DNA MODELLING

CATTOËN Céline

INSA Génie Mathématiques et Modélisation

July - August 2002

CONTENTS

INTRODUCTION

This project consists of modelling DNA molecule to observe twisting and how DNA molecule reacts to applied forces.
The work is separated into two main parts: one deals with searching for documents on the Internet or in the Library, the second deals with coding a program calculating short-range Energy within a molecule.
As this project has been longer than expected, future steps will also be highlighted.

I. Searching for documents.

1. DNA structure:

This chapter is an introduction to DNA structure. DNA is a polymer, the monomer units are nucleotides, and the polymer is known as a "polynucleotide".

- Basic chemical units
  - a five carbon sugar - deoxyribose
  - phosphate – link between sugars
  - bases: purines =adenine and guanine
    pyrimidines = thymine and cytosine

- One strand

Each strand is made up of a sugar covalently linked to a phosphate which is covalently linked to another sugar and so on. A DNA strand may contain thousands to millions of these sugar-phosphate units.
Each sugar also has a purine or pyrimidine base attached to it through a covalent bond.

- Double helix

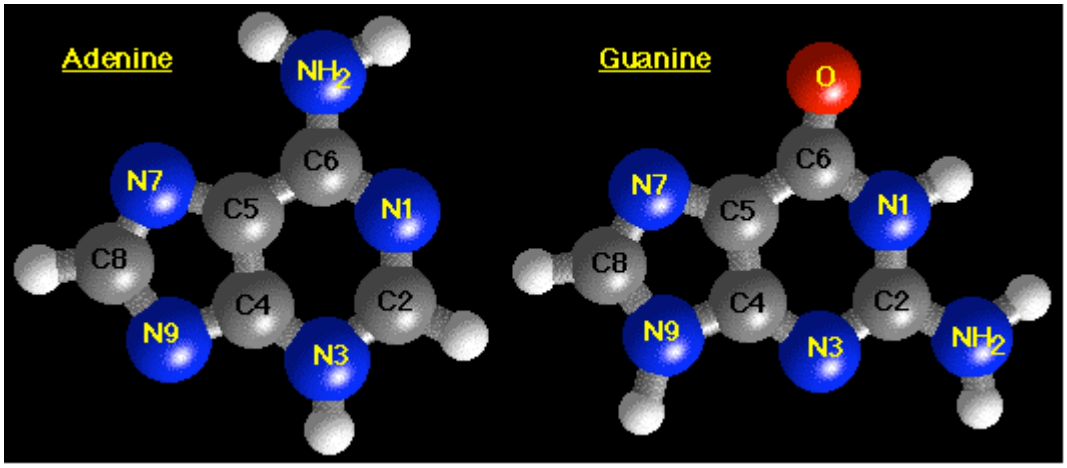
A DNA molecule consists of two strands which are coiled around each other in a double helix. The bases in the opposite strands are arranged such that where there is an adenine in one strand, the other strand has a thymine and where there is a guanine in one strand, the other strand has a cytosine (A-T, C-G).

- Directionality

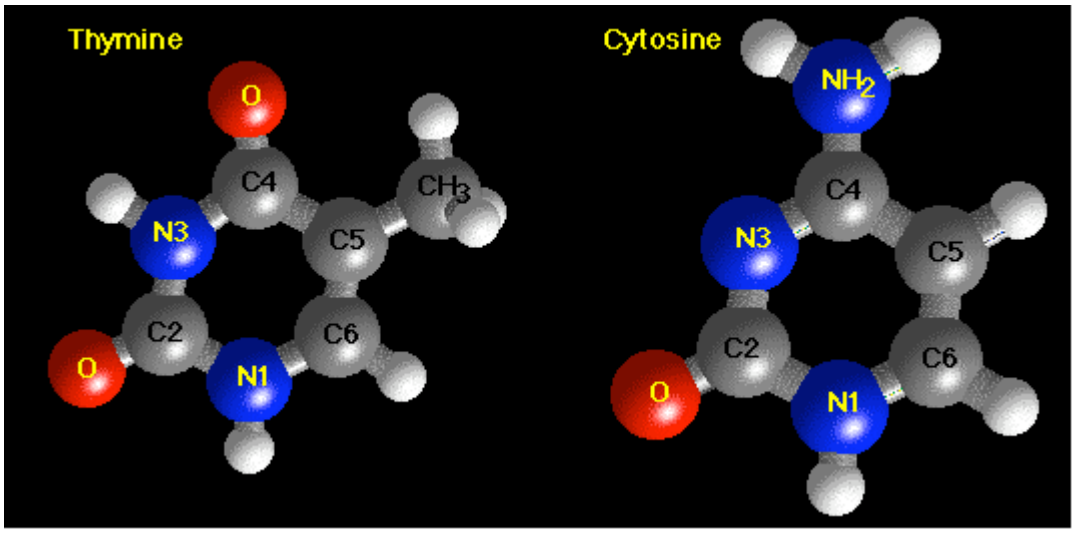
The linkage of the sugar – phosphate "backbone" of a single DNA strand is such that there is a directionality. That is, the phosphate on the 5' carbon of deoxyribose is linked to the 3' carbon of the next deoxyribose. This lends a directionality to a DNA strand which is said to have a 5' to 3' direction. The two strands of a DNA double helix are arranged in opposite directions and are said to be anti-parallel in that one strands is 5' – 3' and the complementary strand is 3' – 5'.

