Twitter Thread by Bert Hubert

Bert Hubert
@PowerDNS_Bert

Recently I learned something about DNA that blew my mind, and in this thread, I'll attempt to blow your mind as well. Behold: Chargaff's 2nd Parity Rule for DNA N-Grams.
If you are into cryptography or reverse engineering, you should love this.
Thread:

DNA consists of four different 'bases', A, C, G and T. These bases have specific meaning within our biology. Specifically, within the 'coding part' of a gene, a triplet of bases encodes for an amino acid

Most DNA is stored redundantly, in two connected strands. Wherever there is an A on one strand, you'll find a T on the other one. And similarly for C and G:

T G T C A G T
A C A G T C A

(note how the other strand is upside down - this matters!)

If you take all the DNA of an organism (both strands), you will find equal numbers of A's and T's, as well as equal numbers of C's and G's. This is true by definition.
This is called Chargaff's 1st parity rule.
https://t.co/jD4cMt0PJ0

Strangely enough, this rule also holds per strand! So even if you take away the redundancy, there are 99% equal numbers of A/T and C/G * on each strand *. And we don't really know why.
This is called Chargaff's 2nd parity rule.

Lots of people have advanced theories for why the number of C's and G's should match up, but as yet no slam dunk explanation has been reported. But, hold on, things are about to get even weirder! https://t.co/hO3ybOdPrR

It turns out the rule also holds for N-grams of bases! That is, as long as you both 'complement' and 'reverse' them. So for N=1, %C and %G are equal.

For N=2, this says that percentage of CC (%CC) and %GG are also equal, as are %AG and %CT (complemented AND reversed) etc.

You can compare this to turning a book upside down and reading it back to front, and finding that all three-letter words occur with equal frequency before and after turning over the book.

For DNA triplets like 'AAA', this looks like this. Left in blue is frequency of 'AAA', the right orange bar shows the reverse complement 'TTT'. And so on for all other 31 triplets. The correspondence is stunning:

And here are the tiny tiny differences for each triplet, all smaller than 0.2%. Note that this plot shows data for _all_ known bacterial chromosomes:

So why is this the case? There are lots and lots of theories, but there is no consensus yet. And that is what makes it so super interesting!

At the very core of life hides a mystery, a mystery that is easy to research from a computer. And I hope that one day soon we'll know for sure what is going on!
/ends