[J(X,Y)] = \frac{J_n^k}{\binom{n}{k}}$. We know that $J_n^0 = \frac{(n+1)n}{2}$. The generating function $J^0(z) = \sum_n J_n^0 z^n = \frac{z}{(1-z)^3}$ for $|z| < 1$. We have the recurrence

$$J_n^k = \sum_{m=0}^{n-k} J_m^0 + J_{n-m-1}^{k-1} \tag{7}$$

which translated in generating function gives $J^k(z) = \frac{z^k}{1-z} J^0(z) + J^{k-1}(z)\frac{z}{1-z}$ which resolves in $J^k(z) = \frac{z^k}{(1-z)^4}\left(\frac{1+z}{(1-z)^{k-1}} - 1 + z\right)$. The asymptotic leading term is contained in $\frac{1+z}{(1-z)^{k+3}}$ which is $\sum_n (n+2-k)\frac{(n+1)n(n-1)\cdots n-k+1}{(k+2)!} z^n$. The coefficient of z^n divided by $\binom{n}{k}$ gives the claimed asymptotic term. ■

III. Accidental pattern matching on genomes

The genome of the Covid totalizes 29,866 bases (first variant 2020). In the figure 1 we show the joint complexity of the SARS-2 Covid genome with "Bat coronavirus HKU2" [8] (discovered in 2007) which has 27,165 bases. The SARS-2 genome is parsed from right to left, and the plot shows the joint complexity between this portion of the genome withh the genome of the bat alpha. In dash we show the average joint complexity between two random genomes obtained via the same uniform memoryless source over the four bases. This last plot is directly obtained via the formula (3). Since the last plot is below the joint complexity with the bat alpha, we can conclude that indeed the SARS-2 Covid and bat alpha are indeed related.

On figure 2 we display the same plot but normalised with formula (3). We add in red the joint complexity with an HIV virus HIV-1 isolate 060SE from Sweeden (1997) [7] (8,732 bases) and see that the genomes are indeed unrelated. In fact we get the surprising result that the plot is below the average between random sequences indicating that that some factors in SARS-2 covid and in HIV are excluding each other.

However in [9] the authors claim to have found 19 short portions of HIV genomes from different sources which appear in the SARS-2 genome. This paper is only a preprint but resulted in lot of noise when it went public. Some found in this insertion the proof that the SARS-2 covid genome would have been forged for malignant. Indeed we have the theorem:

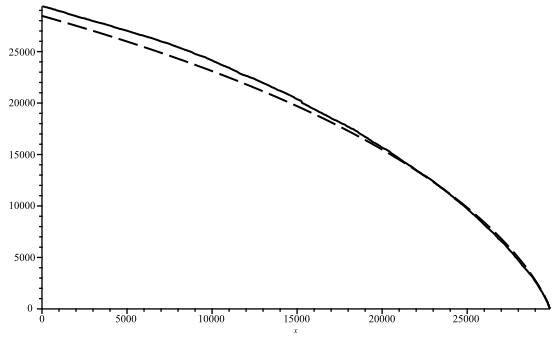

Fig. 1. Joint complexity of SARS-2 genome with bat coronavirus alpha (solid), with random genomes (dashed).

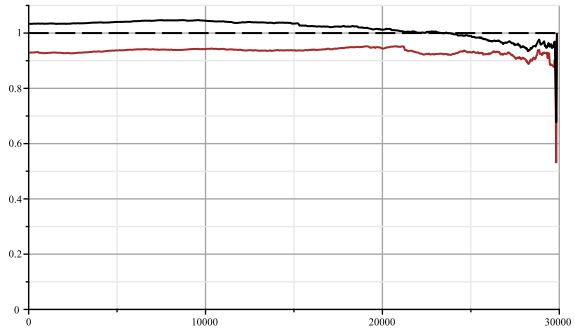

Fig. 2. (Normalized Joint complexity of SARS-2 genome with bat-α, and with HIV genome (red).

Theorem 4 ([6] Chapter 4): Let $\{w_1, w_2, \ldots, w_k\}$ a set of k different sequences. Let X be sequence built on a memoryless source. The probability that the sequence contains all the k factors together is smaller than $|X|^k P(w_1) \cdots P(w_k)$, where the $P(w_i)$'s are the respective probability of occurrence of sequence w_i from the memoryless source.

Since the average putative HIV fragments in SARS-2 genomes depicted in [9] are or length 20 bases or more each, under the archetypal hypothesis that the SARS-2 is from a uniform memoryless source, the probability to have all these 19 copied fragments would be 2.10^{-144}. Thus these accidental insertions would be virtually impossible.

