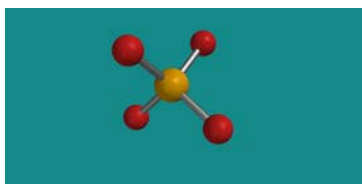


Base pairing in DNA.

Introduction

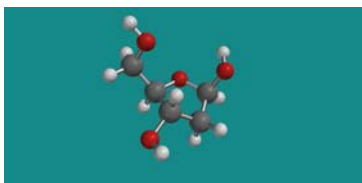
DNA, *deoxyribonucleic acid* are the molecules that contain the genetic information in biological cells. DNA is a so-called biopolymer, with building blocks (monomers) consisting of

a) a phosphate group, PO_4^- (P: orange, O: red)



Phosphate

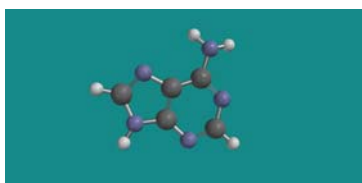
b) a cyclic carbohydrate, i.e., the sugar 2-deoxyribose, $\text{C}_5\text{H}_{10}\text{O}_4$ (C: grey, H: white)



2-deoxyribose

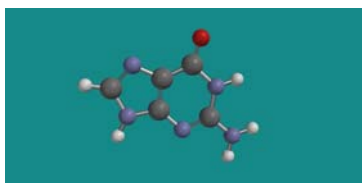
and c) one of four amine bases, either

adenine (A), $\text{C}_5\text{H}_5\text{N}_5$ (N: blue)



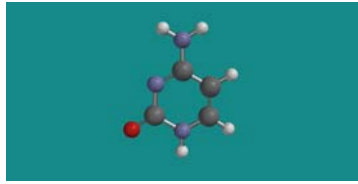
Adenine

guanine (G), $\text{C}_5\text{H}_5\text{N}_5\text{O}$



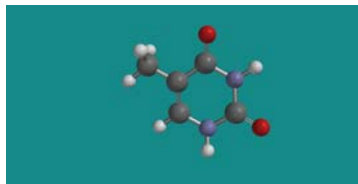
Guanine

cytosine (C), $C_4H_5N_3O$



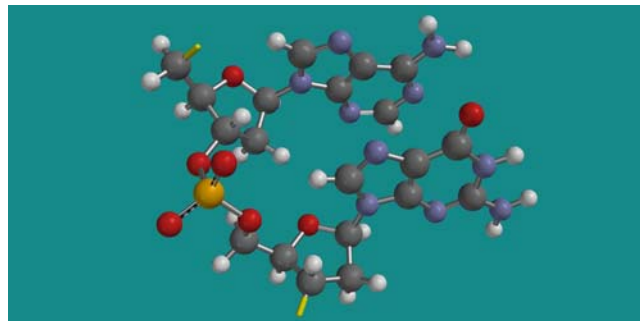
Cytosine

or thymine (T), $C_5H_6N_2O_2$



Thymine

The amine base replaces the hydroxy group (OH) on the carbon atom in position 1 (C1) on the carbohydrate (where we number the C atoms, from 1 for the carbon to the right of the oxygen atom in the five-membered ring, to 5 for the carbon in the CH_2OH group). Next, such “nucleosides” (i.e. amine base + sugar) are linked together into long chains via phosphate groups between C3 in one nucleoside and C5 in another. An example is shown in the figure below, where a nucleoside with the amine base guanine (G, the lower one) is linked to a nucleoside with the amine base adenine (A, the upper one):



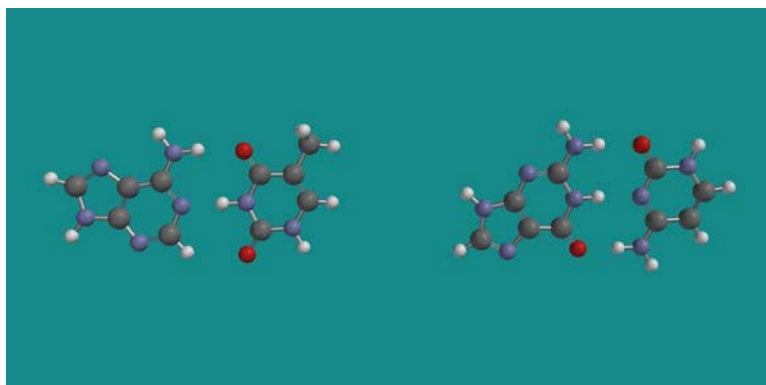
Guanine (bottom) and adenine (top)

The yellow sticks (on C3 of the bottom sugar and C5 on the top sugar) represent open valences where new nucleosides may be attached, all the time with a PO_4 group as the link. The repeating unit {phosphate + sugar + amine base} is called a *nucleotide*. A long chain of such nucleotides is called a *polynucleotide*.

