Reverse Engineering the source code of the BioNTech/Pfizer SARS-CoV-2 Vaccine

Dec 25 2020 *18 mins read*

Translations: ελληνικά / عربى / 中文 / Deutsch / Español / Français / 2עִבְרִית / עִבְרִית / Hrvatski / Italiano / नेपाली / Polskie / русский / Português. / Markdown for translating

Welcome! In this post, we'll be taking a character-by-character look at the source code of the BioNTech/Pfizer SARS-CoV-2 mRNA vaccine.

> *I want to thank the large cast of people who spent time previewing this article for legibility and correctness. All mistakes remain mine though, but I would love to hear about them quickly at bert@hubertnet.nl or @PowerDNS_Bert*

Now, these words may be somewhat jarring - the vaccine is a liquid that gets injected in your arm. How can we talk about source code?

This is a good question, so let's start off with a small part of the very source code of the BioNTech/Pfizer vaccine, also known as BNT162b2, also known as Tozinameran also known as Comirnaty.

WHO
International Nonproprietary Names Programme

9/2020

Sequence / Séquence / Secuencia

```
GAGAAΨAAAC ΨAGΨAΨΨCΨΨ CΨGGΨCCCCA CAGACΨCAGA GAGAACCCGC   50
CACCAΨGΨΨC GΨGΨΨCCΨGG ΨGCΨGCΨGCC ΨCΨGGΨGΨCC AGCCAGΨGΨG  100
ΨGAACCΨGAC CACCAGAACA CAGCΨGCCΨC CAGCCΨACAC CAACAGCΨΨΨ  150
ACCAGAGGCG ΨGΨACΨACCC CGACAAGGΨG ΨΨCAGAΨCCA GCGΨGCΨGCA  200
CΨCΨACCCAG GACCΨGΨΨCC ΨGCCΨΨΨCΨΨ CAGCAACGΨG ACCΨGGΨΨCC  250
ACGCCAΨCCA CGΨGΨCCGGC ACCAAΨGGCA CCAAGAGAΨΨ CGACAACCCC  300
GΨGCΨGCCCΨ ΨCAACGACGG GGΨGΨACΨΨΨ GCCAGCACCG AGAAGΨCCAA  350
CAΨCAΨCAGA GGCΨGGAΨCΨ ΨCGGCACCAC ACΨGGACAGC AAGACCCAGA  400
GCCΨGCΨGAΨ CGΨGAACAAC GCCACCAACG ΨGGΨCAΨCAA AGΨGΨGCGAG  450
ΨΨCCAGΨΨCΨ GCAACGACCC CΨΨCCΨGGGC GΨCΨACΨACC ACAAGAACAA  500
```

First 500 characters of the BNT162b2 mRNA. Source: World Health Organization

The BNT162b mRNA vaccine has this digital code at its heart. It is 4284 characters long, so it would fit in a bunch of tweets. At the very beginning of the vaccine production process, someone uploaded this code to a DNA printer (yes), which then converted the bytes on disk to actual DNA molecules.

A Codex DNA BioXp 3200 DNA printer

Out of such a machine come tiny amounts of DNA, which after a lot of biological and chemical processing end up as RNA (more about which later) in the vaccine vial. A 30 microgram dose turns out to actually contain 30 micrograms of RNA. In addition, there is a clever lipid (fatty) packaging system that gets the mRNA into our cells.

RNA is the volatile 'working memory' version of DNA. DNA is like the flash drive storage of biology. DNA is very durable, internally redundant and very reliable. But much like computers do not

execute code directly from a flash drive, before something happens, code gets copied to a faster, more versatile yet far more fragile system.

For computers, this is RAM, for biology it is RNA. The resemblance is striking. Unlike flash memory, RAM degrades very quickly unless lovingly tended to. The reason the Pfizer/BioNTech mRNA vaccine must be stored in the deepest of deep freezers is the same: RNA is a fragile flower.