#	genome name	#	genome name
1	HIV2-56-Isolate	11	HIV2-UC1
2	HIV1-060SE-Sweden (R)	12	HIV2-Senegal (R)
3	HIV2-Bissau (R)	13	HIV1-Malawi (R)
4	Simian-VSAA2001 (R)	14	HIV1-Russia (R)
5	HIV1-clone-ML1592 (R)	15	Simian-CM545 (R)
6	HIV2-Verde	16	Simian-KM378564
7	HIV2-106	17	HIV1-EU184986 (R)
8	Simian-TAN5	18	HIV1-AY516986
9	Simian-P18	19	HIV1-HQ217329 (R)
10	HIV1-19828	20	Bat-coronavirus-HKU2

The table above lists the 19 matching genomes. The (R) attribute means that the genome must be reversed in order to get the claimed match.

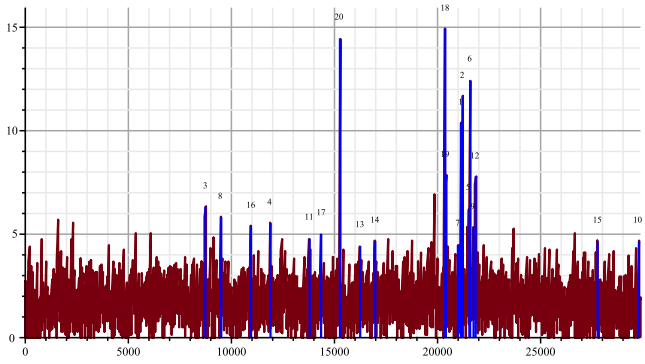

Fig. 3. Joint complexity deviations of SARS-2 genome with the 19 HIV genomes.

Figure 3, shows the repartition of the matching between SARS-2 genome [8] and the 19 HIV genomes (plus the bat coronavirus alpha, which is number 20). The figure has been created the following way. The SARs-2 genome has been cut in

slices of length 24 bases starting every 2 bases. For each HIV candidate genome, we compute its joint complexity with every slices, which give a mean and the variance, then we display for each slice the deviation from the mean in multiple of the standard deviation (it can be negative). This way the accidental matching will be made more apparent. The blue vertical lines are the positions where the maximum deviation appears for each HIV genome. For example for the genome 18 the position of the maximum is 20,400 and is of intensity 15 times the standard deviation, which is very large. The brown plot gives for each slice of SARS-2 genome the maximum of the deviations obtained with the joint complexity over the 20 genomes. Notice that second largest deviation is obtained with "Bat coronavirus HKU2" indicated by index 20. 15 times the standard deviation would mean a probability around $7.2.10^{-100}$ in a pure gaussian context. It should nbe noted that the high peaks correspond to slices with almost exact copy in the other genome, the weaker peaks when there are more mismatches as an illustration of the strong matching theory. The paper [9] lists sequence matching within up to three or four mismatches.

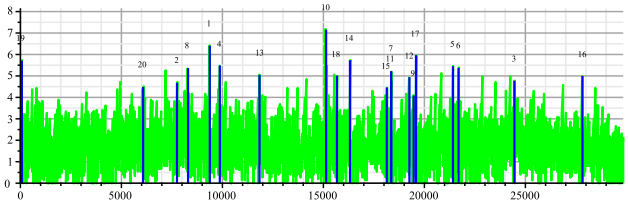

Fig. 4. Joint complexity deviations of Reverse SARS-2 genome with the 19 HIV genomes.

As a matter of comparison, we display in figure 4 the same plot but with the reversed SARS-2 genome. The maxima are way less dramatic. But we notice that all these genomes are coming from very diverse sequences, on HIV-1, other on HIV-2, some on ape origin (the simian IV). Many have been even tested in reversed sequences. We can imagine the author may have tested much more sequences than the 19 selected sequence. Maybe here is the explanation of this paradox. Let M the cumulated number of bases of the tested database. Due to the large sampling of HIV and HIV related sequence in the databases, we can estimate M to the order of half a million bases. The number of positions that can be tested for each match is $M|X|$, X being the SARS-2 genome sequence. If 20 is the size of expected matches, the average number of matches of length 20, is $M|X|4^{-20}$ in the uniform memoryless model. If we include the possibility of up to three errors in the matching, we have to multiply this number by $\binom{20}{3}$. Using Tchebychev inequality we would get

$$P(19 \text{ matches}) \leq \frac{M|X|}{19}\binom{20}{3} \sim 0.8. \tag{8}$$

Under this perspective the matchings are no longer exceptional. But one could argue that the Tchebychev upper bound is very rough and the real probability could be much smaller. But it should be noted that the probability becomes much larger when the data are strongly positively correlated. This is confirmed by figure 5 which display the numerous accidental matching between HIV-2UC1 and the other matcher genomes. Dating of 1993 the DNA editing technology was not existing.