It was discovered early that DNA molecules consist of equal amounts of adenine and thymine, as well as equal amounts of guanine and cytosine, whereas the content of adenine/thymine compared to guanine/cytosine varies from species to species. Human DNA contains about 30% adenine and thymine and about 20% guanine and cytosine.

These observations were key elements when James Watson and Francis Crick in 1953 came up with their model for the DNA structure, namely as a *double helix* with two polynucleotides coiled around each other. The two spirals are held together with so-called *hydrogen bonds* between an A on one spiral and a T on the other, alternatively between a G

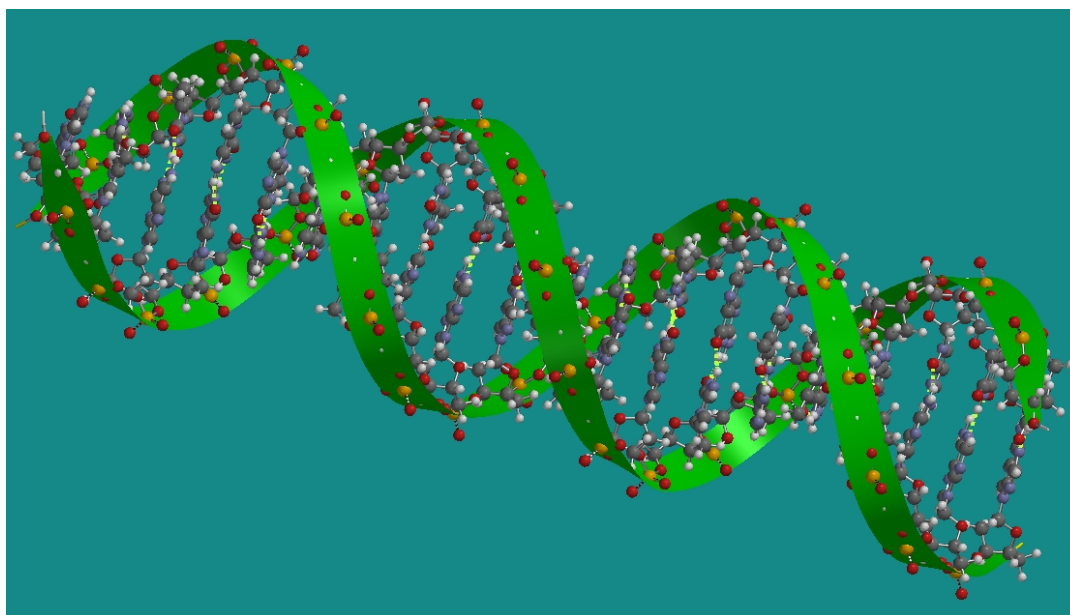
on one spiral and a C on the other. The figure below illustrates these two *Watson – Crick base pairs* A-T (to the left) and G-C (to the right):



Watson – Crick base pair: A – T to the left, G – C to the right

These hydrogen bonds are relatively weak bonds between an H atom on one of the bases and either an O or an N atom on the other. In the base pair A-T, we have two such bonds, one between H on adenine and O on thymine (see above figure, top), the other between N on adenine and H on thymine (“in the middle”). In the base pair G-C, we have three hydrogen bonds, between H on G and O on C (top), between H on G and N on C (middle), and between O on G and H on C (bottom). Such bonds are familiar from water, both in the liquid and in the solid state, where we have weak hydrogen bonds between the O atom in one water molecule and one of the H atoms in another. Without such bonds, water would have much lower melting and boiling points than what it actually has, with dramatic consequences for life as we know it.

The figure below illustrates a DNA double helix with 20 monomers in each of the two polynucleotides:



DNA double helix


The green ribbons pass through the P atoms of the phosphate groups and are visual aids, to make it easier to see the two helices. Yellow dashed lines denote hydrogen bonds between the two bases in each base pair.

In this exercise, you will build and inspect a DNA double helix. You will also take a closer look at a guanine – cytosine (G – C) base pair: The binding energy associated with the three hydrogen bonds will be estimated on the basis of Hartree – Fock calculations on the base pair G – C and the bases G and C separately. The base pair has so many atoms that a Hartree – Fock calculation may take upto half an hour. Therefore, this calculation will be started first.

