Table of comtents

Part I : Presentation of the research of the laboratory 2

Marine Natural Products Chemistry 2

Tropical Island 4

Part II : Introduction to molecular modelling 6

Brief history of Molecular Modelling 6

Calculation of energy 6

Starting points for molecular modelling 9

Where do we get physical data to start modelling? 9

How can we find the smallest energy conformation of a molecule ? 10

Molecular Dynamics 13

How can quantum mechanics help us ? 14

Some applications 14

Part III : Some words about the software I used for my project 15

MacroModel 15

Molden 15

X Cluster 17

Molmol 17

Part IV : My project 18

Conformational search 18

Conformation clustering search 19

Calculation of RMSD 19

Conversion of PDB files 21

The overlap with molmol 31

Thanks 32

Appendices 33

Part I : Presentation of the research of the laboratory

Marine Natural Products Chemistry

Pharmacognosy using Marine Invertebrates and Cyanobacteria

Natural products are traditionally the cornerstone of drug discovery. Despite advances in synthetic chemistry and in the understanding of the mechanisms of drug action, the ideal of rational drug design is still a long way off. Natural product discovery from new sources will continue to be essential to provide novel lead compounds which the synthetic chemist can modify. Studies performed at the National Cancer Institute in the USA have shown that marine organisms represent a significant source of biologically active lead compounds.

We are looking at the isolation of novel drug candidates from soft-bodied marine organisms collected in UK and Indo-Pacific waters. The isolation of bioactive compounds from the crude organism extracts is guided by their biological activity. Once a pure compound is isolated, its structure is defined using one and two dimensional nuclear magnetic resonance methods as well as advanced mass spectrometric techniques.

Chemical Ecology

As well as the discovery of biologically active natural products the laboratory is also interested in the role of these compounds in nature. We are currently investigating production of the microcystin toxins by cyanobacteria (blue green algae) of the genus Microcystis.5,10 We hope to discover the chemical cues that stimulate the production of these toxins.

Marine Bioinorganic Chemistry

The low concentrations of metal ions in the marine environment compared to the terrestrial suggests that marine organisms may have evolved unique mechanisms for the uptake of biologically important metal ions from the ocean and their subsequent storage and utilisation. The laboratory work focuses on the discovery of novel ionophores, the organic compounds responsible for metal uptake and trans-membrane ion transport, from marine organisms. Once discovered the modes of action of these compounds are studied by using various physical methods as well as molecular modeling studies.

Structural Organic Chemistry

The laboratory is also interested on the structure determination by spectroscopic methods, and the use of computer assisted structure elucidation. The laboratory is working on the solution state structure determination of small cyclic peptides.

Completed Projects

Toxic Principles in Saliva of the Octopus Eledone cirrhosa

This project is run in collaboration with Professor Peter Boyle, Department of Zoology, University of Aberdeen.

During this research the paralytic toxin from saliva of the northern octopus Eledone cirrhosa was isolated and partially characterised. Methodology was developed for acquiring the saliva and isolation of the active constituent using HPLC and a locust bioassay. We are currently scaling up the isolation process to obtain enough material for full structure elucidation.

Isolation of Divalent Metal Complexing Agents from Marine Invertebrates.
During this we isolated marine invertebrate metabolites complexed to divalent metal ions (Cu2+, Zn2+), and used spectroscopic methods (CD, MS, NMR) to determine their physical properties. The main focus was on modified cyclic octapeptide metabolites from the seasquirt Lissoclinum patella. We determined binding constants and binding selectivity using circular dichroism spectroscopy and mass spectrometry. The binding

environment has been studied using NOE restrained molecular dynamics studies.

Detoxification of Electrophiles by E. Coli

This project was run in collaboration with Prof Ian Booth, Department of Molecular and Cell Biology, Aberdeen University. We determined the structure and production requirements of two electrophiles which were detoxified by E. coli. The work involved microbiological and molecular biological methods as well as separation technology and spectroscopic methods.

Tropical Island

Cyanobacterial Chemical Ecology

This project is run in collaboration with Dr Linda Lawton, Department of Applied Sciences, Robert Gordon University, Aberdeen. The chief aim of this project is to elucidate whether toxin production in freshwater cyanobacteria affords a competitive advantage to the producer organism and to identify if known toxins and/or other previously unidentified compounds associated with these species exhibit allelopathic properties. Initial mixed culture experiments have shown that toxin production is increased when a toxic strain is mixed with a non toxic strain. Spent medium experiments indicate that the effect is large and reproducible. We are currently engaged in the isolation of the chemical cue involved in eliciting the toxin production.

Molecular Self-Assembly of Marine Toxins

The main aim of the proposed work is determine the degree of supramolecular structuring that occurs in complexes of synthetic alkylpyridinium salt (APS) oligomers with dianions. This will be achieved by the synthesis of 1,2 and 1,3 alkylpyridinium salt oligomers with differing connecting chains. We will study the structuring that has occurred using X-ray crystallography where possible, and complement this with circular dichroism studies and nuclear Overhauser effect NMR spectroscopy. We are currently developing solid phase methodology for the synthesis of 1,3-APS oligomers.

Marine Invertebrates as Sources of Novel Pharmacophores

Exploration of sponges of various genera collected from Fijian waters for novel bioactive compounds. Bioassay guided isolation of novel metabolites using antitumour screens (performed by the Paterson Institute for Cancer Research in Manchester and the Ford Cancer Centre in Detroit). Application of advanced spectroscopic methodology to determine the structure of the bioactive metabolites. We isolated and identified a family of chitinase inhibitors which is currently being tested by Zeneca Agrochemicals in insect and plant fungal screens.

Other compounds include a family of cytotoxic agents from a previously uninvestigated sponge.

Solid supported Cu(II) fluorosensors for environmental and medical applications

An important tool for the study of copper in living systems is the use of fluorescent chemical sensor molecules (chemosensors) which can determine the concentration of copper in living systems. We have discovered that some marine natural products might be suitable for modification to generate a purely copper selective chemosensor by attaching a fluorescent group. We intend to chemically synthesise such a fluorescent copper chemosensor and immobilise it on a solid substrate. This will make it useful for the determination of copper concentrations for medical and environmental applications.

Part II : Introduction to molecular modelling

Brief history of Molecular Modelling

1860	Structural stereochemistry first considered (structural formulae used)
1874	Tetrahedral carbon doscovered by van't Hoff
1953	Barton introduces conformational analysis
1958	3D structure of myoglogin solved by X-ray crystallography (only 300 organics solved at this time).
1959	Drieding stick models developed
1965	CPK space-filling models developed
1970s	Computer models began to be used

Calculation of energy

The goal of modelisation is to know the structure which a molecule can take in the space. The themodynamic laws tell that the most stable conformation of the molecule is the conformation who have the lowest energy. If fact, we should calculate the free enthalpy of a molecule but in fact we will calculate the energy of the molecule U.

G function tell us how a conformation is stable compared to another, il G<0, it means that the other conformation is more stable but it doesn’t mean that the transformation will be quick, it’s just mean the transformation is possible.

We have in fact :

G=H-TS

If we consider that entropy is quite nil, we have

G#H

and H=U+PV, if we have no exterior pressure, we can write

G#H#U

In fact, if we want to know the free enthalpy of a conformation, we just have to calculate intern energy.

How calculate the energy of a molecule ?

Considering the intern energy of a molecule, we can divide it in several parts : energy of bonds, energy of angles, energy of torsion, Van Der Waals energy, energy of charges and so on.

- **Bond energy**

The easiest way to have a view of bond energy is to treat a bond as a spring. We have a minimum energy at the equilibrium position, when the bond is pulled or push, the energy increase.

In we are near the equilibrium position, we can calculate the bond energy by the formula :

$E=k(l-l_0)^2$

E : bond energy
k : constant of the spring
l_0 : length of the bond at the equilibrium position
l : length of the bond

We can know the l_0 value using X-ray spectroscopy and k value using infrared spectroscopy.

- **Bond angle energy**

We choose a similar model as bond energy model. In the easiest way is to use the formula :

$E = k_\theta(\theta - \theta_0)^2$

E : bond angle energy
k_θ : constant of angle spring
θ : bond angle
θ_0 : bond angle equilibrium

Of course, we can use this formula only if we near of the equilibrium but in fact the bond angle don’t vary very much.

- **Van Der Waals interactions**

Atoms can’t be too close each other. Atoms behave as if they were hard spheres. The radius of atoms can be estimated with X-ray spectroscopy.
We can estimate the energy of repulsion of two atom by various term : we can use for example an exponential term or a $1/r^{12}$ term.

- **Torsion angle energy**

Molecules can rotate around single bonds, and there is energy barrier to such rotation. For example rotating around C-C of ethane require energy to allow ethane to be in transition state where hydrogen atoms are close each other, and this energy is the Van Der Waals repulsion of hydrogen atoms.

It’s difficult to determine an expression of torsion energy angle, and this expression is specific of a kind of molecules. A expression who gives good results for ethane will give very bad results for cyclohexane. Truncated Fourrier series are often used.

- **Improper torsion**

Torsion angle are also used to keep sp^2 atoms flats. Improper torsion term is introduced to show the disortion from planarity of double bonds.

- **Charge-charge interactions**

It’s necessary to introduce charge-charge interaction only with polarised molecules. For example a carbonyl group’s electron density is polarised towards the oxygen and the energy of interaction will be different if they are aligned or opposed. Aligning the groups brings two partial negative charges close to each other, which is unfavorable compared with the opposite arrangement which pairs partial positive and partial negative charges.

The easiest way to calculate interactions is to give a charge to every atoms and then calculate the energy using the Coulomb’s laws.

Starting points for molecular modelling

Molecular modelling is usually started through three main methods:

- Building using standard geometries - especially bond lengths and bond angles
- Building using fragments which are known to have sensible geometries - these have usually been corrected by some
- Method of "optimisation"
- Building using data obtained from physical experiment - usually X-ray crystallography, neutron diffraction or structure deduced from nuclear magnetic resonance (NMR) data

Where do we get physical data to start modelling?

- **X-ray Crystallographic Data**

This is the primary source of data, for both small molecules and for large molecules, giving the structure of a compound in the solid state. Positions of heavy atoms are located more accurately than lighter atoms. Hydrogen atom positions are frequently in calculated positions rather than observed positions.

- **Neutron Diffraction Crystallographic Data**

This method also gives the structure of the molecule in the solid state. This method is usually more accurate than X-ray crystallography and gives accurate positions for light atoms such as hydrogen.

- **Nuclear Magnetic Resonance Spectroscopy**

This method is capable of giving information about the solution phase structure. However this information is generally ncomplete and needs extra information from other sources to give a complete picture. This method is becoming considerably ore important in the last few years as a method for determining the structure of medium and large bio-molecules.

- **Other techniques**

There are several other less frequently techniques, such as microwave spectroscopy, for determining molecular structure but hese are not used routinely as starting information for molecular modelling.

How can we find the smallest energy conformation of a molecule ?

Optimization is the term for the mathematical process whereby the structure obtained by a round of calculational processes is ompared to a previous structure. The structure is modified to make it more consistent with the parameter information within he program. Various mathematical procedures are used to determine how the geometry will change from one step to the next.