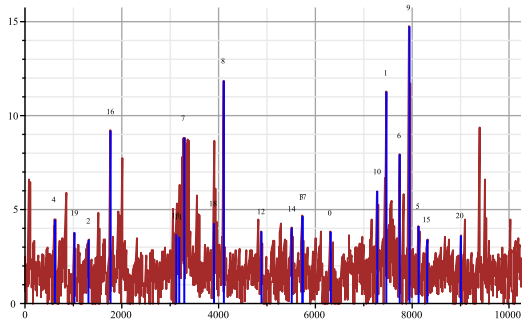

Fig. 5. Joint complexity deviations of HIV-2UC1 genome with other matchers

IV. Strong pattern matching on Covid genomes

In this section we try to analyse the hypotheses of the relation of the covid with its potential ancestors and descendants. The current landscape is summarized in the following figure 6

In short a first putative ancestor is the bat coronavirus "Bat coronavirus HKU2" [10] (let's call it bat-α) which has been discovered in 2007 and has 27,165 bases. The next ancestor is the another bat coronavirus RaTG13 [11], [12] discovered in 2013 and has 29,855 bases (let's call it bat-β). Then the first SARS-2 covid coronavirus for human, discovered late 2019 and another bat coronavirus RaCCS203 [13], discovered early 2020 and has 29,775 bases. Thus the pivot genome is the bat-β

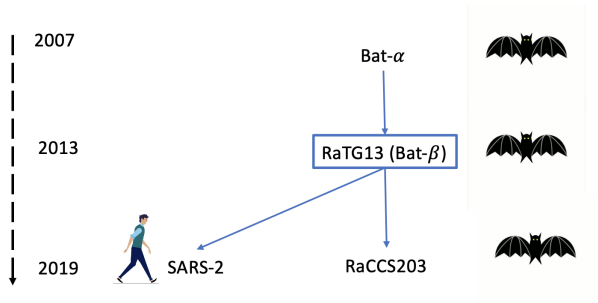

Fig. 6. Putative genealogic tree of SARS-2 Covid.

RaTG13. In the following figure we have sliced the bat-β genome in slices of 50 bases each, and computed the joint complexity with other genomes. Figure 7 shows the bat-β slice joint complexity with the whole genome bat-α. Apparently the ancestor seems too distant to look nothing more than random, we are on a weak pattern matching level indicated by the lower dashed horizontal line determined by the estimate produced in (3). The figure 8 shows the bat-β slice joint complexity with its whole genome. In this case the joint complexity naturally find the position of the slice in the whole genome as its best match and the figure basically shows the complexity of the slice and illustrate the formula of theorem 1 (indicated by the dashed upper line). Between the two lines lies the transition between weak pattern matching and strong pattern matching.

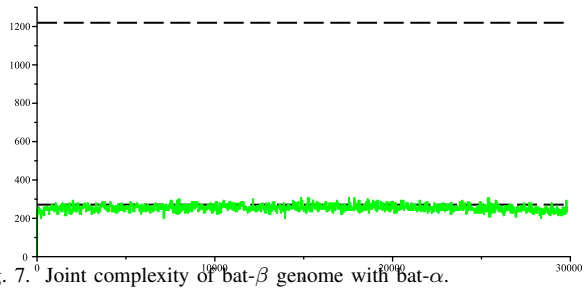

Fig. 7. Joint complexity of bat-β genome with bat-α.

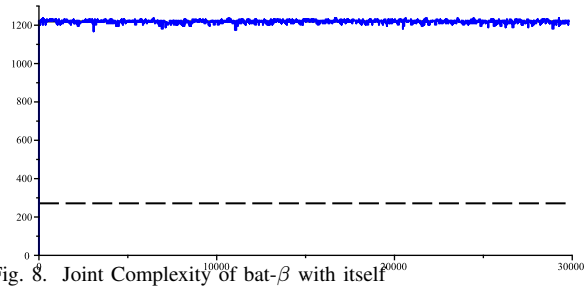

Fig. 8. Joint Complexity of bat-β with itself

Figure 9 shows the bat-β slice joint complexity with its whole SARS-2 covid genome. Surprisingly the slice joint complexity seems to be in strong matching regime (very close to the upper horizontal dashed line), indicating a high degree of similarity. This is unexpectede because there is the same time span between the discovery of bat-α and the discovery of bat-β than there is between the discovery of bat-β and the SARS-2 covid (6 years in both cases). Even more surprising there is even more similarities with SARS-2 than with the genome of the last bat coronavirus RaCCS203, although the latter is for the same specie (bat), and the former is for two different species (human versus bats). Indeed the plot of pattern matching between bat-β and RaCCS203 shows many places where the pattern matching is weak, in particular between the position 21,500 and 24,000 probably indicating the possibility of a large insertion of exogen genetic material.

