

The Statistical Significance of the Correspondence between Genomic Data and Links with Hotel M

One way that the content of the SARS genomes sequenced to date aligns with the clinical data is as follows. After a multiple alignment of the sequences of 13 of the SARS genomes¹, in four different positions, if you segregate the samples based on which nucleotides are present, the division is the same as when the samples are segregated based on whether the patient had any link to the hotel M or not. This document analyzes the statistical significance of this finding.

This analysis concentrates on point mutations occurring on the 26140 loci in which a nucleotide is present and definitively sequenced in all 13 genomes. In the rest of this document, we will restrict ourselves to these loci without further comment, as if they constituted all of the genomes.

Let us refer to the eight samples that are associated with hotel M as the *M*-**samples**, and the other five as the \overline{M} -**samples** (the non-*M* samples).

The null hypothesis is that nucleotides in all 13 genomes are obtained through mutation from a single consensus sequence. Specifically, the null hypothesis is that each position in each genome is mutated independently at random with a common probability q , and each of the 3 possible nucleotides are equally likely.

The value of q used in the analysis was obtained as follows. We took, for the purpose of calculating q , a consensus sequence obtained by choosing the most commonly occurring nucleotide at each position. Then, we calculated the total, over all genomes, of the number of departures from the consensus: this number was 104. Then, in keeping with the null hypothesis, we set $q = \frac{104}{13 \times 26140}$.

Our first calculation will bound the probability p that there are at least four loci with genetic variation that aligns perfectly with (M, \overline{M}) segregation of the samples. Let us begin by evaluating the probability that a single locus will have this property – let us fix our attention on a single, arbitrary, locus L for this purpose for the moment. The ways that this could happen can be divided into three groups:

- the same mutation could happen in each of the M samples, with no mutations in any of the \overline{M} samples,
- the same mutation could happen in each of the \overline{M} samples, with no mutations in any of the M samples, or
- the same mutation could happen in each of the M samples, and the same mutation could happen in all of the \overline{M} samples.

Each of the first two could be further divided based on which of the non-consensus nucleotides is present in all of mutated samples. The last could similarly be divided into which pair of nucleotides are present in the M and \overline{M} samples respectively. In this way, we can see that the probability that the nucleotides at locus L segregate the samples in the same way as the (M, \overline{M}) distinction is at most

$$3(1-q)^{|\overline{M}|}(q/3)^{|M|} + 3(1-q)^{|M|}(q/3)^{|\overline{M}|} + 6(q/3)^{|\overline{M}|+|M|} = 3(1-q)^5(q/3)^8 + 3(1-q)^8(q/3)^5 + 6(q/3)^{13}. \quad (1)$$

The null hypothesis includes the assertion that what happens at different loci are independent. Thus, we can apply the following large deviation bound.

¹One sequence, BJ04, was excluded from this analysis because it appears to be far from complete.

Lemma 1 (Theorem A.12 of [1]) *Suppose a biased coin with probability r of coming up heads is tossed independently at random m times. Then for any $\beta \geq 1$,*

$$\Pr(\text{number of heads} \geq \beta(rm)) \leq \left(e^{\beta-1}\beta^{-\beta}\right)^{rm}.$$

In this case, we are interested in the probability that at least four loci segregate as the (M, \overline{M}) distinction. Thus, the relevant value of β is

$$\frac{4}{rm} = \frac{4}{(3(1-q)^5(q/3)^8 + 3(1-q)^8(q/3)^5 + 6(q/3)^{13}) \times 26140}.$$

Using the above Lemma with this value of β ,

$$r = 3(1-q)^5(q/3)^8 + 3(1-q)^8(q/3)^5 + 6(q/3)^{13}, \text{ and } m = 26140,$$

we get that the probability p of at least four loci segregating as (M, \overline{M}) satisfies,

$$p \leq 10^{-60}.$$

One potential concern about this analysis is its dependence on q , whose value was estimated from the data. However, the conclusion of statistical significance is not sensitive to the choice of q . If, instead of the estimated value of $\frac{104}{13 \times 26140} \approx 0.0003$, we instead took $q = 0.003$, ten times larger, the same analysis would lead to

$$p \leq 10^{-40}.$$

Another limitation of this analysis is that the null hypothesis includes the assertion that different loci are mutated independently. This can be removed by using Markov's inequality in place of Lemma 1. Markov's inequality states that, for any nonnegative random variable X , $\Pr(X > aE(X)) \leq 1/a$. Since, if the samples at each locus are mutated independently of one another, the above analysis implies that the expected number of loci that segregate as does the (M, \overline{M}) distinction is

$$26140 \times (3(1-q)^5(q/3)^8 + 3(1-q)^8(q/3)^5 + 6(q/3)^{13}).$$

Thus, Markov's inequality implies that, under this weaker null hypothesis,

$$p < \frac{26140 \times (3(1-q)^5(q/3)^8 + 3(1-q)^8(q/3)^5 + 6(q/3)^{13})}{4} \leq 10^{-15}.$$

References

- [1] N. Alon, J. H. Spencer, and P. Erdős. *The Probabilistic Method*. Wiley, 1992.