On the trail of viruses: understanding virus research

Info sheet: sequence analysis

All living organisms, whether prokaryotes or eukaryotes, have one thing in common: their nucleic acids. All consist of the same nucleotides. This commonality suggests that there is a common ancestor. How closely one organism is related to another, can be determined by comparing individual DNA segments or by comparing entire genomes. For this to be possible at all, the DNA section must be known, i.e. its exact DNA sequence. This is possible by sequencing either individual DNA segments or entire genomes. Sequencing is a molecular biology method, in which the sequence of a DNA section is determined with the help of specially prepared nucleotides. The sequence stands for the order of the nucleotides in the DNA and the analysis describes the process of comparison.[1]

Why is sequence analysis an important method?

Sequence analyses are required in almost all biosciences, e.g. in the analysis of spike proteins of the SARS-CoV-2 virus to determine possible targets for antibodies, in vaccine development, or to identify mutations of the virus. Sequencing is also a way of making predictions about an altered protein in the case of an altered gene sequence. A gene is a defined section of DNA, that is translated into a protein (amino acid sequence) during protein biosynthesis.[2]

Methodology: Comparison of the base sequence, creation of a family tree

Once the sequence has been determined, a phylogenetic tree can be created by comparing the same segment from different organisms. Figure 1 shows a phylogenetic tree of the SARS-CoV-2 viruses.[3]

A family tree always has a defined beginning. In the case of Figure 1, this is the first SARS-CoV-2 virus strain ***B1***, which is positioned on the far left. The further away a virus is from the origin (from the base of the phylogenetic tree), the more different DNA bases it has compared to the original sequence ***B1***. This fact also coincides with the appearance of the virus mutants. The ***Delta*** mutant was described before the ***Alpha*** mutant and later ***Omikron.***

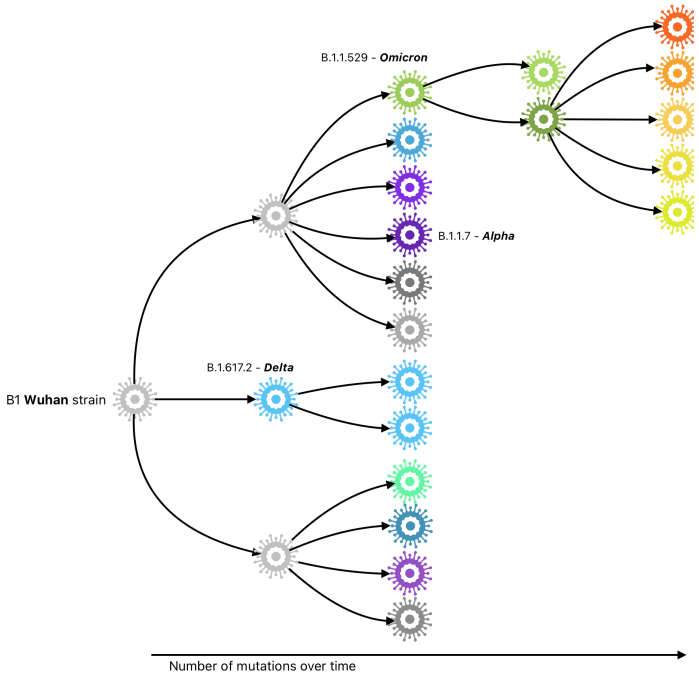

Figure 1: Family tree of SARS-CoV-2 virus strains. The first known strain (**B1** from Wuhan) is shown on the far left. It represents the starting point. The further you move to the right, the more mutation there are. Same colours stand for the same mutation. The strains **Delta, Alpha** and **Omikron** are highlighted in the family tree as they have appeared most frequently during the course of the pandemic.

Image courtesy of the author

Sequence synchronisation is carried out as follows. The original sequence is defined by the user as the standard. The computer now attempts to arrange the sequences to be compared one after the other, in such a way that there is the best possible match between the individual bases of the sequences. Figure 2 shows a sequence comparison of spike nucleotide sequences of nine different SARS-CoV-2 virus variants.