Biochemists use particular notations to describe each atom in DNA molecule, and the following pictures show atom numberings (Hydrogen atoms carry the same numbers as the heavy atoms they are attached to).

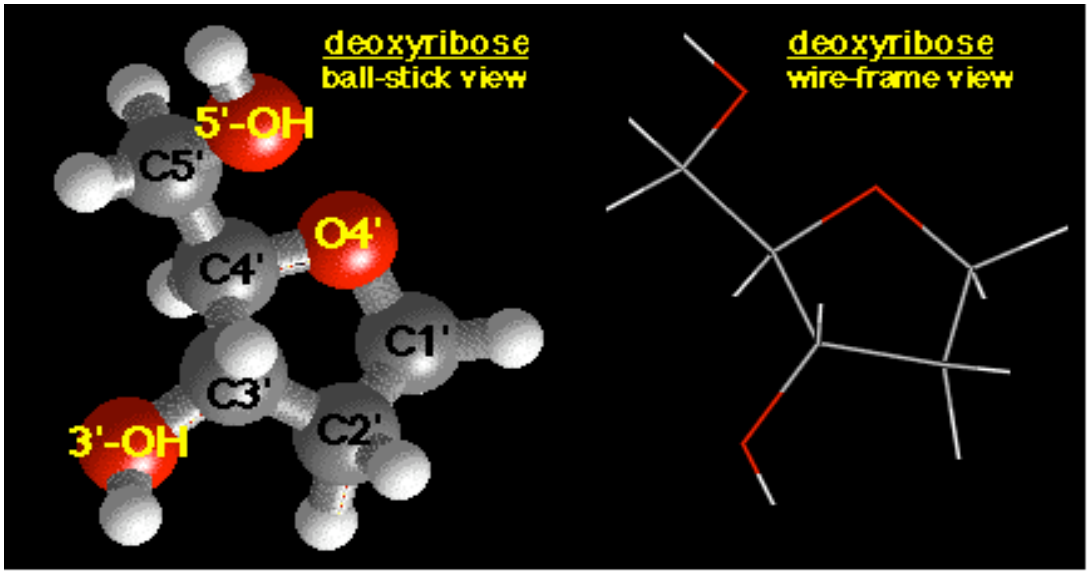
Structure of A and G:



Structure of C and T:

