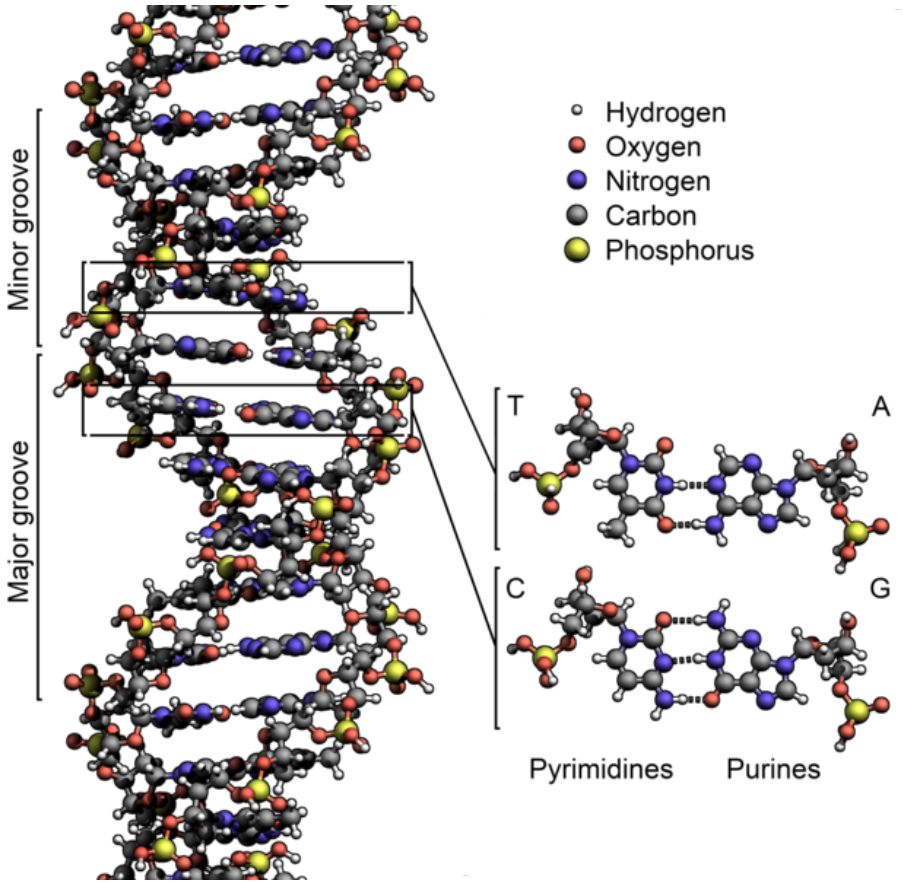
DNA to DATABANK

How do we represent the crazily complex biochemical structure of DNA in the computer?

Below is the complete chemical structure of DNA with all the atoms and bonds for just a small fraction of a typical DNA molecule. Here is the problem: the DNA in just one of your cells is 300,000,000 times longer than this little piece in the picture.