Each RNA character weighs on the order of $0.53 \cdot 10^{-21}$ grams, meaning there are $6 \cdot 10^{16}$ characters in a single 30 microgram vaccine dose. Expressed in bytes, this is around 25 petabytes, although it must be said this consists of around 2000 billion repetitions of the same 4284 characters. The actual informational content of the vaccine is just over a kilobyte. SARS-CoV-2 itself weighs in at around 7.5 kilobytes.

The briefest bit of background

DNA is a digital code. Unlike computers, which use 0 and 1, life uses A, C, G and U/T (the 'nucleotides', 'nucleosides' or 'bases').

In computers we store the 0 and 1 as the presence or absence of a charge, or as a current, as a magnetic transition, or as a voltage, or as a modulation of a signal, or as a change in reflectivity. Or in short, the 0 and 1 are not some kind of abstract concept - they live as electrons and in many other physical embodiments.

In nature, A, C, G and U/T are molecules, stored as chains in DNA (or RNA).

In computers, we group 8 bits into a byte, and the byte is the typical unit of data being processed.

Nature groups 3 nucleotides into a codon, and this codon is the typical unit of processing. A codon contains 6 bits of information (2 bits per DNA character, 3 characters = 6 bits. This means $2^6 = 64$ different codon values).

Pretty digital so far. When in doubt, head to the WHO document with the digital code to see for yourself.

> *Some further reading is available here - this link ('What is life') might help make sense of the rest of this page. Or, if you like video, I have two hours for you.*

So what does that code DO?

The idea of a vaccine is to teach our immune system how to fight a pathogen, without us actually getting ill. Historically this has been done by injecting a weakened or incapacitated (attenuated) virus, plus an 'adjuvant' to scare our immune system into action. This was a decidedly analogue technique involving billions of eggs (or insects). It also required a lot of luck and loads of time. Sometimes a different (unrelated) virus was also used.

An mRNA vaccine achieves the same thing ('educate our immune system') but in a laser like way. And I mean this in both senses - very narrow but also very powerful.

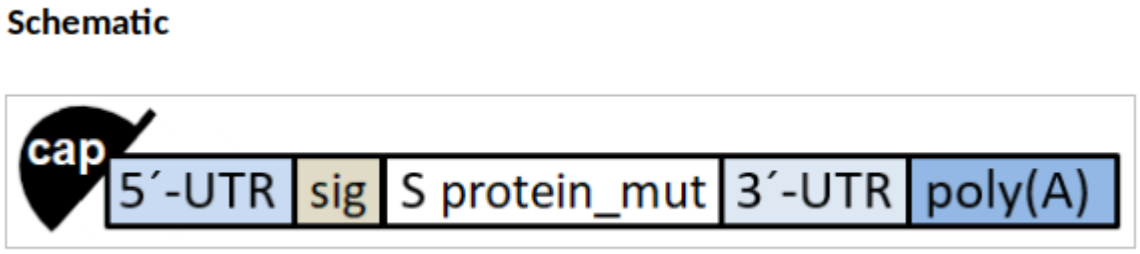
So here is how it works. The injection contains volatile genetic material that describes the famous SARS-CoV-2 'Spike' protein. Through clever chemical means, the vaccine manages to get this genetic material into some of our cells.

These then dutifully start producing SARS-CoV-2 Spike proteins in large enough quantities that our immune system springs into action. Confronted with Spike proteins, and (importantly) tell-tale signs that cells have been taken over, our immune system develops a powerful response against multiple aspects of the Spike protein AND the production process.

And this is what gets us to the 95% efficient vaccine.

The source code!

Let's start at the very beginning, a very good place to start. The WHO document has this helpful picture:



This is a sort of table of contents. We'll start with the 'cap', actually depicted as a little hat.

Much like you can't just plonk opcodes in a file on a computer and run it, the biological operating system requires headers, has linkers and things like calling conventions.

The code of the vaccine starts with the following two nucleotides:

```
GA
```

This can be compared very much to every DOS and Windows executable starting with MZ, or UNIX scripts starting with `#!`. In both life and operating systems, these two characters are not executed in any way. But they have to be there because otherwise nothing happens.