It is important that the two helices in DNA are not *too* strongly bound together: The precise sequence of bases (for example A-G-G-A-C-C-T-A-G-...) represents the genetic information of the organism, and when the organism grows, this information is *copied* by an opening up of the double helix – by breaking the hydrogen bonds of the base pair – and next, each of the two spirals join another helix to form new double helices. The new double helices become exact copies of the original one, because only A and T, alternatively G and C are able to form new base pairs.

In a living cell you will find 23 pairs of chromosomes, i.e., a total of 46 chromosomes. Each chromosome is a long DNA molecule, consisting of many segments called *genes*. The so called *genome* corresponds to the sum of all the genes in a cell. There are probably somewhere between 20000 and 25000 genes in a human cell, and the complete genome consists of about $3 \cdot 10^9$ base pairs. A single chromosome may be more than 10 cm long.

Exercises

1. Start SPARTAN by choosing  under Programs.
2. Choose the *Nucleotide* builder, press "G" once, and then once in the middle of the screen. The choice of guanine automatically gives you the "complementary" base cytosine, and hence the base pair G – C. Remove atoms until you are left with the G – C base pair as in the figure near the end of page 2. Save the base pair in a directory DNA, for example with filename gua_cyt.spartan. Save the same system with filename gua.spartan and cyt.spartan. Remove atoms in each of these so that you are left with guanine and cytosine, respectively. Start Hartree – Fock calculations for these three systems with Setup – Submit. (SPARTAN uses Hartree – Fock with the 3-21G basis set as default.) Close the molecules with File – Close. Do the next exercises while the jobs are running in the background.
3. With the *Nucleotide* builder, make an arbitrary sequence of bases (A=adenine, G=guanine, T=thymine, C=cytosine) for one of the helices in the DNA molecule. Make a sequence with 20 bases. Left-click on the screen. Reduce the size (SHIFT+hm) of the DNA structure until you have all of it on the screen. Choose *Model – Ribbons* to make the two helices more clearly visible. Also, choose *Model – Hydrogen Bonds* to highlight the hydrogen bonds between the two spirals.

Exercises:

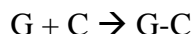
- a) Measure the "period" along the double helix. Simplest is probably to measure the distance between two P atoms. How many P atoms do you have per period in a given helix? (Comment: I believe the built-in model of DNA in SPARTAN is not quite correct – from what I have read, the period should be about 34 Å, with 10 base pairs per period, and not 11 as here.)
- b) Inspect the DNA structure by rotating and scaling the molecule. Switch between Model – Ball and Spoke and Model – Space Filling. Locate both A – T and G – C base pairs and measure the distances O ... H and N ... H of the hydrogen bonds that keep two bases together. Locate the carbohydrate 2-deoxyribose and the phosphate group. Estimate the diameter of the DNA double helix. Estimate the width of the "openings" into the DNA centre from the side of the DNA molecule. (Use Model – Space Filling.) It is believed that both cancer-causing and cancer-preventing agents may affect DNA by entering here.
- c) DNA molecules may be *long*, up to several centimetres. Estimate the mass of a DNA molecule which is 10 cm long. Atomic masses for the constituents can be found in a periodic table. Estimate also the volume of such a DNA molecule. In comparison, a cell typically has a volume somewhere between 10³ and 10⁶ cubic micrometer.

Save the molecule (in the case that you want to inspect it again) as dna and close it.

4. At this stage, the three Hartree – Fock jobs have completed.

Exercises:

- a) Determine the calculated binding energy of the G – C base pair, in other words, the energy gain in the "reaction"



Comment: In reality, the binding energy is much smaller than this, both due to the approximations inherent in the Hartree – Fock method, and because DNA is usually in a solution (water), and not in the gas phase, as anticipated in these calculations. Nevertheless: You see that the binding energy per hydrogen bond, of the order “a few kcal/mol”, is small compared to ordinary chemical bonds, like the ionic bond in HCl and the covalent bond in H₂. (Find these binding energies in the literature, if you do not know them already.)

b) Measure the distances O – N, N – N and N – O (atom in G mentioned first) for the three calculated hydrogen bonds and compare with the experimental values 2.91, 2.95, and 2.86 Å, respectively. (Rosenberg et al, *J. Mol. Biol.*, **104**, 109 (1976).) Comment: The experiment is not on the *exact* same system, so discrepancies are not necessarily due to the Hartree – Fock method.

5. Alternative base pairs?

Build a G-A sequence so that you obtain the base pairs G – C and A – T after one another. Investigate (visually) whether interchange of C and T (resulting in the base pairs G – T and A – C) will give similar good “hydrogen bonding conditions” as in the G – C and A – T base pairs.