The most common methods are:

- "steepest descent"
- Newton-Raphson method
- simplex method
- Fletcher-Powell method
- or a combination of methods (usually two)

Combining methods is done due to the varying methods being more efficient in different circumstances, e.g. steepest descent is easiest to program and understand but is very slow to converge when on a shallow potential energy surface. However it is excellent at correcting major abnormalities at the start of a calculation. The program keeps altering the geometry until a specified cutoff value is reached the molecule is said to be optimized. The specified cutoff value is termed the convergence criterion. A common convergence criterion is that the change in energy, between the last structure calculated and the second last structure calculated, of less than .05 kJoules. Generally the convergence criterion is based on measuring changes in the energy or changes in the geometry or both.

Force fields used for optimization are essentially divided into two classes:

- The first is for use with small molecules with all atoms including hydrogens being included in the calculation. This is an "all atom" approach.

- For large biological molecules, e.g. proteins and nucleic acids, an "essential atoms only" approach is used. Here the majority of hydrogen atoms are removed from the structure in order to decrease computational time. The only hydrogen's maintained are those connected to heteroatoms, the "essential hydrogens". To compensate for this carbons have an expanded van der Waals radius which accommodates the missing hydrogens. This method is known as the "united atom" approach.

The best-known molecular mechanics package for small molecules is MM2 (U. Burkert and N. L. Allinger, "Molecular Mechanics", American Chemical Society, Washington D. C., 1982).

For large molecules the best known program is AMBER (P. W. Weiner and P. A. Kollman, J. Comput. Chem., 1981, 2, 287-303). Two other programs CHARMM and GROMOS are also widely used.

When optimizing "ordinary" organic molecules there are usually no problems encountered obtaining adequate parameters as these will have been invented and tested previously. HOWEVER the most common problem encountered using molecular mechanics is the error message "no parameters for XXXX interaction". Therefore parameter invention is necessary.
This is relatively easy for all parameters except torsion angle parameters (using consideration of hybridization state, commonsense and analogy).

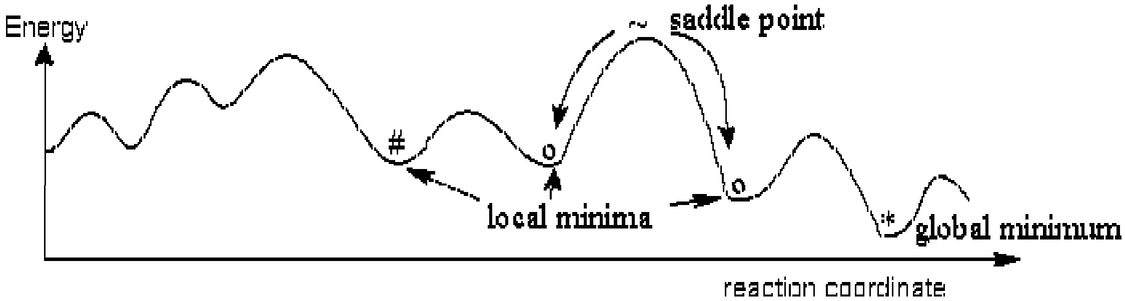
When a united atom force field is being used it is often necessary to include terms for improper torsion angles (a torsion angle where the atoms are not sequentially bonded to each other) to maintain the correct stereochemistry and sometimes to maintain planarity. If this is not done chirality can be changed.

Molecular mechanics calculations, in general, give good geometries though care needs to be taken with strained molecules.

Conformational information can be easily obtained by comparing the difference in energy for different conformations of the same molecule. This may involve "constraining" some particular feature of the geometry, e.g. a torsion angle to a set value to be retained during optimization.

The major problem with molecular mechanics calculations is that they converge on the nearest local minimum which is not necessarily the global minimum.

If, for example, the potential energy surface is depicted as a two dimensional surface:



then an optimization starting at points A or B will converge on local minima and not the global minimum. An optimization starting with at C will optimize to the global minimum. When there are only a few rotatable torsion angles (realistically less than 6) then it is possible to systematically rotate the torsion angles and locate the global minimum. When there are multiple torsion angles then an unoptimized structure is usually subjected to Molecular Dynamics to find low energy conformations by randomly sampling conformational space.

Molecular Dynamics

This method uses the Newtonian equations of motion, a potential energy function and associated force field to follow the displacement of atoms in a molecule over a certain period of time, at a certain temperature and a certain pressure. Calculations of motion are done at discrete and small time intervals and a velocity calculated on each atom position which in turn is used to calculate the acceleration for the next step. Starting velocities can be calculated at random (necessary when starting at 0 Kelvin where the kinetic energy is 0) or by scaling the initial forces on the atoms. Simulations can also be run with differing temperatures to obtain different families of conformers. At higher temperatures more conformers are possible and it becomes feasible to cross energy barriers.

When doing calculations on biological molecules it is becoming more frequent to do the calculations in the presence of solvent (usually water!!). However, this brings further complications due to two main problems. The first being increased CPU time due to the larger number of atoms. The second is that the water molecules surrounding the molecule tend to drift away from the molecule of interest and get "lost" from the calculation if only a certain area of space is being monitored as is usually the case.

This causes nasty "edge effects". There is one method currently used to get around this problem. That is to place your molecule surrounded in water in a box of a specific size and then to surround that box with an image of itself in all directions. The solute in the box of interest only interacts with its nearest neighbour images. Since each box is an image of the other, then when a molecule leaves a box its image enters from the opposite box and replaces it so that there is conservation of the total number of molecules and atoms in the box. This are known as periodic boundary conditions.

Simulated annealing is a special type of dynamics. The molecule is heated and then cooled very slowly so that conformational changes taking place will lead to the global minimum being located.

Related to molecular dynamics are Monte Carlo methods which randomly move to a new geometry/conformation. If it is lower or close in energy it is accepted if not an entirely new conformation is generated. This process is continued until a set of
low energy conformers has been generated a certain number of times

How can quantum mechanics help us ?

Some applications

Part III : Some words about the software I used for my project

MacroModel

MacroModel is a piece of software who is can used to draw molecules, to calculate the energy of a given conformation, and find the minimum energy by various methods, such Monte-Carlo. Morever, Macromodel can give length, angle and angle torsion of a given conformation. It also can calculate the difference between two conformation after overlaping. I learnt Macromodel during the two first weeks of my project. I used Macromodel to solve easy problems such difference in energy between chair and boats cyclohexane conformation, difference between axial and equatorial position of several substituants, and I also used MacroModel for organic chemestry reactivity.

Molden

Molden is an interface to GAMESS, a program who uses quantics mechanics methods.

I used Molden to do a dihedral drive of C-S bond of cysteine.

Molden doesn't used cartesian coordonates, it uses Z-matrics format instead.

Description of Z-matrics format

It is sometimes convenient to describe a molecule in terms of internal coordinates, using a Z-matrix, rather than using Cartesian Coordinates. This means that the position of each atom is expressed in terms of the positions of atoms which have already been defined. Thus a typical line in a Z-matrix description of a molecule looks like:

Atom Type r *atom 1* θ *atom 2* ϕ

The first item is the sort of atom that is being described. The second, *r*, is the distance from this new atom to another atom *atom 1*. This atom must have already been described earlier in the Z-matrix. There is now an angle, θ, which is the angle created by the new atom, *atom 1* and *atom 2.* A second angle, ϕ, describes the torsion angle between the new atom and *atoms 1—3* (Figure 1).

This description depends on there being three atoms that are already defined, so the beginning of the Z-matrix is slightly different. The first atom is usually just given an atom type, and no information about its position. The second atom will be defined simply by its distance from the first atom. The third by its angle. For example, hydrogen peroxide (HOOH), is a simple four-atom molecule whose Z-matrix (minimised by AM1) is as follows:

H1						
O1	1.1	H1				
O2	1.5	O2	107.0	H1		
H2	1.1	O2	107.0	O1	100.0	H1

An internal coordinate description of a molecule has the feature that it is not necessary to define a position or an orientation for the molecule, since all of the atoms are defined relative to each other. It is also possible to rotate around torsion angles, by making simple changes to the Z-matrix. For example, the H-O-O-H torsion angle in the above example is set to 100.0°, but only this single parameter need be altered to adjust the angle. In a description of the molecule using cartesian coordinates, the x, y and *z* coordinates of at least one of the atoms would need to be altered, which would be very much less convenient.

Dihedral drive of cysteine

To do the dihedral drive, I had to keep the C-S torsion angle as a constant (in the beginning a 210 degrees) and let all the other value (all length bond, all angle bonds, all torsion but C-S torsion) variable. So Molden calculate the values to have the minimal energy. And after I would do the same things but use 240 degrees insted 210, after 300 after 330 ... every 30 degrees. At the end, I could have the enedy profile of cysteine if I turn the two part of the cysteine around C-S bond.

You can have some trouble if you have in your molecule 180 degrees torsion angle bond, because the program consider angle from -180 to 180 degrees and if you a little more than 180 degrees, it would be -179.999... degrees and because of this thing, if you have 180 degrees torsion angle, the program may crash. As I had a aromatic cycle, it was impossible to me to avoid 180 degrees torsion angle so I could't make the energy profile.

X Cluster

X Cluster is program designed to compare conformations each other and to tell you how many conformations there are of each shape.

I used this program after I did Monte Carlo conformation searching

Molmol

Molmol is a molecular graphics program for displaying, analysing, and manipulating the three-dimensional structure of biological macromolecules, with special emphasis on the study of protein or DNA structures determined by NMR. Molmol has a graphical user interface with menus, dialogue boxes, and on-line help. The display possibilities include, besides the type of representations found generally in other molecular graphics programs, novel schematic molecular representations. All types of
representations can be combined in the same display. Structures can be manipulated by adding and removing atoms and bonds and by interactive rotation about dihedral angles. Special efforts were made to allow for appropiate display and analysis of sets of (typically 20-40) conformers that are conventionally used to represent the result of a NMR structure determination. Thus, Molmol has functions for superimposing sets of conformers, calculating RMSD values, identifying hydrogen bonds, checking and displaying violations of NMR constraints, and listing short proton-proton distances.

Part IV : My project

Marines sponges contains peptides who they have some anti-cancer properties. All of this peptides are cyclics, they have like a dozen of amino-acids and all of them are many proline amino-acids. The goal of my project was to compare these molecules each other, in fact I had nine molecules to compare.

Conformational search

In the first part of my project, I searched the conformations of the nine peptides of the global minimum and local minimums who didn’t have more than 50 kJ more than global minimum.

I used for that Monte-Carlo algorithm. I asked to the computer to try 5000 molecules. Of course, most of them (don’t forget it’s semi-random process) was rejected (by energy, by constrainst, by chirallity change). It took one day of calculation per molecule.

I did conformation searching using implicit hydrogens, i.e., I didn’t put missing protons to the carbons. There only were hydrogens on hydroxyl and amino groups. I used AMBER force field.

Results

Molecule	**number of unique confor-mations**	**Number of conformations found of lowser uner a given energy below the global minimum (in kcal/mol)**					**Energy of the global minum (kcal)**
		1	**2**	**3**	**5**	**10**	
A	666	10	13	22	54	435	-537.26
B	906	1	8	17	70	570	-394.73
C	735	4	8	30	121	540	-454.17

D	858	1	4	21	117	658	-462.03
E	799	8	34	68	177	646	-576.69
F	1110	11	32	70	224	911	-299.31
G	1179	5	22	59	218	960	-245.59
H	1260	27	86	195	564	1189	-338.45
P8	987	10	29	74	220	789	-529.08

Conformation clustering search

Once the conformation search was finish, I had to see how the conformations found ressamble each other, i.e how many stable conformation the molecule has.