The three genomes are so close that we can make in correspondence the segments of each genome with the segment in the other genome. Via a straightforward adaptation of the joint complexity program we can compute the offset between the segments in one genome with the segments in the other genome. It consists to spot the largest common factor instead of enumerating the common factors. In term of programming it is just replacing the operator of summation evaluation with the operator of maximum evaluation. Thus for each slice of bat-β we detect the position of its largest match in the other genome. The difference of positions between the two matches in their respective genome is the offset. If the offset is positive then the match is in advance in the first genome compared with the second genome, otherwise when it is negative it is in advance in the second genome.

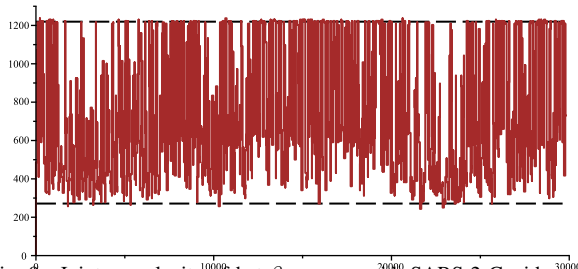

Fig. 9. Joint complexity of bat-β genome with SARS-2 Covid genome.

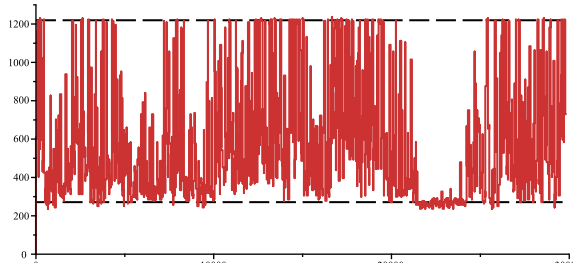

Fig. 10. Joint Complexity of bat-β with with the RaCCS203 genome.

The figure 11 shows the offset per bat-β slice with the SARS-2. Two surprises. First surprise is that except for an extreme minority of slices marked by the three dotted vertical blue lines, the offset is constant and flat and increases from -26 to -16. The offset stability indicates that the mutation sequence between the two genomes are mostly substitution. The three slices which does not fit well are slices where the substituted bases are too numerous and corrupt the largest match to make it jump by a large value, since the correspondence can be anywhere in the genome sequence: in the interval $[-29,855,+29,855]$. For the readability of the figure we have truncated the abscissas. The second surprise is that the offset monotonically increases, indicating that the mutation happened by insertions and never by deletion. That is against the common belief that virus mutations mostly proceed by deletion. Maybe it is the consequence of the inter-specie transfer from bat-β to SARS-2.

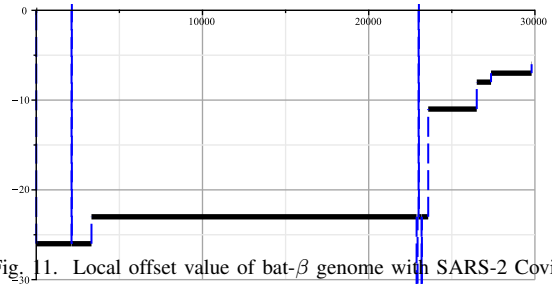

Fig. 11. Local offset value of bat-β genome with SARS-2 Covid genome.

The figure 12 shows the result of the same exercise of offset determination from bat-β genome to the last bat coronavirus RaCCS203. Contrary to the transition between bat-β genome to SARS-2 genome, the offset value is decreasing, sometimes sharply, indicating that the mutation proceeded more by deletion than by insertion, confirming the natural trends in virus evolution. However we notice some insertion small at some positions where the offset value slightly bumps up. We again notice the large corrupted area between position 21,500 and position 24,000. But the offset value drops too after, thus this exogen insertion does not push the material to the right. Figure 13 displays the mismatch rate between each slice of bat-β

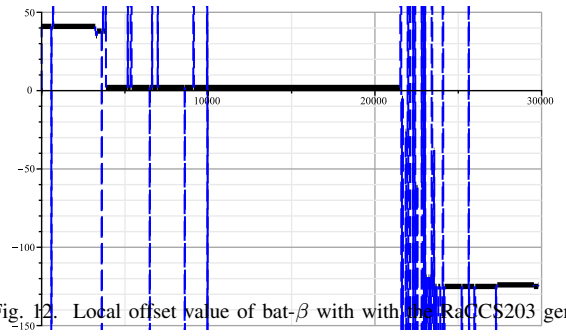

Fig. 12. Local offset value of bat-β with with the RaCCS203 genome.