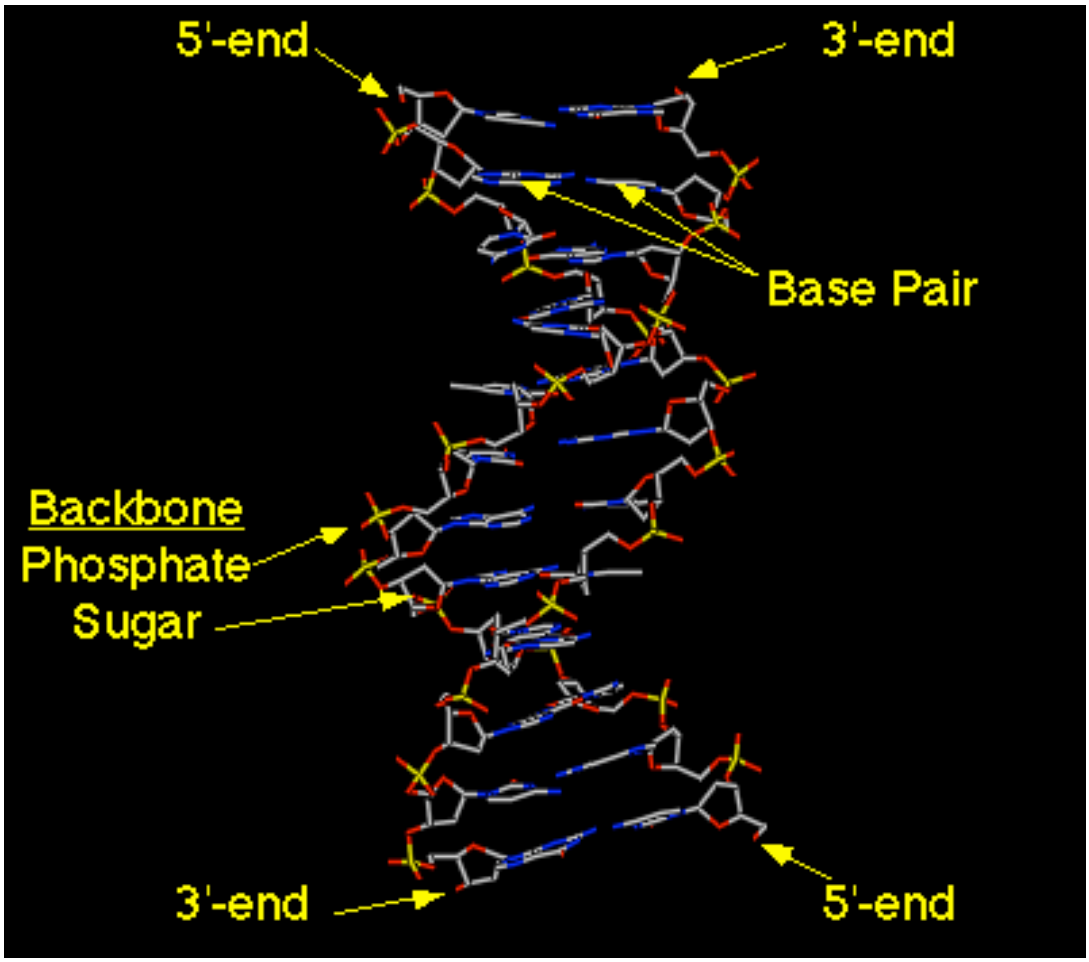


Structure of deoxyribose:



The atoms belonging to the sugar are distinguished from those of the base by the subscript prime mark on the atom number. The numbering scheme for the bases and sugars is the same as that recommended by IUPAC (8).

Nucleotides:



Remark:
In thymine base the CH_3 part has several representations for different pdb files:

- For the 2 main files C is referenced by C5M, H is 1H5M, 2H5M or 3H5M, for it is a Methyl group.
- C can be referenced by C7, H by 1H7,2H7 or 3H7.

Warning:
The writing can be slightly different for H or other atoms depending on the pdb file: instead of 1H5M it can be H5M1. This remark also applies to any other atoms (1H2' can be H2'1, O1P can be OP1).
An internet site shows a list of all Hetero-Atoms naming in PDB files:

- http://daisy.bio.nagoya-u.ac.jp/golab/het_grep.html

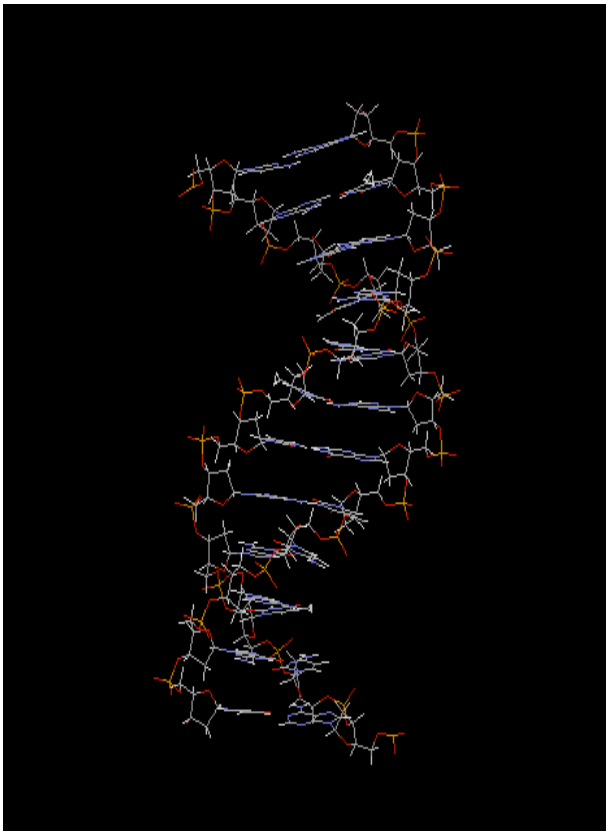
2. Atom coordinates:

The two main files involved in the research have been downloaded on the internet at the following address:
http://www.genevue.com/A_MModel/JACS_2.html