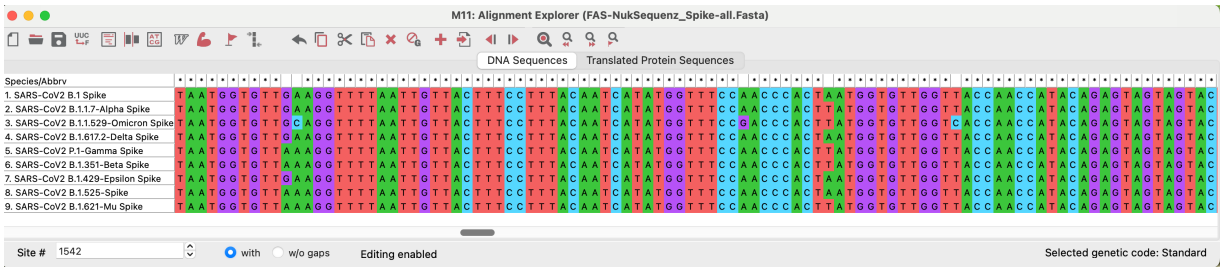

Figure 2: Sequence analysis of nine different SARS-CoV-2 nucleotide sequences of the spike protein. The four nucleotides (adenine A, thymine T, guanine G and cytosine C) are each shown in a specific colour.

Image courtesy of the author

The top sequence is the original sequence. If you look at the individual columns, you can see that there are similarities (same bases) and differences (different bases). Differences may indicate mutations. During sequence analysis, three bases are always combined into one codon and considered as a unit. A codon, i.e. three bases, is translated into one amino acid by the cell during translation. An overview of all possible codons and the resulting amino acids can be found at the end of the document.[4]

Types of mutations and their influence on selection pressure

But how do these differences arise? All organisms are constantly under selection pressure. This is the effect of a selection factor, i.e. an environmental factor, that has an influence on the survival of an organism in a certain environment. Let's take the SARS-CoV-2 viruses again. The selection factor that could have led to a change in the genome was, for example, the vaccine. After immunisation, the human immune system can recognise the virus and actively fight it. Random changes in the virus genome gave rise to new virus variants, that possessed, for example, different spike proteins. The random changes in the genome are known as mutations.[5]

Mutations can occur at genome, chromosome or gene level and have different triggers such as radiation (UV or radioactive), chemicals, errors in replication or translation. Gene mutations occur most frequently, as these often have the least effect. At the gene level, only individual base pairs (nucleotides) are usually altered. These are also known as point mutations. In the case of point mutations, a distinction is made between substitutions, insertions and deletions. They are specified in a certain format. G102Y is an example of this format. Here, the first letter indicates the original amino acid (here the amino acid glycine), the number stands for the position within the gene and the last letter indicates the changed amino acid (here tyrosine). In a substitution, individual bases are exchanged for others, whereas in an insertion, additional bases are added. In a deletion, bases are removed. If the coding region of a gene is altered by insertions or deletions of a number of base pairs that is not a multiple of three a frameshift occurs. A frameshift mutation is very likely to result in a protein with altered activity or an inactive protein.[6]

The individual point mutations have different effects on the gene product and thus ultimately on the phenotypes. Here, too, a distinction is made between three types:

1. The silent or neutral mutation. This occurs, for example, in intron regions of a gene or at the third position of the DNA triplet. Mutations in intron areas very rarely have an effect, as these areas have seldom a relevance for the gene. Due to the redundancy of the genetic code (several DNA triplets code for one and the same amino acid), the third base of a codon can mutate without having an effect on the function of the gene. ACA, ACG, ACC and ACU, for example, all code for the amino acid threonine.
2. If the mutation occurs at the first or second position of a coding triplet, another (incorrect) amino acid is incorporated, with a few exceptions. This is then a missense mutation. These mutations can lead to a change in the gene, which can have a positive or negative effect. In relation to the spike protein, for example, the missense mutation can lead to better or worse binding.
3. If a mutation changes the triplet to a stop codon, this usually has drastic consequences for the gene product. In this case, we speak of a nonsense mutation. Nonsense mutations very often lead to the translation of the mRNA being terminated earlier than actually intended. The result is a shorter amino acid chain. This then usually leads to a non-functional protein.[6]