For that kind of things, I used X cluster, described in third part of this report. For each structure, I looked for highest level cluster. After that I selected this cluster, I could see the main conformation(s).

I saved the cluster structure file and this file can be open by Monte-Carlo.

Calculation of RMSD

At this moment, I determined the main conformation(s) of my structures. If there was two main conformations, I only took the first.

I was ready for comparison, but as the nine molecules haven’t the same structure, it was ridiculous the compare some of them. Don’t forget that we are interested by proline properties.

So proline amino acid will be called P and other amino acid will be called X.

I only represent the part of the molecule where there is not more than two X-amino-acid between two proline amino-acid.

Molecule	Structure
A	PPXP
B	PPXXP

C	PXPXXP
D	PXXPXXP
E	PXPXP
F	PXPXP
G	PXPXP
H	PXPXP
P8	PPXXPP

As the E, F, G, H as the same PXPXP sequence, I decided to compare this four molecules each other.

To do this, I used the SUPerposition Atom function of Macromodel. In a first time, I only overlap nitrogen of amino groups of proline amino-acids.

This is the results (in A)

	E		
F	1.046	**F**	
G	0.747	0.646	**G**
H	0.422	0.983	1.470

You can see that only E and G, E and H and F and H are close enough. So, I carried on the comparisons with these pairs of compunds.

I had to do a better overlap to have a better comparison. I had to overlap all atoms of proline and the NH-CH-CO part of X amino acids where are between two prolines.

It was very very very difficult to me to use Macromodel to overlap a great number of atoms because if you make a mistake during your selection, you had to restart the selection from the beginning. It's quiet easy to do overlap when you have only three atoms (e.g. three nitrogens) but as I had in that case many atoms, it was quite impossible.

To do this work, I used Molmol.

There is only one kind of file format that both Macromodel and Molmol reconize : it's PDB file format. Unfortunatelly, there are a few differences between Macromodel PDB files and Molmol PDB files format.

Conversion of PDB files

So, I had to make a program to do the conversion. I chose for that QBasic.

The thirst to in this case is to compare the two format files. (The beginning of this files are given in appendice)

As you can see it, the are a few differences (this differences are only in the lines beginnig with HETATM in Macromodel PDB file) :

Columns	Macromodel PDB file	MolMol PDB file
1-6	“HEDATM” string	“ATOM ” string
14-16	Type of atom followed by a atom number	The atom name according to IUPAC notation
18-20	“UNK” string meaning unknow	The amino
22	A position letter of amino acid in the molecule	Nothing
26	Always "1"	A position number amino acid in the molecule

The description of ATOM format was a good help to find meaning of the differents coloums.

These is the description :

Colums	Data type	Field	Definition
1 - 6	Record name	"ATOM "	
7 - 11	Integer	serial	Atom serial number.
13 - 16	Atom	name	Atom name.
17	Character	altLoc	Alternate location indicator.
18 - 20	Residue name	resName	Residue name.
22	Character	chainID	Chain identifier.
23 - 26	Integer	resSeq	Residue sequence number.
27	Achar	iCode	Code for insertion of residues.

31 - 38	Real(8.3)	x	Orthogonal coordinates for X in Angstroms.
39 - 46	Real(8.3)	y	Orthogonal coordinates for Y in Angstroms.
47 - 54	Real(8.3)	z	Orthogonal coordinates for Z in Angstroms.
55 - 60	Real(6.2)	occupancy	Occupancy.
61 - 66	Real(6.2)	temp Factor	Temperature factor.
73 - 76	LString(4)	segID	Segment identifier, left-justified.
77 - 78	LString(2)	element	Element symbol, right-justified.
79 - 80	LString(2)	charge	Charge on the atom.

I hadn't yet informations about atoms names and residue names (residue name is the three-letter code of an amino acid).

I easely found the residue name in the appendice 3 of the PDB Guide.

The residue name are given in appendice.

For each amino acid, we have in atom name coloumn the kind of atom (C, O, N, H) and its place in the chain. The atom names for each amino acid are given in appendice.

In fact, when I made the program, I didn't find this table yet, I used data contained in the PDB example file of Molmol. Unfortunatelly, this file didn't contain any tryptophan data. (The usefull information I was needed is also given in appendice).

Once I understood the meaning of colunms and the part who must be changed, I could write the program.

This is the program :

```
REM Program of conversion of PDB files
REM ----------------------------------------------
REM input file : PDB Macromodel file
REM output file : PDB Molmol file
REM ---------------------------------------------

DIM n$(7) 'array of three-letter code amino-acids
DIM n(7) 'array of number of atoms for each atom
         '(implicit hydrogen rule)
DIM d$(7, 20) 'array of atoms names for each
              'aminoacid


'filling of arrays

FOR i = 1 TO 7 'i is the amino-acid number
READ n$(i)
READ n(i)
FOR j = 1 TO n(i)
READ d$(i, j)
NEXT j, i

OPEN "i", #1, "molg"
OPEN "o", #2, "molg2"

FOR i = 1 TO 3 'let the compund description
               'unchanged
LINE INPUT #1, a$
PRINT #2, a$
NEXT i

READ n    'read amino-acid number of the molecule

FOR i = 1 TO n

READ aa    'read amino acid number

FOR j = 1 TO n(aa)

LINE INPUT #1, a$ 'read amino ac

REM fields of line beginnig by HEDATM
REM colomns 1-6 : HEDATM keyword must be changed
REM in ATOM
REM colomns 14-16 : atom name
REM colomns 18-20 : three-letter amino-acid code
REM colomn 22 : it's a letter of the order amino
REM acid in the molecule and must be deleted
REM colomn 26 : order in amino acid (always 1 in
PDB Macromodel file)

MID$(a$, 1) = "ATOM  "
MID$(a$, 14) = d$(aa, j)
MID$(a$, 18) = n$(aa)
MID$(a$, 22) = " "
MID$(a$, 25) = STR$(i)    'don't forget the space

PRINT #2, a$ 'write the line in the output file

NEXT j, i

DO   'let the end of file unchanged
LINE INPUT #1, a$
PRINT #2, a$
LOOP UNTIL EOF(1)

CLOSE


'atom names data for each amino acid

DATA GLU
DATA 10,"N  ","CA ","C  ","O  ","CB ","CG "
DATA "CD ","OE1","OE2","H  "

DATA ILE
DATA 9,"N  ","CA ","C  ","O  "
DATA "CB ","CG1","CG2","CD1","H  "

DATA LEU
DATA 9,"N  ","CA ","C  ","O  ","CB "
DATA "CG ","CD1","CD2","H  "

DATA PHE
DATA 12,"N  ","CA ","C  ", "O  ","CB "
DATA "CG ","CD1","CD2","CE1 ","CE2 ","CZ ","H  "

DATA PRO
DATA 7,"CD ","N  ","CA ","C  ","O  ","CB ","CG "

DATA TYR
DATA 14,"N  ","CA ","C  ","O  ","CB "
DATA "CG ","CD1","CD2","CE1","CE2"
DATA "CZ ","OH ","HH ","H  "

DATA TRP
DATA 14,"N  ","CA ","C  ","O  ","CB ","CG "
DATA "CD1","CD2","CE2","CE3"
DATA "CZ2","CZ3","CH2","H  "

REM ***************************
REM molecule description
REM ***************************
REM 1. glutamic acid
REM 2. isoleucine
REM 3. leucine
REM 4. phenylalanine
REM 5. proline
REM 6. tyrosine
REM 7. tryptophan
REM ***************************
REM this is amino-acid chain of the
REM molecule previoused by the
REM number of amino acid in the
REM molecule.
REM e.g. for Pro-Tyr-Iso
REM you put
REM data 4,5,6,2
REM ***************************

DATA 7,5,4,5,2,2,5,6
```

End of the program

Some explanations

All informations about the instructions are given in appendice.

For the number of the columns, I used the information given below and I used a text editior (EDIT supplied with DOS system) to exactly know the colomns number (I viewed the PDB Molmol example file).

To replace one part of string by another, I used MID$ instruction. It's very powerfull but only works if the number of characters remains unchanged.

The program assumed that the input file is really made by Macromodel. In that kind of file, there exactely three lines of compound descrip tion before atom description.

In the first part, the program reads all amino acid data.

The program reads the three first lines of the input file and puts them in the output file without change.

Then, the program reads the number of amino acid.

After, there is a loop for each amino acid (the number of repetion of the loop is the number of amino acids in the molecule)

There is another loop for each atom of each amino acid

The program reads the line of input file, makes the changes using MID$ instruction (all the changes mades are written in REM lines in the program) ant put them in the output file.

End of the loops here

In the end, the program reads the last lines of the input files and puts them output file without changes (it's the bonds data).

Note that the program assumes that the peptid is cyclic. It's why we only consider amino acid in with neither amino nor carboxylic group.

The PDB file made with this program should work, in fact Molmol read it. But Molmol can't compare properly two molecules who they don't have the same number of amino acid and who the wmino acid chian (remind that all molecules are cyclic and you start the description with whatever amino acid you want) start with a random amino aicid. To compare two peptides, I had to write another program in order to the description begin with PXPXP amino acid chain. (With molmol we can compare two residue of the same number).

That's the program :

```
REM program for modifing atom order
REM (we want PXPXP first, with first P
REM number one)
REM ---------------------------------------------
REM input file : initial Molmol PDB file
REM output file : Molmol file with PXPXP first
REM ----------------------------------------------

DIM n(7)'array of number of atom
        'for each amino acid
DIM naa(100)'array for know where the new atom
            'is when you know old atom place

'filling of array

FOR i = 1 TO 7
READ n(i)
NEXT i

OPEN "i", #1, "molg3"
OPEN "o", #2, "molg4"

FOR i = 1 TO 3 'let the compound description
               'unchanged
LINE INPUT #1, a$
PRINT #2, a$
NEXT i

READ n     'read the number of amino acid
```

```
           'of the molecule

na = 0  'counter for atoms, (in fact the number
         of lines written) na : new atom

FOR i = 1 TO n

READ aa    'read amino acid number (see the
           'table at the end of the program)

FOR j = 1 TO n(aa)

LINE INPUT #1, a$

REM uselful information in PDB file
REM in ATOM section
REM colonnes 10-11 : counter of atom
REM colonne 26 : position of amino acid in the
molecule

na = na + 1'increase the counter of atoms
naa = VAL(MID$(a$, 10, 2))'put the value of old
                         'atom in naa variable
naa(naa) = na'put the new atom number in the
             'old atom array

MID$(a$, 10) = RIGHT$(STR$(na), 2)
'replace the old atom number by the new one
MID$(a$, 25) = STR$(i)'put the new position of
                      'amino acid
                      'don't forget the space
PRINT #2, a$

NEXT j, i

DO 'now correct the bonds connections
LINE INPUT #1, a$

FOR c = 10 TO 25 STEP 5
'c is the number of the column
```

```
aa = VAL(MID$(a$, c, 2))'read old atom
                                 'number
'aa is nil is there is not number in the column
IF aa <> 0 THEN   'if there is a number ...
   MID$(a$, c) = RIGHT$(STR$(naa(aa)), 2)
   'put it in the new in the place of old
   'using the old atom array
END IF
NEXT c

PRINT #2, a$
LOOP UNTIL EOF(1)

CLOSE


'number of atom (implicit H rules)
'in each amino acid
DATA 10,9,9,12,7,14,14

REM ***************************
REM way of writing of the mol.
REM ***************************
REM 1. glutamine
REM 2. isoleucine
REM 3. leucine
REM 4. phenylalanine
REM 5. proline
REM 6. tyrosine
REM 7. trytophan
REM ***************************
REM put here the new
REM structure. The beginning
REM of the molecule must be
REM PXPXP and the first proline
REM must be number one.
REM ***************************

DATA 7,5,6,5,4,5,2,2
```

Explanations

The program only changes residue and atoms numbers.