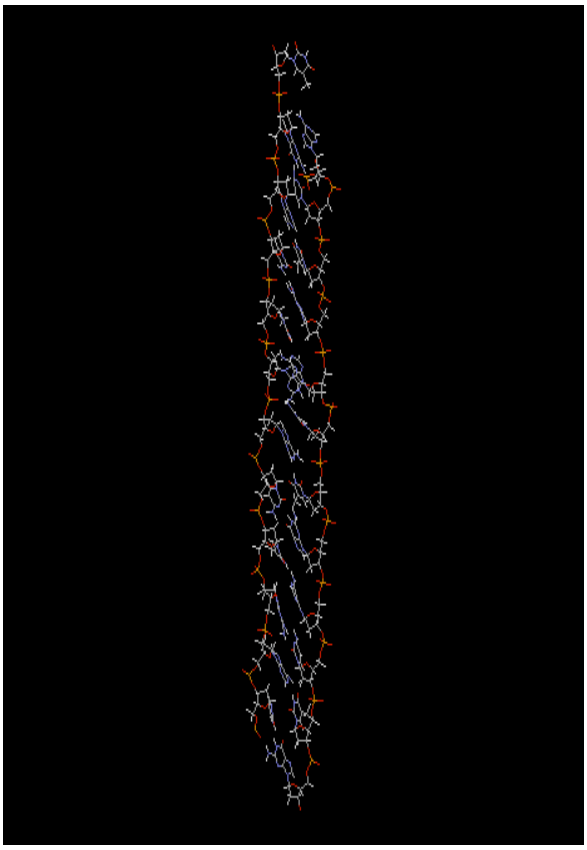
- UNSTRETCHED DNA
- STRETCHED DNA

These two files (actg_000M.pdb and actg_200M.pdb) contain 762 atom coordinates of a DNA molecule with a pdb format and thus are viewable with a software such as RasMol.

The two following figures represent the two coordinate files viewed with RasMol.



UNSTRECHED DNA



STRETCHED DNA

However, five other files containing DNA or DNA plus protein coordinates are recorded in the same folder DNA/COORDINATES as the two main DNA files:

- path.pdb
- monstra_completo_1.pdb
- monstra_completo_2.pbd
- pdb1msf.pdb
- pdb1d18.pdb

Remark: Some of these files do not have exactly the same notation as the two main files for atoms naming, instead of a subscript prime mark ’ in C1’ it is C1* for example. To calculate the energy with these files some changes in the program would be needed.

3. Electro potential force fields:

The modelling of DNA molecule has been based on one document above all to start, and is referenced in the list of documents (1):

The model used to calculate the energy within a molecule is described as “minimalist” in its functional form.

$$E_{total} = \sum_{bonds} K_r (r - r_{eq})^2 + \sum_{angles} K_\theta (\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 + \cos(n\phi - \gamma)] + \sum_{i<j} [\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon R_{ij}}]$$

The last term refers to Van Der Waals interactions represented by a 6-12 potential.
This simple representation of bond and angle energies is adequate for modelling most unstrained systems. The goal of this force field is to accurately model conformational energies and intermolecular interactions involving nucleic acids, bases like in DNA molecule.

To simplify even more calculus, only short-range energy will be calculated in a first step, which gives the following formula:

$$E_{total} = \sum_{bonds} K_r (r - r_{eq})^2 + \sum_{angles} K_\theta (\theta - \theta_{eq})^2$$

The r_{eq}, θ_{eq}, K_r and K_θ values are given in Table 14 in the previous article, and r, θ must be evaluated from the atom coordinates file.
The atom types employed are described in Table 1.

II. Coding.

1. Program's description:

   The main goal of this program is to calculate the short-range energy within a given molecule through its atom coordinates set in a pdb file.

   The program is composed of 7 modules, 1 program, and 5 data files, and creates 2 output files, and is located in the following folder DNA/Fortran90/NEW:

   - principal.f90 : this program calls different modules to calculate the Energy of a given molecule from a specific file containing atom coordinates and atom types.

   - global_var.f90 : this module contains all the global variables and a derived type "ATOM" which describes atoms, the components are atom name, atom type, atom coordinates.

   - interface.f90 : this module is an interface with the user of the program and propose to calculate an energy for a given or chosen file.

   - reading.f90 : this module reads a pdb file containing a molecule and stocks values in a vector of ATOM type.

   - bond.f90 : this module searches for bonded atoms and calculates the energy of this bond Ebond.

   - angle.f90 : this module searches the value of an angle between 3 bonded atoms and calculates the energy of the angle Eangle.

   - total.f90 : this module calculates the total Energy (short-range energy) for a molecule.