The mRNA 'cap' has a number of functions. For one, it marks code as coming from the nucleus. In our case of course it doesn't, our code comes from a vaccination. But we don't need to tell the cell that. The cap makes our code look legit, which protects it from destruction.

The initial two `GA` nucleotides are also chemically slightly different from the rest of the RNA. In this sense, the `GA` has some out-of-band signaling on it.

The "five-prime untranslated region"

Some lingo here. RNA molecules can only be read in one direction. Confusingly, the part where the reading begins is called the 5' or 'five-prime'. The reading stops at the 3' or three-prime end.

Life consists of proteins (or things made by proteins). And these proteins are described in RNA. When RNA gets converted into proteins, this is called translation.

Here we have the 5' untranslated region ('UTR'), so this bit does not end up in the protein:

```
GAAΨAAACΨAGΨAΨΨCΨΨCΨGGΨCCCCACAGACΨCAGAGAGAACCCGCCACC
```

Here we encounter our first surprise. The normal RNA characters are A, C, G and U. U is also known as 'T' in DNA. But here we find a Ψ, what is going on?

This is one of the exceptionally clever bits about the vaccine. Our body runs a powerful antivirus system ("the original one"). For this reason, cells are extremely unenthusiastic about foreign RNA and try very hard to destroy it before it does anything.

This is somewhat of a problem for our vaccine - it needs to sneak past our immune system. Over many years of experimentation, it was found that if the U in RNA is replaced by a slightly modified molecule, our immune system loses interest. For real.

So in the BioNTech/Pfizer vaccine, every U has been replaced by 1-methyl-3'-pseudouridylyl, denoted by Ψ. The really clever bit is that although this replacement Ψ placates (calms) our immune system, it is accepted as a normal U by relevant parts of the cell.

In computer security we also know this trick - it sometimes is possible to transmit a slightly corrupted version of a message that confuses firewalls and security solutions, but that is still accepted by the backend servers - which can then get hacked.

We are now reaping the benefits of fundamental scientific research performed in the past. The discoverers of this Ψ technique had to fight to get their work funded and then accepted. We should all be very grateful, and I am sure the Nobel prizes will arrive in due course.

> *Many people have asked, could viruses also use the Ψ technique to beat our immune systems? In short, this is extremely unlikely. Life simply does not have the machinery to build 1-methyl-3'-pseudouridylyl nucleotides. Viruses rely on the machinery of life to reproduce themselves, and this facility is simply not there. The mRNA vaccines quickly degrade in the human body, and there is no possibility of the Ψ-modified RNA replicating with the Ψ still in there. "No, Really, mRNA Vaccines Are Not Going To Affect Your DNA" is also a good read.*

Ok, back to the 5' UTR. What do these 51 characters do? As everything in nature, almost nothing has one clear function.

When our cells need to *translate* RNA into proteins, this is done using a machine called the ribosome. The ribosome is like a 3D printer for proteins. It ingests a strand of RNA and based on that it emits a string of amino acids, which then fold into a protein.

0:00 / 0:30

Source: Wikipedia user Bensaccount

This is what we see happening above. The black ribbon at the bottom is RNA. The ribbon appearing in the green bit is the protein being formed. The things flying in and out are amino acids plus adaptors to make them fit on RNA.

This ribosome needs to physically sit on the RNA strand for it to get to work. Once seated, it can start forming proteins based on further RNA it ingests. From this, you can imagine that it can't yet read the parts where it lands on first. This is just one of the functions of the UTR: the ribosome landing zone. The UTR provides 'lead-in'.

In addition to this, the UTR also contains metadata: when should translation happen? And how much? For the vaccine, they took the most 'right now' UTR they could find, taken from the alpha globin gene. This gene is known to robustly produce a lot of proteins. In previous years, scientists

had already found ways to optimize this UTR even further (according to the WHO document), so this is not quite the alpha globin UTR. It is better.

