Playing With K-Mers

You are helping a biologist find the cure of a genetic disease via a new genetic drug. He explained you a *k-mer* is a DNA sequence with k base pairs, or *bps*. He identified an important segment with 23 bps (a *23-mer*), that he knows has 9 `A-T` bps, and 14 `G-C` bps.

He needs you to run simulations on all possible *23-mers* that have this distribution of base pairs. How many different segments will you have to run simulations on?

First, we calculate all possible orderings of the 23 base pairs, and then we divide the result to account for the 9 repeated `A-T` base pairs, and for the 14 repeated `G-C` base pairs:

$23!/(9! \times 14!) = 817,190$ sequences.

But the problem isn't over. We didn't consider each base pair of the sequence can be in any of two possible orientations. A given sequence of base pairs can give many different DNA sequences, depending on the orientation of the bases in the sequence.

```
--A--G--T--                          --A--C--T--
  |  |  |      isn't the same as       |  |  |
--T--C--A--                          --T--G--A--
```

Since each base pair gives the sequence 2 different possibilities, we have to double the number of possibilities 23 times:

$[23!/(9! \times 14!)] \times 2^{23} \approx 7$ trillion sequences.

And that's for a tiny 23 base pair sequence with a known distribution. The smallest replicable DNA known so far are from the minuscule *Porcine circovirus*, and it has 1,800 base pairs! DNA code and life is a truly amazing thing from the technological point of view. To blow your mind, remember that the human DNA has about 3 billion base pairs, replicated in each of the 3 trillion cells of the human body.

```
TODO: incorporate feedback from
https://www.reddit.com/r/bioinformatics/comments/3wopms/
```