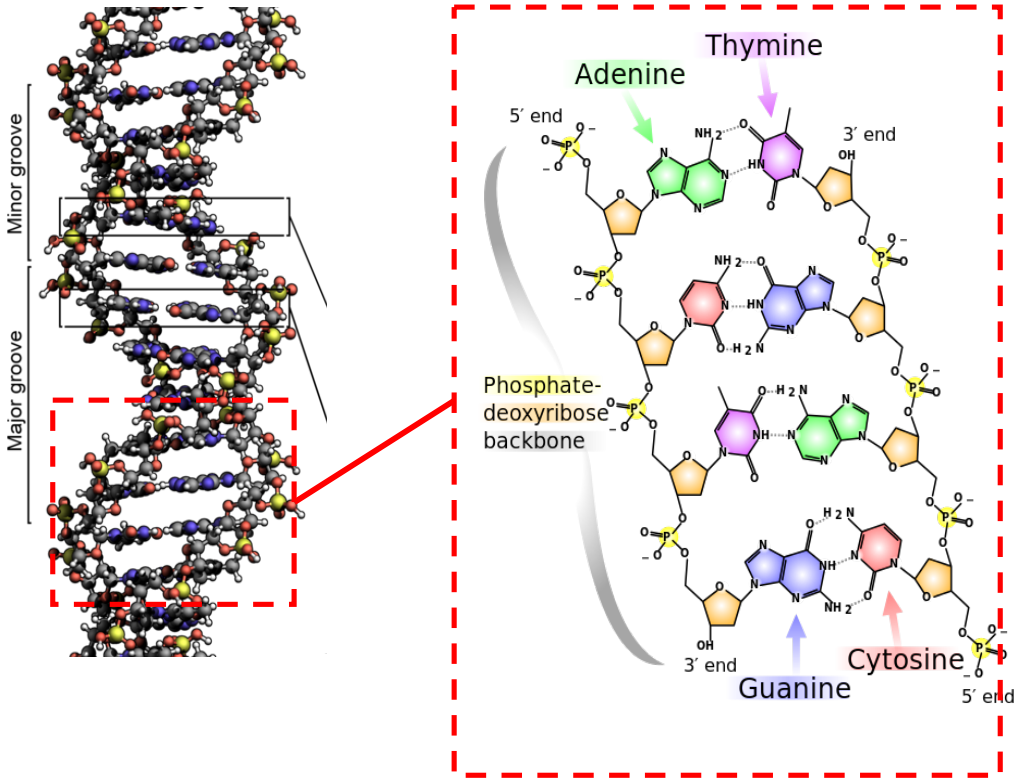


Simplify, Simplify

Saving all this data even in a big computer is not very practical. Clearly, we need an alternative to storing every atom of every DNA molecule.

Let's simplify the structure a bit. If we take a little piece of the structure, flatten it and zoom in, we can see the four molecular parts, the four nucleotides, that make up DNA:

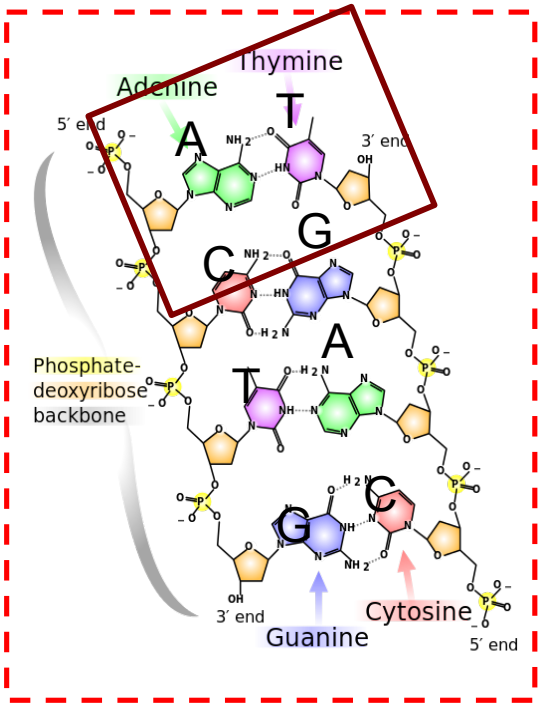
Adenine, Thymine, Guanine, Cytosine



The Bare Minimum

As Watson and Crick figured out, DNA molecules are elegantly simple. DNA is a long string of these four nucleotides base-paired to complementary nucleotides on the opposite strand. The solid box shows the A and T binding to one another making a “base-pair”.

To make it simpler and easier to store in a computer, we can use single letters in place of the nucleotides: A (Adenine), T (Thymine), C (Cytosine), G (Guanine).