The S glycoprotein signal peptide

As noted, the goal of the vaccine is to get the cell to produce copious amounts of the Spike protein of SARS-CoV-2. Up to this point, we have mostly encountered metadata and "calling convention" stuff in the vaccine source code. But now we enter the actual viral protein territory.

We still have one layer of metadata to go however. Once the ribosome (from the splendid animation above) has made a protein, that protein still needs to go somewhere. This is encoded in the "S glycoprotein signal peptide (extended leader sequence)".

The way to see this is that at the beginning of the protein there is a sort of address label - encoded as part of the protein itself. In this specific case, the signal peptide says that this protein should exit the cell via the "endoplasmic reticulum". Even Star Trek lingo is not as fancy as this!

The "signal peptide" is not very long, but when we look at the code, there are differences between the viral and vaccine RNA:

(Note that for comparison purposes, I have replaced the fancy modified Ψ by a regular RNA U)

```
            3   3   3   3   3   3   3   3   3   3   3   3   3   3   3   3
Virus:   AUG UUU GUU UUU CUU GUU UUA UUG CCA CUA GUC UCU AGU CAG UGU GUU
Vaccine: AUG UUC GUG UUC CUG GUG CUG CUG CCU CUG GUG UCC AGC CAG UGU GUU
               !   !   !   !   ! ! ! !     !   !   !   !   !
```

So what is going on? I have not accidentally listed the RNA in groups of 3 letters. Three RNA characters make up a codon. And every codon encodes for a specific amino acid. The signal peptide in the vaccine consists of *exactly* the same amino acids as in the virus itself.

So how come the RNA is different?

There are 4^3=64 different codons, since there are 4 RNA characters, and there are three of them in a codon. Yet there are only 20 different amino acids. This means that multiple codons encode for the same amino acid.

Life uses the following nearly universal table for mapping RNA codons to amino acids:

1st base	2nd base								3rd base
	U		C		A		G		
U	UUU	(Phe/F) Phenylalanine †	UCU	(Ser/S) Serine †	UAU	(Tyr/Y) Tyrosine †	UGU	(Cys/C) Cysteine †	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L) Leucine †	UCA		UAA	Stop (*Ochre*)[note 2]	UGA	Stop (*Opal*)[note 2]	A
	UUG		UCG		UAG	Stop (*Amber*)[note 2]	UGG	(Trp/W) Tryptophan †	G
C	CUU		CCU	(Pro/P) Proline †	CAU	(His/H) Histidine ‡	CGU	(Arg/R) Arginine ‡	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	(Gln/Q) Glutamine †	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	(Ile/I) Isoleucine †	ACU	(Thr/T) Threonine †	AAU	(Asn/N) Asparagine †	AGU	(Ser/S) Serine †	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K) Lysine ‡	AGA	(Arg/R) Arginine ‡	A
	AUG	(Met/M) Methionine †	ACG		AAG		AGG		G
G	GUU	(Val/V) Valine †	GCU	(Ala/A) Alanine †	GAU	(Asp/D) Aspartic acid ↓	GGU	(Gly/G) Glycine †	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GUG		GCG		GAG		GGG		G

The RNA codon table (Wikipedia)

In this table, we can see that the modifications in the vaccine (UUU -> UUC) are all *synonymous*. The vaccine RNA code is different, but the same amino acids and the same protein come out.

If we look closely, we see that the majority of the changes happen in the third codon position, noted with a '3' above. And if we check the universal codon table, we see that this third position indeed often does not matter for which amino acid is produced.

So, the changes are synonymous, but then why are they there? Looking closely, we see that all changes *except one* lead to more C and Gs.

So why would you do that? As noted above, our immune system takes a very dim view of 'exogenous' RNA, RNA code coming from outside the cell. To evade detection, the 'U' in the RNA was already replaced by a Ψ.

However, it turns out that RNA with a higher amount of Gs and Cs is also converted more efficiently into proteins,

And this has been achieved in the vaccine RNA by replacing many characters with Gs and Cs wherever this was possible.