   - creating.f90 : this module creates a file with the number of atoms, atom symbols and atom coordinates (this file does not contain anymore atom types and can be used later to find back atom types by only knowing atom coordinates).

Data in the following files come from table 14 of the document referenced in I. 3. :

- Angle_Parameters : this file gives parameters concerning angles between 3 given atom.

- Bond_Parameters : this file gives parameters concerning bonds between 2 given atoms.

- Atom_name : this file gives the atom symbol of an atom in DNA notation.

- cut_off_distance : this file gives the maximum cut off distance between 2 atoms in Angstrom.

- convert : this file gives atom types in function of atom notations in DNA molecule.

- out : this file contains output data, that is energy of each bond, each angle, the neighbour array, and total Energy of the molecule.

- file_out : this file contains the number of atoms, atom symbols and atom coordinates.

2. Approximations:

Some approximations have been necessary for this program.

- Data in the cut_off_distance file are approximations done from table 14 in the previous referenced article. The cut off distance represented by 0.0_ means that this kind of bond is not represented in the DNA molecule or means that it was not given in this article.
- Angle_Parameters and Bond_Parameters are not complete, only data needed for the DNA molecule are written in these documents.
- Concerning atom types described in table 1, N* represents sp2 nitrogen in 5-membered ring with lone pair, but it also appears in 6-membered ring in thymine and cytosine from figure 4. Subsequently N* will concern sp2 nitrogen in 5-membered ring or 6-membered ring with lone pair.

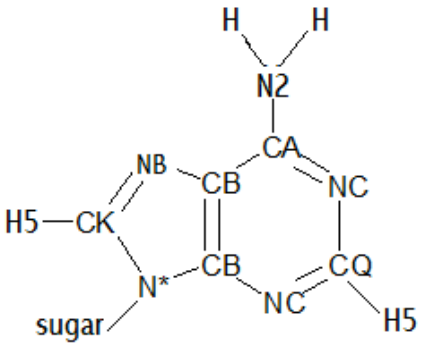
3. Atom types:

A description of atom types is given in Table 1 and the four DNA bases are drawn with these types in figure 4.
However some explanations are described below:

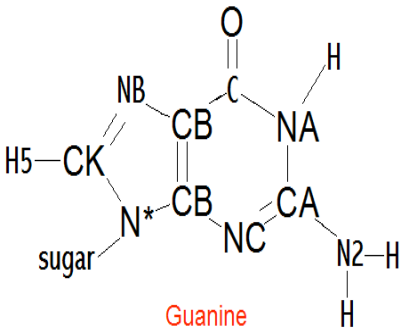
• HC H attached to aliphatic carbon with no electron-withdrawing substituent means the Hydrogen is bonded to an aliphatic carbon which does not have electronegative neighbour like O or N.

• H1 H attached to aliphatic carbon with one electron-withdrawing means Hydrogen is bonded to an aliphatic carbon which has exactly one electronegative neighbour like O or N.
Similarly for H2 and H3 description.

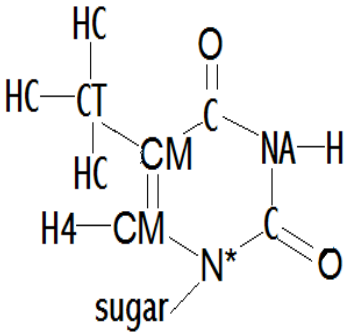
The following pictures represent the 4 bases with chemical notations for each atom:



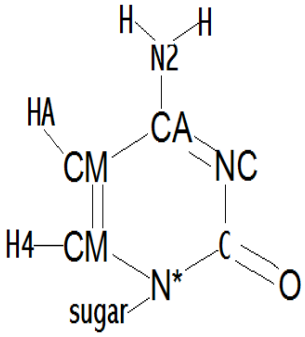
Adenine



Guanine

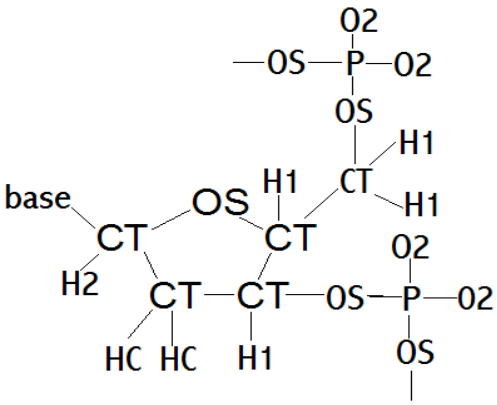


Thymine



Cytosine

Notations for the sugar deoxyribose and the backbone composed of phosphate links are presented in the next figure:



sugar+backbone

III. Results.

1. Applications:

The program has been running with the two DNA molecule main files, that is UNSTRETCHED DNA (DNA.pdb extract from actg_000M.pdb) and STRETCHED DNA (DNA2.pdb extract from actg_200M.pdb). Some results are described below:

UNSTRETCHED DNA:

This file contains 762 atoms and 24 bases of a DNA molecule.