genome and its corresponding slice in SARS-2 genome in blue. We see the very few strongly corrupted slices with poor

correspondence, elsewhere the substitution rate oscillates between 0 and 10% per 50 base slices. In green we do the same exercise with the RaCCS203 genome. Although in the same lineage, the corruption are much more important. We notice again the area between 21,500 and 24,000 where the Hamming distance is around 65%, ten point below the expected 75% with both portions uniformly and independently generated, but which can be explained by the fact that the largest match should isolate around 5-6 bases.

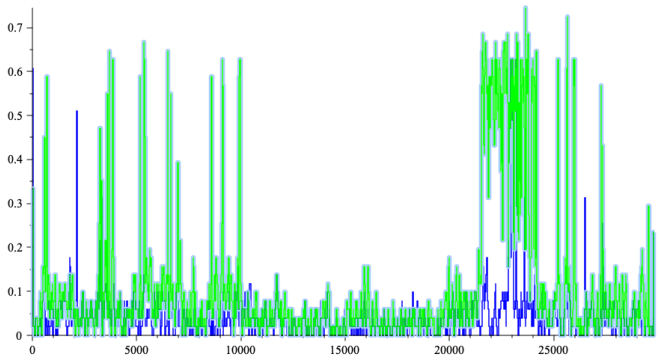

Fig. 13. Mismatch rates between bat-β genome slices and corresponding SARS-2 genome slices (blue) and corresponding RaCCS203 genome slices (green).

V. Conclusion

We have presented an analysis of the covid genome and its possible origins via pure information theoretic tool. Our investigations are not addressing any medical and biogenetic considerations and are mainly based on the pure randomness in the genome mutation process. Therefore they cannot lead to definite answers. Anyhow we can establish that the accidental insertion of HIV segment in the SARS-2 covid is not so exceptional and be easily explained by the abundance of existing materials in the genetic database of HIV. On the other size the strong pattern matching with the putative ancestor bat coronavirus RATG13 is a surprise since the two sequences are more than 6 years apart and two different species. The matching are much weaker with the putative descendant of RATG13, RACS203, despite they are both related to bats.

References

[1] P. Flajolet, Philippe, X. Gourdon, and P. Dumas. "Mellin transforms and asymptotics: Harmonic sums." Theoretical computer science 144.1-2 (1995): 3-58.
[2] S. Janson, S. Lonardi, W. Szpankowski "On the average sequence complexity", Annual Symposium on Combinatorial Pattern Matching, Springer, Berlin, Heidelberg, 2004.
[3] P. Jacquet, Common words between two random strings. In 2007 IEEE International Symposium on Information Theory (pp. 1481-1485).
[4] Jacquet, P., Milioris, D., Szpankowski, W. (2013, July). Classification of Markov sources through joint string complexity: Theory and experiments. In 2013 IEEE International Symposium on Information Theory (pp. 2289-293).
[5] Milioris, D. (2018). Joint Sequence Complexity: Introduction and Theory. In Topic Detection and Classification in Social Networks (pp. 21-56). Springer.
[6] P. Jacquet, Philippe, and W. Szpankowski. Analytic pattern matching: from DNA to Twitter. Cambridge University Press, 2015.
[7] Neogi,U., Siddik,A.B., et al., Recent increased identification and transmission of HIV-1 unique recombinant forms in Sweden, Sci Rep 7 (1), 6371 (2017)
[8] Wu,F., et al., A new coronavirus associated with human respiratory disease in China, Nature 579 (7798), 265-269 (2020)
[9] JC Perez, and L. Montagnier, (2020). COVID-19, SARS and Bats Coronaviruses Genomes Unexpected Exogenous RNA Sequences. OSF Preprints
[10] Lau,S.K., et al. Complete genome sequence of bat coronavirus HKU2 from Chinese horseshoe bats revealed a much smaller spike gene with a different evolutionary lineage from the rest of the genome, Virology 367 (2), 428-439 (2007)
[11] Zhou,P., et al. "A pneumonia outbreak associated with a new coronavirus of probable bat origin", Nature 579 (7798), 270-273 (2020)
[12] Zhou, P., Yang, XL., Wang, XG. et al. Addendum: A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 588, E6 (2020).
[13] Wacharapluesadee,S., et al.. Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia, Nat Commun 12 (1), 972 (2021)