The program uses an array to know where is new number atom for an old given one.

It's useful for lines beginning by CONNECT.

When the programs read the structure (lines beginnig by ATOM), it fills the array.

Note I call old atom atom the number that the atom have in the input file and the new one the number in the output file.

So it read the old atom number.
It put it in the array and it puts the new atom number instead.

To know the new atom number, the program uses a atom counter (it's in fact a line counter).

So we use naa(8)=91
means that we put the 91 (new atom number) in the positon 8 in old atom number.

So if we want to know where the new atom is we just have to look at in the table at the old atome place to have the new one.

The last line of the program is only useful to write residue number.

Once all lines beginning with ATOM have been read, the program modifies atom numbers in the connections.

In these lines there are four columns of number, at columns 10, 15, 20, 25 so it's easier to make a loop to read a line.

The program reads the old atom number in the line and in the column c.

If it's a space, the old atom number variable will have a nil value. If the ther is a number (the old atom number is not nil) the program looks at in the table and puts the new atom number instead.

The overlap with molmol

I used SelectAtom command to select atoms and Fit command to do the overlap.

The syntax of SelectAtom is

SelectAtom 'expression'

I call simple expression is the character # followed by a list of molecule names and numbers or the character followed by a list of residue names and numbers or the character @ followed by a list of atom names

If you use several simple expressions without space between you selct atom who would selcect by the two simple expression. (logical AND)

If you use several simple expressions with a space between you selct atom who would selcect by one of the two simple expression. (logical OR)

Example of expressions

Expression	**Selection**
`@CA`	all atoms named CA
`:10-20`	all atoms in residues number 10 to 20
`#1-3,5:10-20,25,LYS@N,CA,C`	all atoms named N, CA or C in residues number 10 to 20 residue number25, and residues named LYS, in molecules number 1 to 3 and 5
`:10@HN :17@HA`	atom named HN in residue number 10 and atom named HA in residue number 17

My expressions

	atoms overlap	**expression**
a	the three proline nitrogens	`#1-2:1,3,5@N`
b	all atoms named N, C and CA in the five first residues	`#1-2:1-5@N,CA,C`
c	all prolines atoms and all atoms named N, C, CA, H and O in the five first residues	`#1-2:1-5@N,CA,C,O,H #1-2:1,3,5`

Results (RMSD values)

molecules overlap	**a**	**b**	**c**
E and G	0.853	1.901	2.656
F and G	0.423	1.941	2.213
F and H	0.540	1.308	2.078

The letters refers to three kinds of overlap listed in the previous table.

Thanks

Special thanks to Dr Jaspars to have supervised my project, to Dr Marr to have receive me and to Dr De roy and Dr Fournier to have permit me to study in a Scotich laboratory in computoring chemestry.

Appendices

Appendix 1 : Molecules of my project 34

Appendix 2 : Amnio acid conventions 38

Appendix 3 : Example of a Macrodel PDB file 39

Appendix 4 : Example of a Molmol PDB file (extract) 43

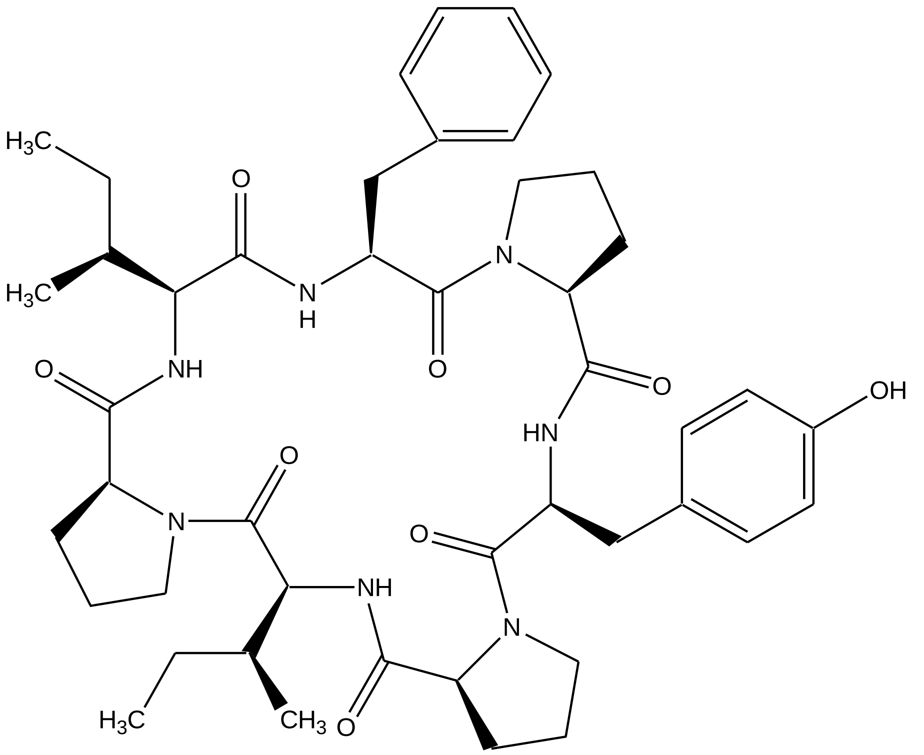
Appendix 5 : Amino acid atom description table A 48

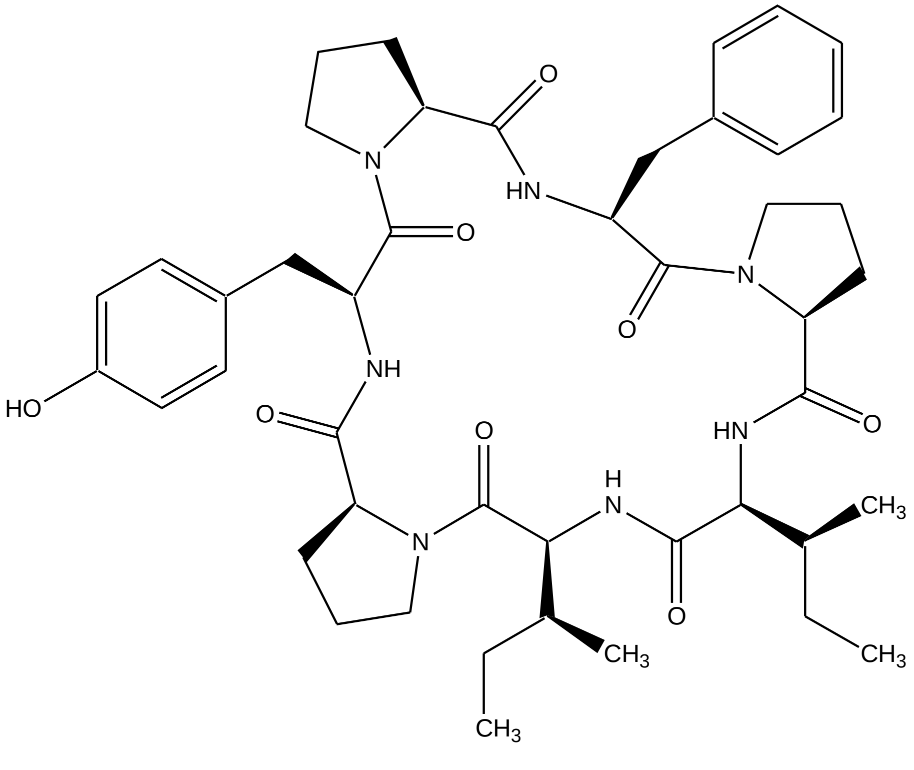
Appendix 6 : Amino acid IUPAC notation 57

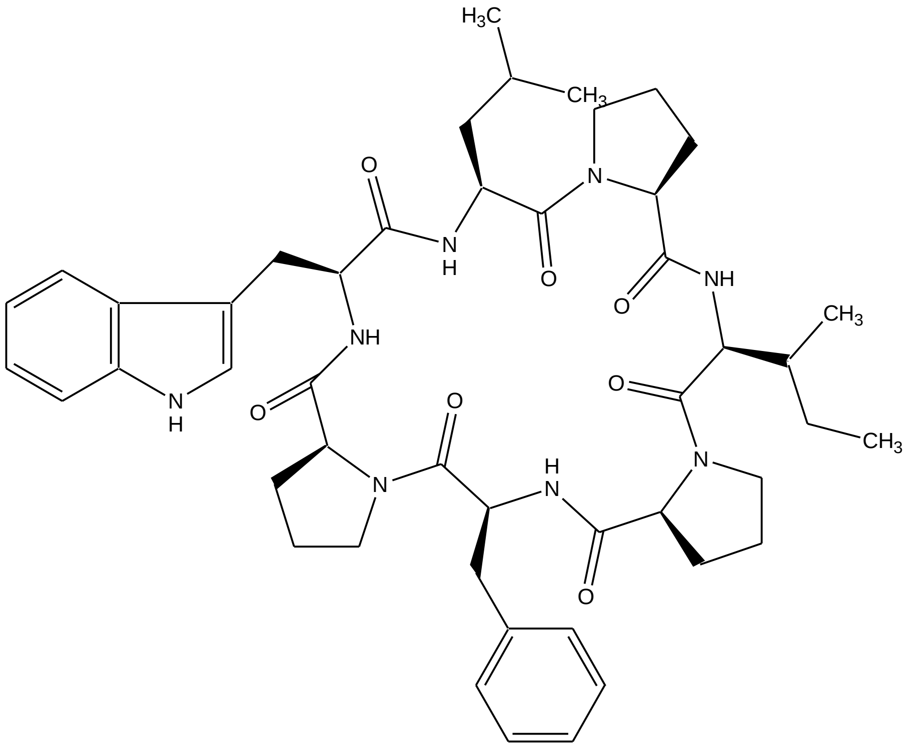
Appendix 7 : Amino acid atom description table B 58