Total bond Energy: $E_{bond} = 2908.9572176553197 \quad kcal/mol$

Total angular Energy: $E_{angle} = 15752.556883288682 \quad kcal/mol$

Total short-range Energy: $E_{total} = 18661.514100944001 \quad kcal/mol$

STRETCHED DNA:

This file contains 762 atoms and 24 bases of a DNA molecule.

Total bond Energy: $E_{bond} = 256.69024670306646 \quad kcal/mol$

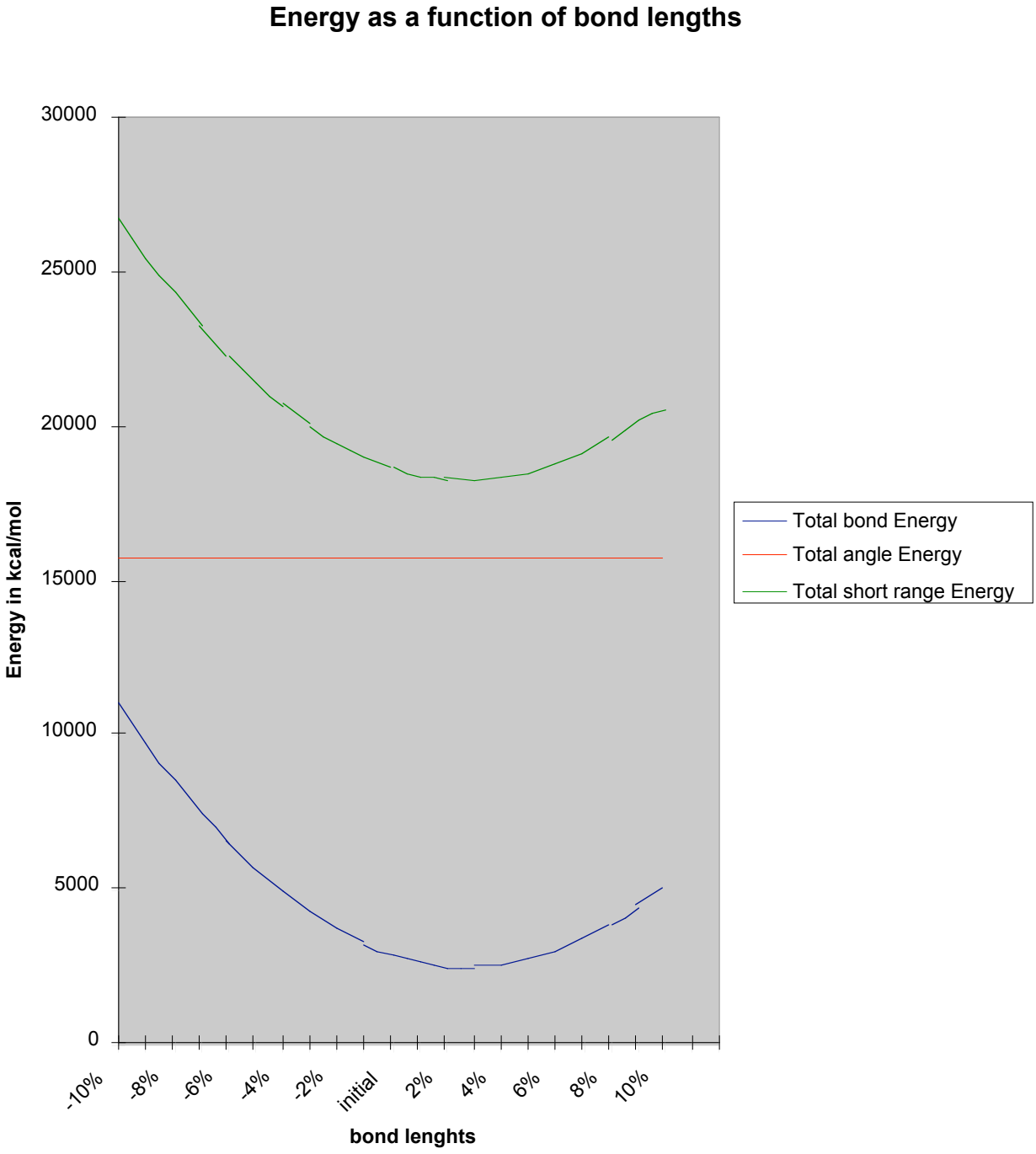
Total angular Energy: $E_{angle} = 15172.43867189256 \quad kcal/mol$