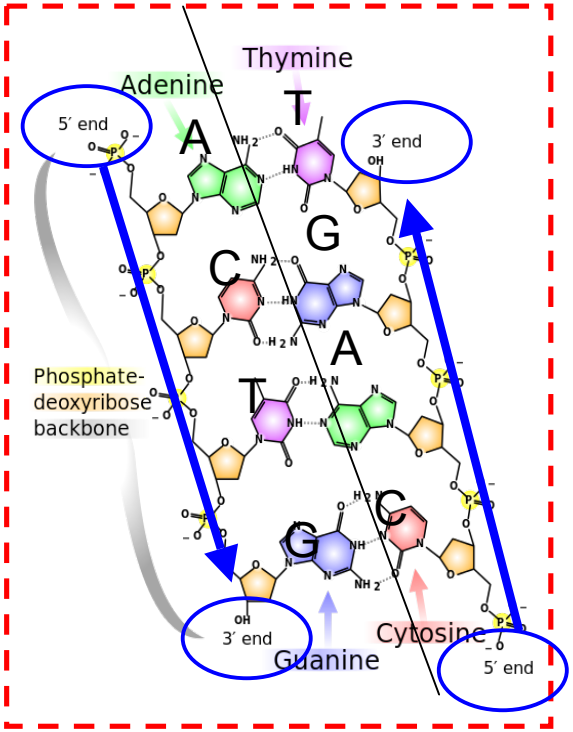
Letters are **perfect for data storage**. In fact, all the letters of all the books *ever written* could fit on a single computer.

The whole human genome is “only” 3 billion letters - that’s just 6 MB of data, The size of most of my family photos, and you can store a lot of photos on a typical laptop.

Simple Text File

Storage is even easier when you realize one must only store half the data. DNA has two parallel strands running opposite directions. The line separate the strands below. The rightmost strand runs from 5' (top) to 3' (bottom), while the leftmost strand runs the opposite direction.

Since Adenine ALWAYS binds Thymine, and Guanine ALWAYS binds Cytosine, if you know one strand you automatically know the other. So why write or store both?



To represent the DNA in the picture we store the nucleotide data as text. The leftmost strand would be: **ACTG**

To store the other strand we would type: **CAGT**

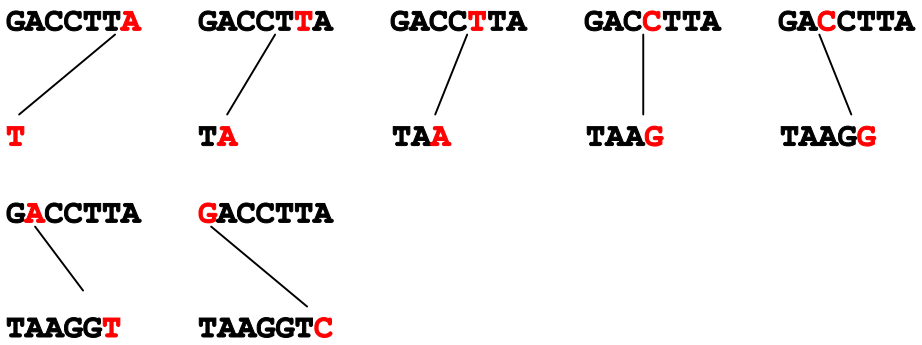
Note: We write the order of nucleotides in the 5' to 3' direction.

Reverse Complement

The DNA nucleotides are always written left to right, from the 5' to 3' direction. This is the "sequence" (the order) in which the nucleotides are written. When Biologists talk about a DNA sequence, this is what they are talking about.

When you have one strand of a sequence, you can determine the other by finding its "reverse complement": Move in reverse , from the end back to the beginning, and determine the matching base for each nucleotide. For instance, you have the DNA sequence: **GACCTTA**

To reverse complement this sequence, go to the last letter, in this case " A", and write the complementary base "T". Move backwards until you reach the beginning.



The reverse complement is: **TAAGGTC**