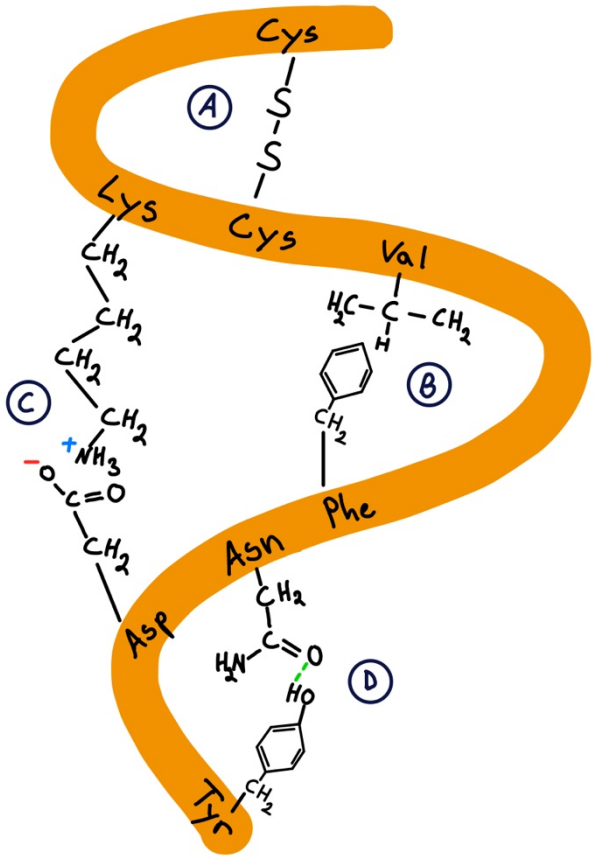

Figure 3: Various interactions between amino acid side chains (disulphide bridges Ⓐ, Van-der-Waals forces Ⓑ, ionic bonds Ⓒ, and hydrogen bonds Ⓓ)

Image courtesy of the author

Whether a mutation has an effect depends to a large extent on the change in the amino acid sequence, i.e. the primary structure. The side chains of the amino acids play a major role here. They are responsible for the intermolecular bonds that ultimately determine the structure of the protein (see figure 3). A distinction is made between real bonds (disulphide bonds Ⓐ and ionic bonds Ⓒ) and molecular forces (hydrogen bonds Ⓓ and van der Waals forces Ⓑ).[7]

Mutations are therefore also seen as a motor of evolution, as they can cause a change in the organism. This change can then be an advantage and increase the fitness of the organism, or it can be a disadvantage and worsen the survival of the organism.

Mutations can be used to distinguish between different organisms of a species by comparing the sequences. The more differences there are, the more distantly the organisms are related to each other (figure 1). Sequence comparisons are also carried out across species boundaries, for example to determine whether there were common ancestors in the past or whether a relationship exists at all.

Furthermore, sequence comparisons can be used, for example, to make predictions about virulence (strength of pathogenicity). This is particularly important for viruses and bacteria. This is because it can be used to deduce how quickly a virus or bacterium can spread. Sequence comparisons can also provide information on the possible affinity of an antibody to its antigen (spike protein). This in turn says something about the effectiveness of a vaccine.[5]

References

[1] Heather JM, Chain B (2016) The sequence of sequencers: The history of sequencing DNA. *Genomics* **107**: 1–8. doi: 10.1016/j.ygeno.2015.11.003

[2] Harvey WT et al. (2021) SARS-CoV-2 variants, spike mutations and immune escape. *Nature Reviews Microbiology* **19**: 409–424. doi: 10.1038/s41579-021-00573-0

[3] Hadfield J et al. (2018) Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* **34**: 4121–4123. doi: 10.1093/bioinformatics/bty407

[4] Alberts B et al. (2002) *Molecular Biology of the Cell* 4th edition. Garland Science. ISBN: 0-8153-3218-1

[5] Carabelli AM et al. (2023) SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nature Reviews Microbiology* **21**: 162–177. doi: 10.1038/s41579-022-00841-7

[6] Lodish H et al. (2016) Chapter 5: Molecular Genetic Techniques. *Molecular Cell Biology* 8th edition pp 171–216. W.H. Freeman. ISBN: 1-4641-8339-2

[7] Ragupathi A, Mastrogiannis AJ, Rahimi N (2026) *Biochemistry, Tertiary Protein Structure*. StatPearls Publishing.

Acknowledgements

This resource has been sponsored by the Joachim Herz Stiftung and produced by the European XFEL.