Total short-range Energy: $E_{total} = 15429.128918595627 \quad kcal/mol$

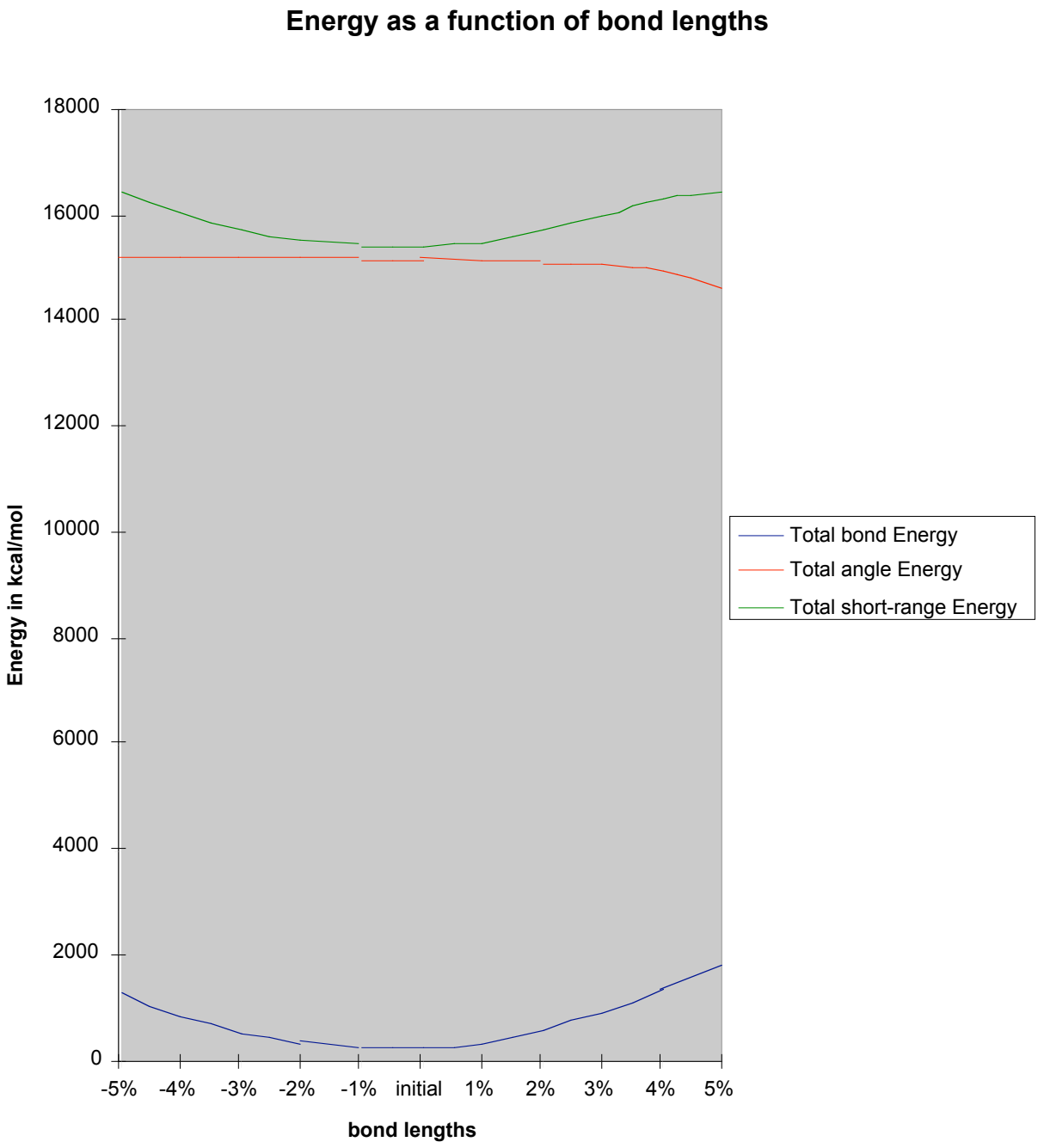
Logically UNSTRETCHED DNA should have a lower total energy than STRETCHED DNA has. But as this is only a short-range energy, some terms are missing which could probably explain the results.