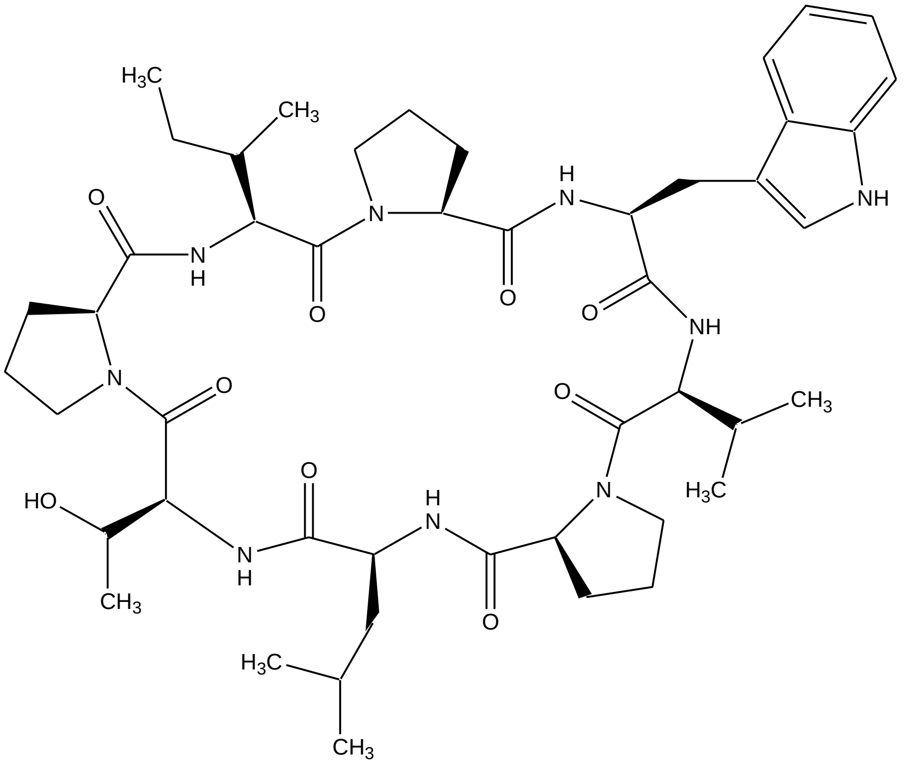
Appendix 8 : Basic instruction syntax 61

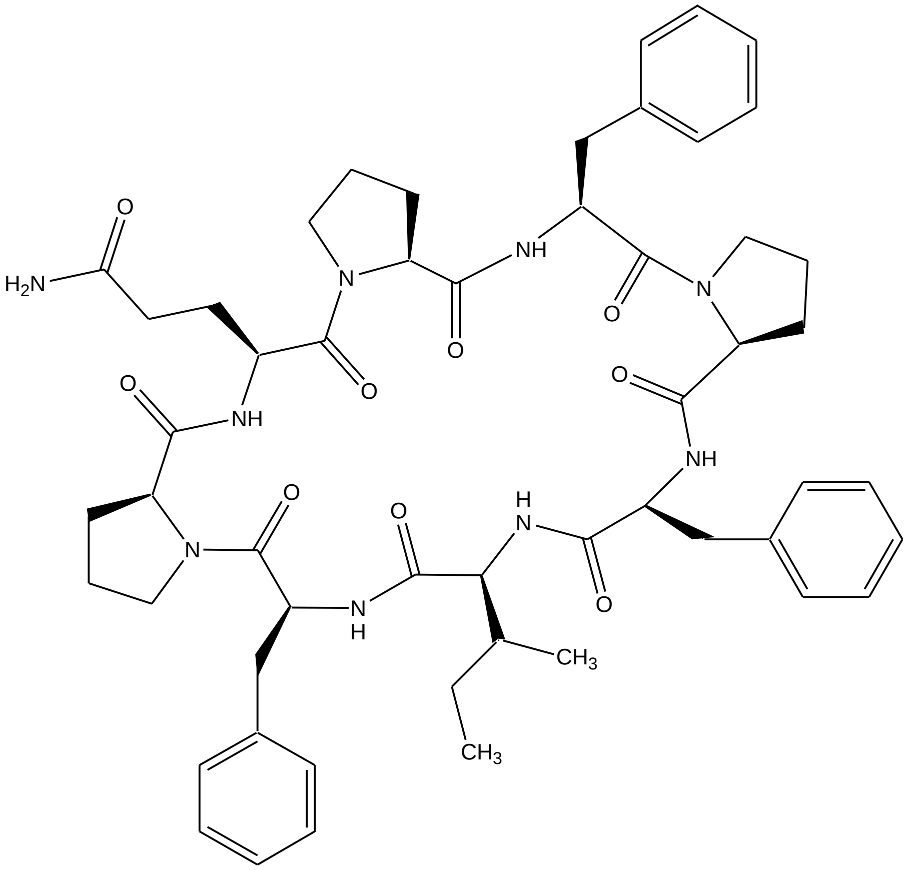
Appendix 1 : Molecules of my project

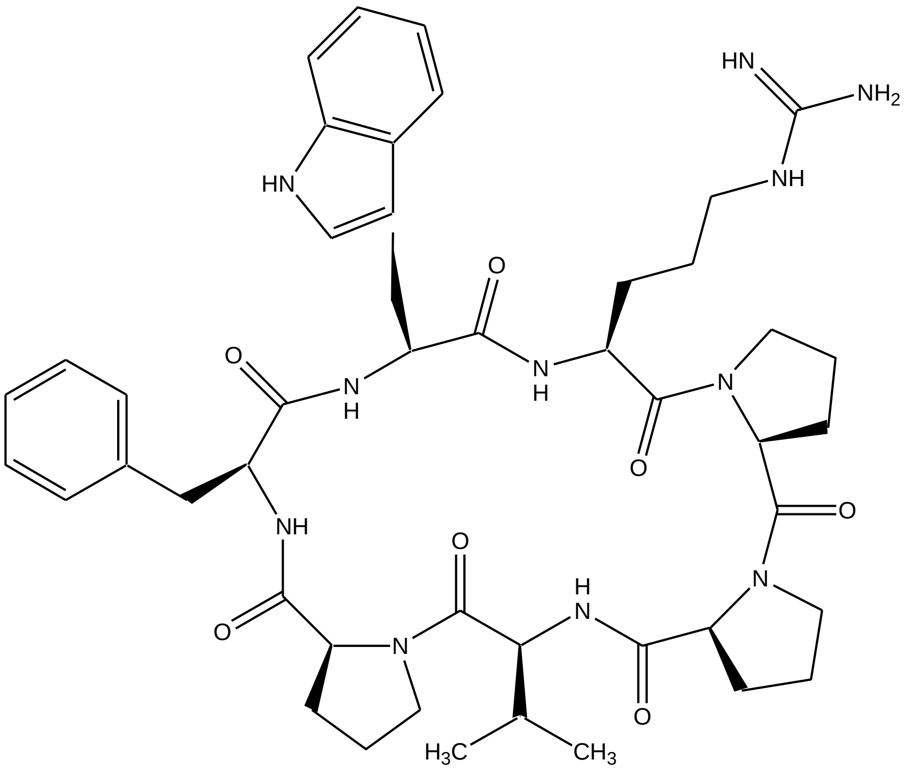


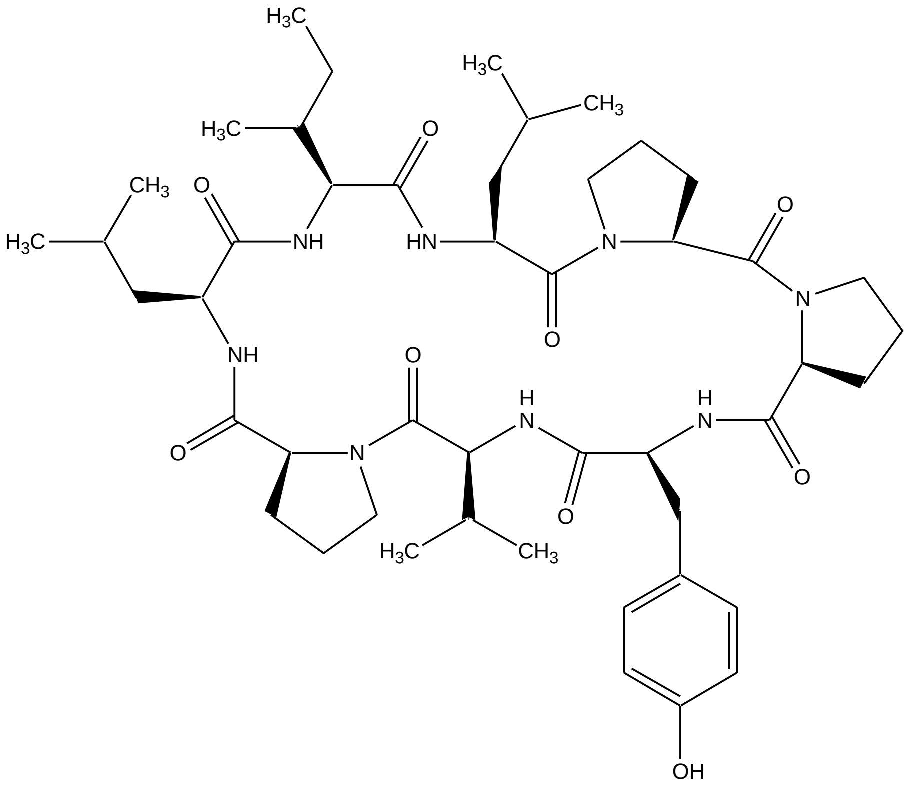


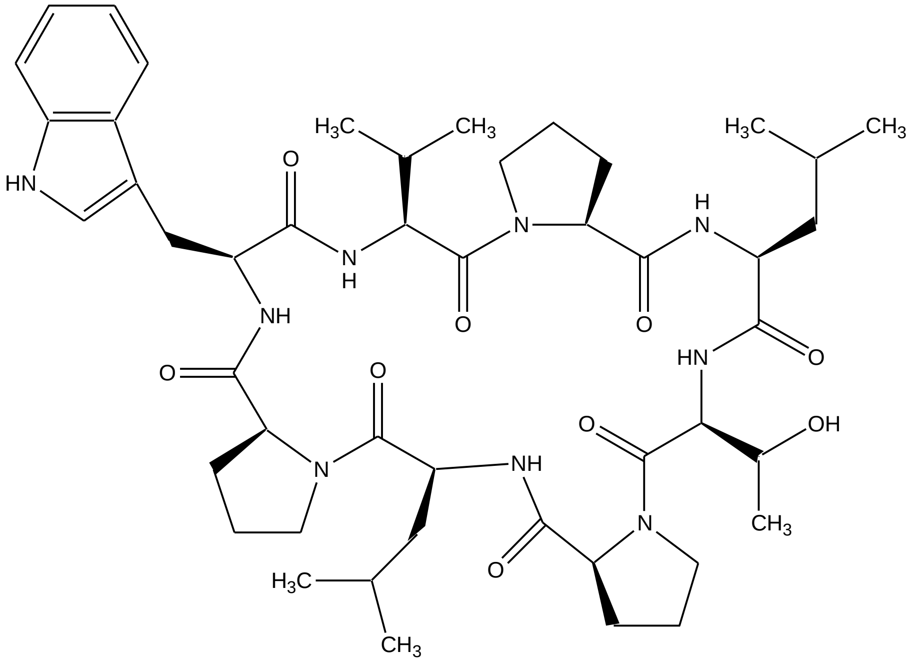

H_3C
CH_3
O
N
NH
H
O
NH
O
CH_3
NH
O
N
H
O
O
N
N
H
CH_3
O


H_3C
CH_3
O
N
H
N
H
O
O
NH
NH
O
N
O
NH
O
CH_3
HO
N
H
O
H
N
N
H_3C
CH_3
O
H_3C
CH_3






H_3C
H_3C
CH_3
H_3C
O
CH_3
H_3C
H_3C
O
NH
HN
N
O
N
NH
O
O
H
N
H
N
N
O
H_3C
CH_3
O
OH


HN
H_3C
CH_3
O
H_3C
CH_3
H
N
N
N
H
O
O
NH
NH
HN
O
O
O
O
OH
N
NH
N
CH_3
H_3C
O
CH_3

Appendix 2 : Amnio acid conventions

Residue (Amino acid)	**three-letter code**
Alanine	ALA
Arginine	ARG
Asparagine	ASN
Aspartic acid	ASP
Cysteine	CYS
Glutamine	GLN
Glutamic acid	GLU
Glycine	GLY
Histidine	HIS
Isoleucine	ILE
Leucine	LEU
Lysine	LYS
Methionine	MET
Phenylalanine	PHE
Proline	PRO
Serine	SER
Threonine	THR
Tryptophan	TRP
Tyrosine	TYR
Valine	VAL