> *I'm slightly fascinated by the one change that did not lead to an additional C or G, the CCA -> CCU modification. If anyone knows the reason, please let me know! Note that*

> *I'm aware that some codons are more common than others in the human genome, but I also read that this does not influence translation speed a lot.*

The actual Spike protein

The next 3777 characters of the vaccine RNA are similarly 'codon optimized' to add a lot of C's and G's. In the interest of space I won't list all the code here, but we are going to zoom in on one exceptionally special bit. This is the bit that makes it work, the part that will actually help us return to life as normal:

```
                 *   *
          L   D   K   V   E   A   E   V   Q   I   D   R   L   I   T   G
Virus:   CUU GAC AAA GUU GAG GCU GAA GUG CAA AUU GAU AGG UUG AUC ACA GGC
Vaccine: CUG GAC CCU CCU GAG GCC GAG GUG CAG AUC GAC AGA CUG AUC ACA GGC
          L   D   P   P   E   A   E   V   Q   I   D   R   L   I   T   G
           !     !!! !!       !   !       !   !   !   ! !
```

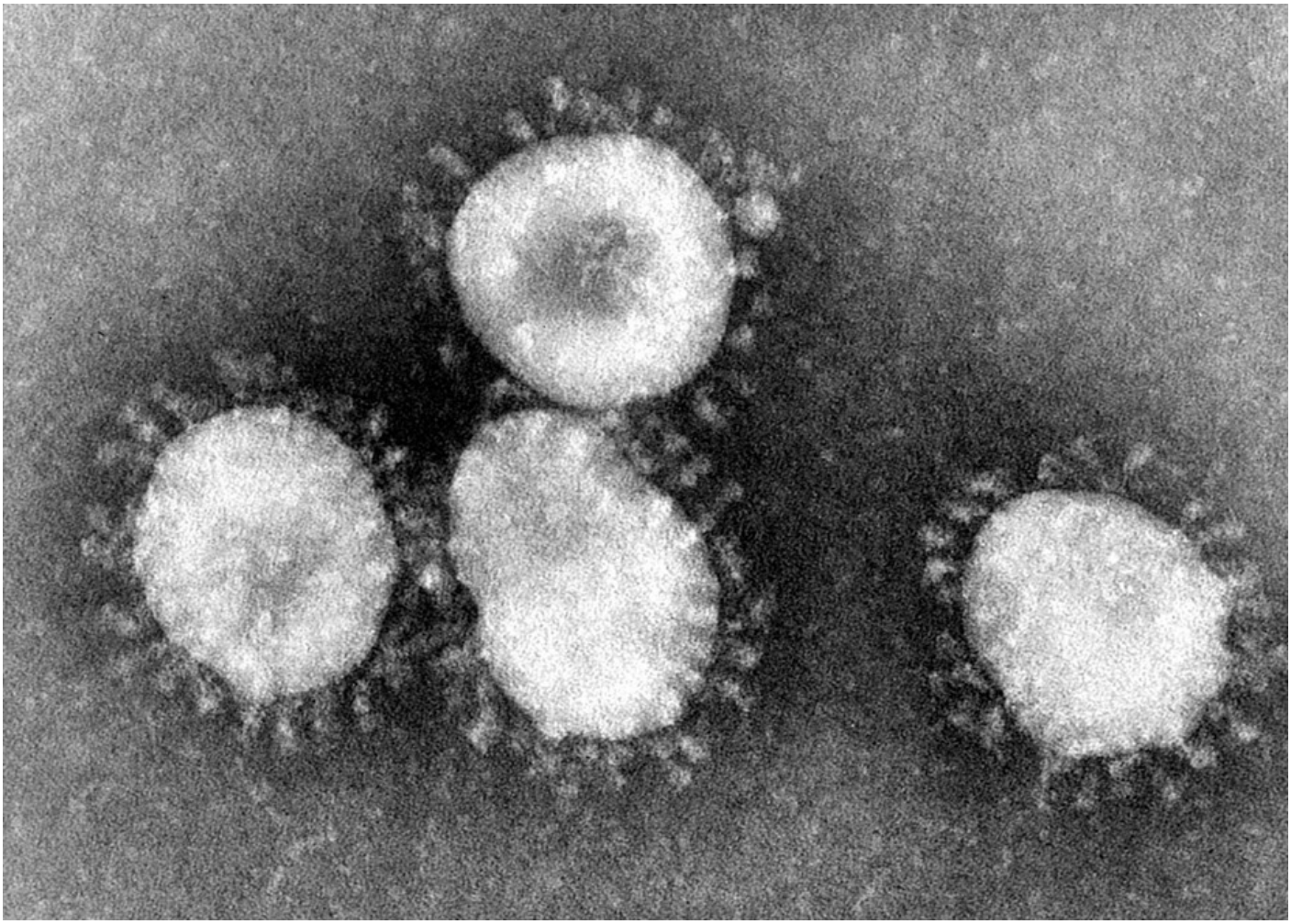
Here we see the usual synonymous RNA changes. For example, in the first codon we see that CUU is changed into CUG. This adds another 'G' to the vaccine, which we know helps enhance protein production. Both CUU and CUG encode for the amino acid 'L' or Leucine, so nothing changed in the protein.