2. Energy as a function of bond lengths:

The following diagram shows energies as a function of bond lengths. From an initial file “UNSTRETCHED DNA”, bond lengths are increased or decreased of 1% to 10%. Thus, consequences on bond energies, angle energies, and total short-range energies are represented in this diagram. A minimum energy is reached for both bond energy and total short-range energy between 2.75% and 3% of bond lengths increasing. The angular energy is not affected by bond lengths increasing.



The following diagram shows energies as a function of bond lengths. From an initial file “STRETCHED DNA”, bond lengths are increased or decreased of 1% to 5%. Thus, consequences on bond energies, angle energies, and total short-range energies are represented in this diagram. A minimum energy is reached for both bond energy and total short-range energy between -1% of bond lengths increasing and initial bond lengths. The angular energy should not be affected by bond lengths increasing, but it seems to decrease from 1% of bond lengths increasing. Cut off distances may be reached, and thus less angle would be calculated which could explain the angular energy decreasing from its initial value.



IV. Future goals.

1. Documents:

   Some documents would need to be studied deeper to see correlations or disagreements about electrostatic force fields.

2. Formula:

   Only the short-range formula has been used to calculate the total energy. The last part of it, dihedrals and Van der Waals forces, should be integrated in calculus for a more accurate result on the energy within a molecule.

3. Atom types:

   A program describing atom types just knowing atom symbols, coordinates and the neighbourhood of each atom should be written. Thus, by imposing a force field to a DNA molecule it would be possible to observe consequences on the molecule's structure (twisting).

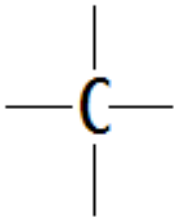
CONCLUSION

This work on DNA modelling has been an amazing experience of research. From a force field model for the simulation of DNA, an approach of an energy function has been carried out. Some results show that calculating only short-range energy is not as an accurate reproduction of electrostatic interactions as the whole energy function would be.
Further applications will be required to develop a simulation of DNA modelling. Indeed, this project appears to be much more complex than expected, and thus becomes a kind of starting to another more important project.

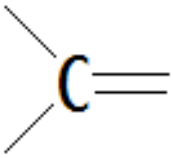
APPENDIX: Hybridization

Carbon:

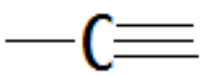
sp3



sp2

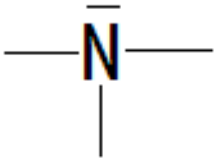


sp

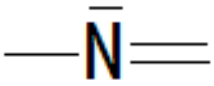


Nitrogen:

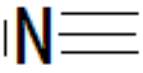
sp3



sp2

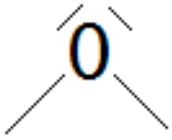


sp

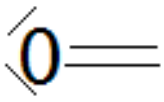


Oxygen:

sp3



sp2



List of documents

(1) “A second generation Force field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules.”
W.C. Cornell, P. Cieplack, C.I. Bayley, I.R. Gould, K.M. Merz, Jr., D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, P.A. Kollman
J. Am. Chem. Soc. 1995, 117, 5179-5197

(2) “Molecular Dynamics Simulation of DNA Stretching Is Consistent with the Tension Observed for Extension and Strand Separation and Predicts a Novel Ladder Structure.”
Michael W. Konrad, Joel I. Bolonick
J. Am. Chem. Soc. 1996, 118, 10989-10994

(3)“Modeling of Long-Range Electrostatic Interactions in DNA.”
A.Vologodskii, N. Cozzarelli
Biopolymers, Vol. 35, 289-296 (1995)

(4) “Electrostatic Potentials of DNA. Comparative Analysis of Promoter and Nonpromoter Nucleotide Sequences.”
R. V. Polozov, T.R. Dzhelyadin, A.A. Sorokin, N.N. Ivanova, V.S. Sivozhelezov, S.G. Kamzolova
Journal of Biomolecular Structure and Dynamics, Vol. 16, Issue number 6, 1999

(5)“model. it: building three dimensional DNA models from sequence data.”
K. Vlahovicek, S. Pongor
Bioinformatics applications note, Vol. 16, no 11 2000, pages 1044-1045, (2000)

(6)“Simulating DNA at low resolution.”
Wilma K Olson
Current Opinion in Structural Biology 1996, 6:242-256

(7)“Computer Simulation of DNA Double-helix Dynamics.”
M. LEVITT
Cold Spring Harbor Symposia on Quantitative Biology, Vol. XLVII

(8) IUPAC-IUB Joint Commission on Biochemical Nomenclature (JCBN) Symbols for specifying the conformation of polysaccharide chains, Recommendations 1981, Eur.J.Biochem. 131,5-7 (1983)

APPENDIX

Program used to calculate short-range energy within a molecule