Appendix 3 : Example of a Macrodel PDB file

```
COMPND     Clustering=1 Rep=556 Members=795 Leading=1 Max_Rg=3.01
REMARK   1  PDB:     73     74     75     76
REMARK   1 MMOD:     73     74     75     76 /
HETATM    1  N02 UNK A   1      -23.971 -24.409 -39.785 -0.52 -0.52    0
HETATM    2  C03 UNK A   1      -22.979 -24.514 -40.842  0.25  0.25    0
HETATM    3  C04 UNK A   1      -23.414 -25.374 -42.036  0.53  0.53    0
HETATM    4  O05 UNK A   1      -23.380 -26.605 -41.959 -0.50 -0.50    0
HETATM    5  C06 UNK A   1      -21.681 -25.095 -40.259  0.00  0.00    0
HETATM    6  C07 UNK A   1      -21.194 -24.432 -38.962 -0.21 -0.21    0
HETATM    7  C08 UNK A   1      -21.358 -22.922 -38.934  0.62  0.62    0
HETATM    8  O09 UNK A   1      -21.664 -22.398 -37.844 -0.71 -0.71    0
HETATM    9  OM0 UNK A   1      -21.192 -22.294 -39.996 -0.71 -0.71    0
HETATM   10  H01 UNK A   1      -23.580 -24.137 -38.897  0.25  0.25    0
HETATM   11  C11 UNK B   1      -24.243 -25.473 -44.327  0.08  0.08    0
HETATM   12  N12 UNK B   1      -23.745 -24.738 -43.173 -0.26 -0.26    0
HETATM   13  C13 UNK B   1      -24.360 -23.415 -43.127  0.11  0.11    0
HETATM   14  C14 UNK B   1      -23.405 -22.271 -42.811  0.53  0.53    0
HETATM   15  O15 UNK B   1      -23.847 -21.248 -42.302 -0.50 -0.50    0
HETATM   16  C16 UNK B   1      -24.994 -23.206 -44.496  0.00  0.00    0
HETATM   17  C17 UNK B   1      -25.317 -24.608 -44.974  0.04  0.04    0
HETATM   18  N19 UNK C   1      -22.113 -22.442 -43.092 -0.52 -0.52    0
HETATM   19  C20 UNK C   1      -21.078 -21.473 -42.761  0.21  0.21    0
HETATM   20  C21 UNK C   1      -19.752 -22.039 -43.265  0.53  0.53    0
HETATM   21  O22 UNK C   1      -19.725 -22.574 -44.377 -0.50 -0.50    0
HETATM   22  C23 UNK C   1      -21.399 -20.144 -43.461  0.04  0.04    0
HETATM   23  C24 UNK C   1      -20.493 -18.978 -43.151  0.01  0.01    0
HETATM   24  C25 UNK C   1      -19.454 -18.628 -44.032 -0.01 -0.01    0
HETATM   25  C26 UNK C   1      -20.694 -18.213 -41.989 -0.01 -0.01    0
HETATM   26  C27 UNK C   1      -18.625 -17.527 -43.754  0.00  0.00    0
HETATM   27  C28 UNK C   1      -19.868 -17.111 -41.705  0.00  0.00    0
HETATM   28  C29 UNK C   1      -18.832 -16.767 -42.590  0.00  0.00    0
HETATM   29  H18 UNK C   1      -21.803 -23.287 -43.547  0.25  0.25    0
HETATM   30  C30 UNK D   1      -17.409 -22.602 -42.850  0.08  0.08    0
HETATM   31  N31 UNK D   1      -18.663 -21.955 -42.490 -0.26 -0.26    0
HETATM   32  C32 UNK D   1      -18.730 -21.716 -41.053  0.11  0.11    0
HETATM   33  C33 UNK D   1      -18.383 -20.259 -40.737  0.53  0.53    0
HETATM   34  O34 UNK D   1      -17.736 -19.589 -41.538 -0.50 -0.50    0
HETATM   35  C35 UNK D   1      -17.695 -22.676 -40.473  0.00  0.00    0
HETATM   36  C36 UNK D   1      -16.639 -22.814 -41.554  0.04  0.04    0
HETATM   37  N38 UNK E   1      -18.782 -19.733 -39.579 -0.52 -0.52    0
HETATM   38  C39 UNK E   1      -19.377 -20.464 -38.473  0.21  0.21    0
HETATM   39  C40 UNK E   1      -20.125 -19.445 -37.610  0.53  0.53    0
HETATM   40  O41 UNK E   1      -19.697 -18.292 -37.530 -0.50 -0.50    0
HETATM   41  C42 UNK E   1      -18.217 -21.098 -37.697  0.04  0.04    0
HETATM   42  C43 UNK E   1      -18.545 -21.995 -36.528  0.01  0.01    0
HETATM   43  C44 UNK E   1      -18.943 -23.327 -36.739 -0.01 -0.01    0
HETATM   44  C45 UNK E   1      -18.414 -21.522 -35.211 -0.01 -0.01    0
HETATM   45  C46 UNK E   1      -19.218 -24.173 -35.651  0.00  0.00    0
HETATM   46  C47 UNK E   1      -18.684 -22.364 -34.118  0.00  0.00    0
HETATM   47  C48 UNK E   1      -19.088 -23.692 -34.338  0.00  0.00    0
HETATM   48  H37 UNK E   1      -18.630 -18.750 -39.408  0.25  0.25    0
HETATM   49  N50 UNK F   1      -21.228 -19.833 -36.971 -0.52 -0.52    0
HETATM   50  C51 UNK F   1      -21.983 -18.891 -36.151  0.20  0.20    0
HETATM   51  C52 UNK F   1      -23.448 -18.634 -36.558  0.53  0.53    0
HETATM   52  O53 UNK F   1      -24.063 -17.753 -35.962 -0.50 -0.50    0
HETATM   53  C54 UNK F   1      -21.998 -19.267 -34.650  0.03  0.03    0
HETATM   54  C55 UNK F   1      -22.730 -20.585 -34.317  0.02  0.02    0
HETATM   55  C56 UNK F   1      -20.563 -19.248 -34.112  0.00  0.00    0
HETATM   56  C57 UNK F   1      -22.069 -21.893 -34.756  0.00  0.00    0
HETATM   57  H49 UNK F   1      -21.555 -20.788 -37.119  0.25  0.25    0
HETATM   58  N59 UNK G   1      -24.086 -19.304 -37.521 -0.52 -0.52    0
HETATM   59  C60 UNK G   1      -23.614 -20.376 -38.376  0.21  0.21    0
HETATM   60  C61 UNK G   1      -24.829 -21.196 -38.817  0.53  0.53    0
HETATM   61  O62 UNK G   1      -25.658 -20.720 -39.592 -0.50 -0.50    0
HETATM   62  C63 UNK G   1      -22.866 -19.788 -39.586  0.04  0.04    0
HETATM   63  C64 UNK G   1      -23.407 -18.506 -40.185  0.01  0.01    0
HETATM   64  C65 UNK G   1      -24.342 -18.527 -41.235 -0.01 -0.01    0
HETATM   65  C66 UNK G   1      -22.957 -17.264 -39.699 -0.01 -0.01    0
HETATM   66  C67 UNK G   1      -24.823 -17.329 -41.791  0.00  0.00    0
HETATM   67  C68 UNK G   1      -23.434 -16.063 -40.250  0.00  0.00    0
HETATM   68  C69 UNK G   1      -24.369 -16.095 -41.297  0.00  0.00    0
HETATM   69  H58 UNK G   1      -25.032 -19.013 -37.711  0.25  0.25    0
HETATM   70  C70 UNK H   1      -24.456 -22.836 -37.004  0.08  0.08    0
HETATM   71  N71 UNK H   1      -24.984 -22.417 -38.291 -0.26 -0.26    0
HETATM   72  C72 UNK H   1      -25.947 -23.371 -38.818  0.11  0.11    0
HETATM   73  C73 UNK H   1      -25.270 -24.148 -39.951  0.53  0.53    0
HETATM   74  O74 UNK H   1      -25.903 -24.453 -40.957 -0.50 -0.50    0
HETATM   75  C75 UNK H   1      -26.295 -24.268 -37.632  0.00  0.00    0
HETATM   76  C76 UNK H   1      -25.090 -24.189 -36.703  0.04  0.04    0
CONECT    1    2   10   73
CONECT    2    1    3    5
CONECT    3    2   12
CONECT    3    4
CONECT    3    4
CONECT    4    3
CONECT    4    3
CONECT    5    2    6
CONECT    6    5    7
CONECT    7    6    9
CONECT    7    8
CONECT    7    8
CONECT    8    7
CONECT    8    7
CONECT    9    7
CONECT   10    1
CONECT   11   12   17
CONECT   12   11   13    3
CONECT   13   12   14   16
CONECT   14   13   18
CONECT   14   15
CONECT   14   15
CONECT   15   14
CONECT   15   14
CONECT   16   13   17
CONECT   17   11   16
CONECT   18   19   29   14
CONECT   19   18   20   22
CONECT   20   19   31
CONECT   20   21
CONECT   20   21
CONECT   21   20
CONECT   21   20
CONECT   22   19   23
CONECT   23   22   25
CONECT   23   24
CONECT   23   24
CONECT   24   26
CONECT   24   23
CONECT   24   23
CONECT   25   23
CONECT   25   27
CONECT   25   27
CONECT   26   24
CONECT   26   28
CONECT   26   28
CONECT   27   28
CONECT   27   25
CONECT   27   25
CONECT   28   27
CONECT   28   26
CONECT   28   26
CONECT   29   18
CONECT   30   31   36
CONECT   31   30   32   20
CONECT   32   31   33   35
CONECT   33   32   37
CONECT   33   34
CONECT   33   34
CONECT   34   33
CONECT   34   33
CONECT   35   32   36
CONECT   36   30   35
CONECT   37   38   48   33
CONECT   38   37   39   41
CONECT   39   38   49
CONECT   39   40
CONECT   39   40
CONECT   40   39
CONECT   40   39
CONECT   41   38   42
CONECT   42   41   44
CONECT   42   43
CONECT   42   43
CONECT   43   45
CONECT   43   42
CONECT   43   42
CONECT   44   42
CONECT   44   46
CONECT   44   46
CONECT   45   43
CONECT   45   47
CONECT   45   47
CONECT   46   47
CONECT   46   44
CONECT   46   44
CONECT   47   46
CONECT   47   45
CONECT   47   45
CONECT   48   37
CONECT   49   50   57   39
CONECT   50   49   51   53
CONECT   51   50   58
CONECT   51   52
CONECT   51   52
CONECT   52   51
CONECT   52   51
CONECT   53   50   54   55
CONECT   54   53   56
CONECT   55   53
CONECT   56   54
CONECT   57   49
CONECT   58   59   69   51
CONECT   59   58   60   62
CONECT   60   59   71
CONECT   60   61
CONECT   60   61
CONECT   61   60
CONECT   61   60
CONECT   62   59   63
CONECT   63   62   65
CONECT   63   64
CONECT   63   64
CONECT   64   66
CONECT   64   63
CONECT   64   63
CONECT   65   63
CONECT   65   67
CONECT   65   67
CONECT   66   64
CONECT   66   68
CONECT   66   68
CONECT   67   68
CONECT   67   65
CONECT   67   65
CONECT   68   67
CONECT   68   66
CONECT   68   66
CONECT   69   58
CONECT   70   71   76
CONECT   71   70   72   60
CONECT   72   71   73   75
CONECT   73   72    1
CONECT   73   74
CONECT   73   74
CONECT   74   73
CONECT   74   73
CONECT   75   72   76
CONECT   76   70   75
END
```