When we compare the entire Spike protein in the vaccine, all changes are synonymous like this.. except for two, and this is what we see here.

The third and fourth codons above represent actual changes. The K and V amino acids there are both replaced by 'P' or Proline. For 'K' this required three changes ('!!!') and for 'V' it required only two ('!!').

It turns out that these two changes enhance the vaccine efficiency enormously.

So what is happening here? If you look at a real SARS-CoV-2 particle, you can see the Spike protein as, well, a bunch of spikes:



SARS virus particles (Wikipedia)

The spikes are mounted on the virus body ('the nucleocapsid protein'). But the thing is, our vaccine is only generating the spikes itself, and we're not mounting them on any kind of virus body.

It turns out that, unmodified, freestanding Spike proteins collapse into a different structure. If injected as a vaccine, this would indeed cause our bodies to develop immunity.. but only against the collapsed spike protein.

And the real SARS-CoV-2 shows up with the spiky Spike. The vaccine would not work very well in that case.

So what to do? In 2017 it was described how putting a double Proline substitution in just the right place would make the SARS-CoV-1 and MERS S proteins take up their 'pre-fusion' configuration, even without being part of the whole virus. This works because Proline is a very rigid amino acid. It acts as a kind of splint, stabilising the protein in the state we need to show to the immune system.

The people that discovered this should be walking around high-fiving themselves incessantly. Unbearable amounts of smugness should be emanating from them. And it would all be well

deserved.

> *Update! I have been contacted by the McLellan lab, one of the groups behind the Proline discovery. They tell me the high-fiving is subdued because of the ongoing pandemic, but they are pleased to have contributed to the vaccines. They also stress the importance of many other groups, workers and volunteers.*

The end of the protein, next steps

If we scroll through the rest of the source code, we encounter some small modifications at the end of the Spike protein:

```
          V   L   K   G   V   K   L   H   Y   T   s
Virus:   GUG CUC AAA GGA GUC AAA UUA CAU UAC ACA UAA
Vaccine: GUG CUG AAG GGC GUG AAA CUG CAC UAC ACA UGA UGA
          V   L   K   G   V   K   L   H   Y   T   s   s
              !   !   !   !     ! !   !           !
```