Appendix 4 : Example of a Molmol PDB file (extract)

```
HEADER    PROTEINASE INHIBITOR (TRYPSIN)          30-APR-92   1PIT
COMPND    TRYPSIN INHIBITOR
SOURCE    BOVINE (BOS TAURUS) PANCREAS
EXPDTA    NMR
AUTHOR    K.D.BERNDT,P.GUNTERT,L.P.M.ORBONS,K.WUTHRICH
JRNL        AUTH   K.D.BERNDT,P.GUNTERT,L.P.M.ORBONS,K.WUTHRICH
JRNL        TITL   DETERMINATION OF A HIGH-QUALITY NUCLEAR MAGNETIC
JRNL        TITL 2 RESONANCE SOLUTION STRUCTURE OF THE BOVINE
JRNL        TITL 3 PANCREATIC TRYPSIN INHIBITOR AND COMPARISON WITH
JRNL        TITL 4 THREE CRYSTAL STRUCTURES
JRNL        REF    J.MOL.BIOL.                   V. 227   757 1992
JRNL        REFN   ASTM JMOBAK  UK ISSN 0022-2836                  070
REMARK   1
REMARK   1 REFERENCE 1
REMARK   1  AUTH   G.WAGNER,W.BRAUN,T.F.HAVEL,T.SCHAUMANN,N.GO,
REMARK   1  AUTH 2 K.WUTHRICH
REMARK   1  TITL   PROTEIN STRUCTURES IN SOLUTION BY NUCLEAR
REMARK   1  TITL 2 MAGNETIC RESONANCE AND DISTANCE GEOMETRY:  THE
REMARK   1  TITL 3 POLYPEPTIDE FOLD OF THE BASIC PANCREATIC TRYPSIN
REMARK   1  TITL 4 INHIBITOR DETERMINED USING TWO DIFFERENT
REMARK   1  TITL 5 ALGORITHMS, DISGEO AND DISMAN
REMARK   1  REF    J.MOL.BIOL.                   V. 196   611 1987
REMARK   1  REFN   ASTM JMOBAK  UK ISSN 0022-2836                  070
REMARK   1 REFERENCE 2
REMARK   1  AUTH   G.WAGNER,K.WUTHRICH
REMARK   1  TITL   SEQUENTIAL RESONANCE ASSIGNMENTS IN PROTEIN 1H
REMARK   1  TITL 2 NUCLEAR MAGNETIC RESONANCE SPECTRA.  BASIC
REMARK   1  TITL 3 PANCREATIC TRYPSIN INHIBITOR
REMARK   1  REF    J.MOL.BIOL.                   V. 155   347 1982
REMARK   1  REFN   ASTM JMOBAK  UK ISSN 0022-2836                  070
REMARK   2
REMARK   2 RESOLUTION. NOT APPLICABLE.  SEE REMARK 4.
REMARK   3
REMARK   3 REFINEMENT.  NONE.
REMARK   3
REMARK   3 THREE-DIMENSIONAL STRUCTURE IN AQUEOUS SOLUTION AS
REMARK   3 DETERMINED BY NUCLEAR MAGNETIC RESONANCE AND DISTANCE
REMARK   3 GEOMETRY. DATA WERE COLLECTED AT PH 4.6, AND A TEMPERATURE
REMARK   3 OF 36 DEGREES CELSIUS. INPUT DATA CONSISTS OF 642 UPPER
REMARK   3 DISTANCE LIMIT CONSTRAINTS FROM NOE DATA; 41 PHI, 41 PSI,
REMARK   3 AND 33 CHI1 DIHEDRAL ANGLE CONSTRAINTS; 9 UPPER AND
REMARK   3 9 LOWER DISTANCE LIMIT CONSTRAINTS TO ENFORCE THE THREE
REMARK   3 DISULFIDE BONDS. THESE INPUT DATA ARE ALSO AVAILABLE
REMARK   3 FROM THE PROTEIN DATA BANK. A TOTAL OF 36 STEREOSPECIFIC
REMARK   3 PROTON RESONANCE ASSIGNMENTS WERE MADE.
REMARK   4
REMARK   4 THESE COORDINATES WERE GENERATED FROM SOLUTION NMR DATA.
REMARK   4 PROTEIN DATA BANK CONVENTIONS REQUIRE THAT *CRYST1* AND
REMARK   4 *SCALE* RECORDS BE INCLUDED, BUT THE VALUES ON THESE
REMARK   4 RECORDS ARE MEANINGLESS.
REMARK   5
REMARK   5 DISTANCE GEOMETRY CALCULATIONS WERE PERFORMED WITH THE
REMARK   5 PROGRAM DIANA (P.GUNTERT, W.BRAUN AND K.WUTHRICH,
REMARK   5 J.MOL.BIOL. (1991) VOL. 217, 517-530). FOR THE RESTRAINED
REMARK   5 ENERGY MINIMIZATION, A MODIFIED VERSION OF THE PROGRAM
REMARK   5 AMBER 3.0 (U.C.SINGH, P.K.WEINER, J.W.CALDWELL,P.A.KOLLMAN,
REMARK   5 UNIVERSITY OF CALIFORNIA, SAN FRANCISCO (1986)) WAS USED.
REMARK   5 FOR THE PRESENT STRUCTURES, THE NMR DISTANCE CONSTRAINTS
REMARK   5 WERE WEIGHTED SUCH THAT A VIOLATION OF AN UPPER DISTANCE
REMARK   5 LIMIT OF 0.2 ANGSTROM CORRESPONDS TO AN ENERGY OF KT/2 AND
REMARK   5 THE CONSTRAINTS ON DIHEDRAL ANGLES RESULTING FROM
REMARK   5 MEASUREMENT OF VICINAL COUPLING CONSTANTS WERE WEIGHTED
REMARK   5 SUCH THAT A VIOLATION OF 5 DEGREES CORRESPONDS TO AN ENERGY
REMARK   5 OF KT/2.
REMARK   6
REMARK   6 DEPOSITED COORDINATES ARE THOSE OF CONFORMERS 1 TO 20 OF
REMARK   6 REFERENCE JRNL WHICH ARE INDICATED WITH THE KEYWORD MODEL
REMARK   6 1 TO 20. THE AVERAGE VIOLATION OF THE NOE UPPER LIMIT
REMARK   6 DISTANCE CONSTRAINTS DERIVED FROM NOE DATA WAS 0.005
REMARK   6 ANGSTROMS PER CONSTRAINT FOR THE 20 CONFORMERS. THE AVERAGE
REMARK   6 VIOLATION OF THE DIHEDRAL ANGLE CONSTRAINTS WAS 0.05
REMARK   6 DEGREES PER CONSTRAINT FOR THE 20 CONFORMERS. THE AVERAGE
REMARK   6 MAXIMAL VIOLATION OF THE NOE UPPER DISTANCE LIMITS WAS 0.22
REMARK   6 ANGSTROMS IN THE 20 CONFORMERS. THE AVERAGE MAXIMAL
REMARK   6 VIOLATION OF THE DIHEDRAL ANGLE CONSTRAINTS WAS 2.2 DEGREES
REMARK   6 IN THE 20 CONFORMERS. THE AVERAGE ENERGY ACCORDING TO THE
REMARK   6 AMBER FORCE FIELD (S.J.WEINER, P.A.KOLLMAN, D.T.NGUYEN,
REMARK   6 D.A.CASE, J.COMP.CHEM. (1986) VOL. 7, 230-252) WAS -734
REMARK   6 KCAL/MOL.
REMARK   7
REMARK   7 ATOM NAMES HAVE BEEN ASSIGNED FOLLOWING THE RECOMMENDATIONS
REMARK   7 OF THE IUPAC-IUB COMMISSION AS PUBLISHED IN BIOCHEMISTRY
REMARK   7 (1970) VOL. 9, 3471-3479, EXCEPT THAT BACKBONE AMIDE
REMARK   7 HYDROGENS ARE DENOTED BY HN INSTEAD OF H. THE INDIVIDUAL
REMARK   7 NUMBERS OF THE HYDROGEN ATOMS IN METHYL AND METHYLENE
REMARK   7 GROUPS ARE INDICATED AS THE FIRST CHARACTER RATHER THAN
REMARK   7 THE LAST CHARACTER OF THE ATOM NAMES.
REMARK   7 IN THIS FILE THE AMINO ACID RESIDUES ARE NUMBERED
REMARK   7 CONSECUTIVELY FROM 1 TO 58.
REMARK   8
REMARK   8 PSEUDO-ATOMS DESIGNATED AS Q ARE DIMENSIONLESS REFERENCE
REMARK   8 POINTS REPRESENTING A GROUP OF HYDROGEN ATOMS. THEY ARE
REMARK   8 PLACED IN THE CENTER OF THE POSITIONS OF THE HYDROGEN ATOMS
REMARK   8 THEY REPRESENT. QA REPRESENTS THE TWO METHYLENE HYDROGEN
REMARK   8 ATOMS OF GLY. QB, QG, ... REPRESENT BETA, GAMMA, ...
REMARK   8 METHYLENE OR METHYL GROUPS IN THE SIDE CHAINS. IN CASE OF
REMARK   8 BRANCHES IN THE SIDE CHAINS THE NUMBERS OF THE PSEUDO-ATOMS
REMARK   8 ARE THE SAME AS THE NUMBERS OF THE CARBONS TO WHICH THE
REMARK   8 HYDROGEN  ATOMS ARE ATTACHED.
REMARK   8 QQG AND QQD DENOTE THE PSEUDO-ATOMS FOR THE 6 HYDROGEN
REMARK   8 ATOMS OF THE ISOPROPYL METHYL GROUPS OF VAL AND LEU.
REMARK   8 QR IS THE PSEUDO-ATOM FOR THE DELTA AND EPSILON HYDROGENS
REMARK   8 OF THE AROMATIC RINGS OF TYR AND PHE.
REMARK   8 (K.WUTHRICH, M.BILLETER AND W.BRAUN, J. MOL. BIOL. (1983)
REMARK   8 VOL. 169, 949-961)
REMARK   9
REMARK   9 THE AVERAGE OF THE RMSD VALUES TO THE MEAN OF THE 20 NMR
REMARK   9 CONFORMERS AS DESCRIBED IN REFERENCE JRNL IS 0.43 ANGSTROMS
REMARK   9 FOR THE HEAVY ATOMS OF THE BACKBONE OF RESIDUES 2-56 AND
REMARK   9 THE 28 BEST-DEFINED SIDECHAINS. THE CONFORMATIONS OF THE
REMARK   9 CHAIN TERMINI CONSISTING OF RESIDUES 1 AND 57-58 ARE LESS
REMARK   9 WELL DETERMINED. THE AVERAGE STRUCTURAL CHANGE DURING THE
REMARK   9 RESTRAINED AMBER REFINEMENT CORRESPONDS TO A RMSD OF 0.25
REMARK   9 ANGSTROMS FOR ALL HEAVY ATOMS. IN THE COLUMNS 55 TO 60 THE
REMARK   9 ENTRY 1.00 IDENTIFIES THE AMINO ACID RESIDUES THAT WERE
REMARK   9 USED IN THE CALCULATION OF THE GLOBAL RMSDS. FOR ALL OTHER
REMARK   9 RESIDUES THE ENTRY IS 0.00. NOTE: IN THE X-RAY CRYSTAL
REMARK   9 STRUCTURE FILES THESE COLUMNS CONTAIN THE OCCUPANCY VALUES.
REMARK  10
REMARK  10 IN THE COLUMNS 55 TO 60 THE ENTRY 1.00 IDENTIFIES THE