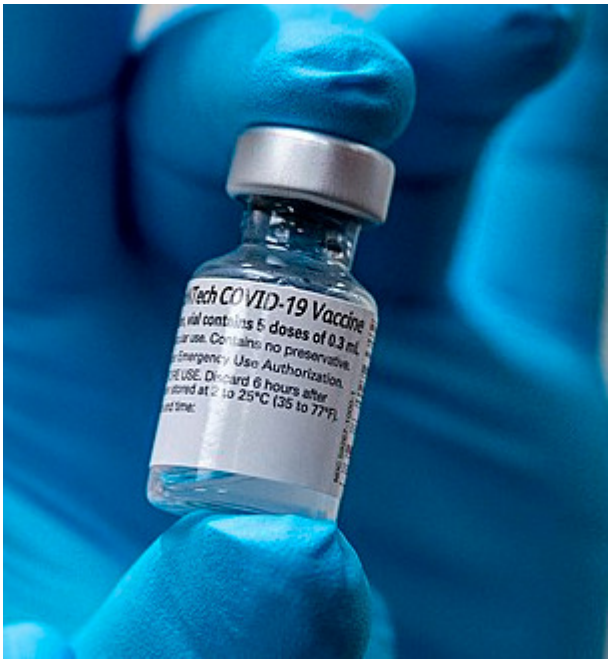
At the end of a protein we find a 'stop' codon, denoted here by a lowercase 's'. This is a polite way of saying that the protein should end here. The original virus uses the UAA stop codon, the vaccine uses two UGA stop codons, perhaps just for good measure.

The 3' Untranslated Region

Much like the ribosome needed some lead-in at the 5' end, where we found the 'five prime untranslated region', at the end of a protein coding region we find a similar construct called the 3' UTR.

Many words could be written about the 3' UTR, but here I quote what the Wikipedia says: "The 3'-untranslated region plays a crucial role in gene expression by influencing the localization, stability, export, and translation efficiency of an mRNA .. **despite our current understanding of 3'-UTRs, they are still relative mysteries**".

What we do know is that certain 3'-UTRs are very successful at promoting protein expression. According to the WHO document, the BioNTech/Pfizer vaccine 3'-UTR was picked from "the amino-terminal enhancer of split (AES) mRNA and the mitochondrial encoded 12S ribosomal RNA to confer RNA stability and high total protein expression". To which I say, well done.



The AAAAAAAAAAAAAAAAAAAAA end of it all

The very end of mRNA is polyadenylated. This is a fancy way of saying it ends on a lot of AAAAAAAAAAAAAAAAAA. Even mRNA has had enough of 2020 it appears.

mRNA can be reused many times, but as this happens, it also loses some of the A's at the end. Once the A's run out, the mRNA is no longer functional and gets discarded. In this way, the 'poly-A' tail is protection from degradation.

Studies have been done to find out what the optimal number of A's at the end is for mRNA vaccines. I read in the open literature that this peaked at 120 or so.

The BNT162b2 vaccine ends with:

```
                                        ****** ****
UAGCAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAGCAUAU GACUAAAAAA AAAAAAAAAA
AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAAAAAAAA AAAA
```

This is 30 A's, then a "10 nucleotide linker" (GCAUAUGACU), followed by another 70 A's.

I suspect that what we see here is the result of further proprietary optimization to enhance protein expression even more.

Summarising

With this, we now know the exact mRNA contents of the BNT162b2 vaccine, and for most parts we understand why they are there:

- The CAP to make sure the RNA looks like regular mRNA
- A known successful and optimized 5' untranslated region (UTR)
- A codon optimized signal peptide to send the Spike protein to the right place (copied 100% from the original virus)
- A codon optimized version of the original spike, with two 'Proline' substitutions to make sure the protein appears in the right form
- A known successful and optimized 3' untranslated region
- A slightly mysterious poly-A tail with an unexplained 'linker' in there

The codon optimization adds a lot of G and C to the mRNA. Meanwhile, using Ψ (1-methyl-3'-pseudouridylyl) instead of U helps evade our immune system, so the mRNA stays around long enough so we can actually help train the immune system.

Further reading/viewing

In 2017 I held a two hour presentation on DNA, which you can view here. Like this page it is aimed at computer people.

In addition, I've been maintaining a page on 'DNA for programmers' since 2001.

You might also enjoy this introduction to our amazing immune system.

Finally, this listing of my blog posts has quite some DNA, SARS-CoV-2 and COVID related material.

← Previous
History of PowerDNS: 2013-2020 (Technology)

Dekonstruktion des Programmcodes des BioNTech/Pfizer SARS-CoV-2 Impfstoffes
Next →