REMARK  10 AMINO ACID RESIDUES THAT WERE USED IN THE CALCULATION OF
REMARK  10 THE GLOBAL RMSD'S. FOR ALL OTHER RESIDUES THE ENTRY IS
REMARK  10 0.00.  NOTE: IN THE X-RAY CRYSTAL STRUCTURE FILES THESE
REMARK  10 COLUMNS CONTAIN THE OCCUPANCY VALUES.
REMARK  11
REMARK  11 AVERAGES OF THE ROOT-MEAN-SQUARE DEVIATIONS IN ANGSTROMS
REMARK  11 OF THE INDIVIDUAL ATOMS OF EACH CONFORMER RELATIVE TO THE
REMARK  11 19 OTHER CONFORMERS ARE LISTED IN THE COLUMNS 61 TO 66 OF
REMARK  11 THE ATOM RECORDS. THEY WERE OBTAINED AFTER THE BACKBONE OF
REMARK  11 RESIDUES 2-56 OF THE OTHER CONFORMERS HAD BEEN OPTIMALLY
REMARK  11 FIT TO THE CONFORMER FOR WHICH THE ATOMIC DEVIATION IS
REMARK  11 GIVEN.  NOTE: IN THE X-RAY CRYSTAL STRUCTURE FILES COLUMNS
REMARK  11 61 TO 66 CONTAIN THE TEMPERATURE FACTORS.
SEQRES   1     58  ARG PRO ASP PHE CYS LEU GLU PRO PRO TYR THR GLY PRO
SEQRES   2     58  CYS LYS ALA ARG ILE ILE ARG TYR PHE TYR ASN ALA LYS
SEQRES   3     58  ALA GLY LEU CYS GLN THR PHE VAL TYR GLY GLY CYS ARG
SEQRES   4     58  ALA LYS ARG ASN ASN PHE LYS SER ALA GLU ASP CYS MET
SEQRES   5     58  ARG THR CYS GLY GLY ALA
HELIX    1  H1 ASP      3  GLU      7  5 ALL DONORS,ACCEPTORS INCLUDED
HELIX    2  H2 SER     47  GLY     56  1 ALL DONORS,ACCEPTORS INCLUDED
SHEET    1  S1 3 LEU    29  TYR    35  0
SHEET    2  S1 3 ILE    18  ASN    24 -1  N  ILE    18   O  TYR    35
SHEET    3  S1 3 PHE    45  PHE    45 -1  N  PHE    45   O  TYR    21
SSBOND   1 CYS      5    CYS     55
SSBOND   2 CYS     14    CYS     38
SSBOND   3 CYS     30    CYS     51
SSBOND   4 CYS      5    CYS     55
SSBOND   5 CYS     14    CYS     38
SSBOND   6 CYS     30    CYS     51
SSBOND   7 CYS      5    CYS     55
SSBOND   8 CYS     14    CYS     38
SSBOND   9 CYS     30    CYS     51
SSBOND  10 CYS      5    CYS     55
SSBOND  11 CYS     14    CYS     38
SSBOND  12 CYS     30    CYS     51
SSBOND  13 CYS      5    CYS     55
SSBOND  14 CYS     14    CYS     38
SSBOND  15 CYS     30    CYS     51
SSBOND  16 CYS      5    CYS     55
SSBOND  17 CYS     14    CYS     38
SSBOND  18 CYS     30    CYS     51
SSBOND  19 CYS      5    CYS     55
SSBOND  20 CYS     14    CYS     38
SSBOND  21 CYS     30    CYS     51
SSBOND  22 CYS      5    CYS     55
SSBOND  23 CYS     14    CYS     38
SSBOND  24 CYS     30    CYS     51
SSBOND  25 CYS      5    CYS     55
SSBOND  26 CYS     14    CYS     38
SSBOND  27 CYS     30    CYS     51
SSBOND  28 CYS      5    CYS     55
SSBOND  29 CYS     14    CYS     38
SSBOND  30 CYS     30    CYS     51
SSBOND  31 CYS      5    CYS     55
SSBOND  32 CYS     14    CYS     38
SSBOND  33 CYS     30    CYS     51
SSBOND  34 CYS      5    CYS     55
SSBOND  35 CYS     14    CYS     38
SSBOND  36 CYS     30    CYS     51
SSBOND  37 CYS      5    CYS     55
SSBOND  38 CYS     14    CYS     38
SSBOND  39 CYS     30    CYS     51
SSBOND  40 CYS      5    CYS     55
SSBOND  41 CYS     14    CYS     38
SSBOND  42 CYS     30    CYS     51
SSBOND  43 CYS      5    CYS     55
SSBOND  44 CYS     14    CYS     38
SSBOND  45 CYS     30    CYS     51
SSBOND  46 CYS      5    CYS     55
SSBOND  47 CYS     14    CYS     38
SSBOND  48 CYS     30    CYS     51
SSBOND  49 CYS      5    CYS     55
SSBOND  50 CYS     14    CYS     38
SSBOND  51 CYS     30    CYS     51
SSBOND  52 CYS      5    CYS     55
SSBOND  53 CYS     14    CYS     38
SSBOND  54 CYS     30    CYS     51
SSBOND  55 CYS      5    CYS     55
SSBOND  56 CYS     14    CYS     38
SSBOND  57 CYS     30    CYS     51
CRYST1    1.000    1.000    1.000  90.00  90.00  90.00 P 1           1
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      1.000000  0.000000  0.000000        0.00000
SCALE2      0.000000  1.000000  0.000000        0.00000
SCALE3      0.000000  0.000000  1.000000        0.00000
MODEL        1
ATOM      1  N   ARG     1      -8.544   3.578  14.046  0.00  2.37
ATOM      2  CA  ARG     1      -7.776   3.484  12.790  0.00  1.79
ATOM      3  C   ARG     1      -8.492   4.333  11.742  0.00  1.59
ATOM      4  O   ARG     1      -9.713   4.432  11.844  0.00  1.71
ATOM      5  CB  ARG     1      -7.640   2.025  12.309  0.00  1.57
ATOM      6  CG  ARG     1      -8.996   1.382  11.955  0.00  2.49
ATOM      7  CD  ARG     1      -8.875  -0.093  11.554  0.00  2.14
ATOM      8  NE  ARG     1      -8.120  -0.280  10.303  0.00  1.94
ATOM      9  CZ  ARG     1      -7.792  -1.484   9.802  0.00  3.00
ATOM     10  NH1 ARG     1      -8.140  -2.588  10.477  0.00  3.63
ATOM     11  NH2 ARG     1      -7.139  -1.609   8.640  0.00  4.23
ATOM     12  H   ARG     1      -8.193   3.022  14.799  0.00  2.72
ATOM     13  HA  ARG     1      -6.780   3.877  12.993  0.00  1.81
ATOM     14 1HB  ARG     1       -6.995   2.010  11.428  0.00  1.97
ATOM     15 2HB  ARG     1       -7.148   1.441  13.089  0.00  1.67
ATOM     16 1HG  ARG     1       -9.660   1.426  12.819  0.00  3.58
ATOM     17 2HG  ARG     1       -9.473   1.918  11.131  0.00  3.34
ATOM     18 1HD  ARG     1       -8.402  -0.641  12.369  0.00  2.67
ATOM     19 2HD  ARG     1       -9.887  -0.479  11.411  0.00  3.27
ATOM     20  HE  ARG     1       -7.881   0.558   9.793  0.00  2.23
ATOM     21 1HH1 ARG     1       -8.653  -2.502  11.341  0.00  3.57
ATOM     22 2HH1 ARG     1       -7.954  -3.514  10.126  0.00  4.64
ATOM     23 1HH2 ARG     1       -6.921  -0.841   8.004  0.00  4.55
ATOM     24 2HH2 ARG     1       -6.870  -2.518   8.306  0.00  5.23
ATOM     25  QB  ARG     1       -7.072   1.726  12.259  0.00  1.50
ATOM     26  QG  ARG     1       -9.567   1.672  11.975  0.00  3.23
ATOM     27  QD  ARG     1       -9.145  -0.560  11.890  0.00  2.69
ATOM     28  QH1 ARG     1       -8.304  -3.008  10.734  0.00  3.99
ATOM     29  QH2 ARG     1       -6.896  -1.679   8.155  0.00  4.81
ATOM     30  N   PRO     2       -7.785   4.939  10.781  1.00  1.36
ATOM     31  CA  PRO     2       -8.423   5.636   9.681  1.00  1.20
ATOM     32  C   PRO     2       -9.076   4.653   8.705  1.00  0.99
ATOM     33  O   PRO     2       -8.809   3.451   8.734  1.00  1.08
ATOM     34  CB  PRO     2       -7.326   6.451   8.997  1.00  1.19
ATOM     35  CG  PRO     2       -6.002   5.839   9.465  1.00  1.26
ATOM     36  CD  PRO     2       -6.338   5.014  10.708  1.00  1.37
ATOM     37  HA  PRO     2       -9.189   6.315  10.062  1.00  1.32
ATOM     38 1HB  PRO     2       -7.420   6.411   7.912  1.00  1.04
ATOM     39 2HB  PRO     2       -7.394   7.491   9.312  1.00  1.37
ATOM     40 1HG  PRO     2       -5.595   5.206   8.677  1.00  1.16
ATOM     41 2HG  PRO     2       -5.276   6.613   9.713  1.00  1.40
ATOM     42 1HD  PRO     2       -5.903   4.019  10.633  1.00  1.29
ATOM     43 2HD  PRO     2       -5.944   5.525  11.587  1.00  1.60
ATOM     44  QB  PRO     2       -7.407   6.951   8.612  1.00  1.19
ATOM     45  QG  PRO     2       -5.436   5.910   9.195  1.00  1.28
ATOM     46  QD  PRO     2       -5.924   4.772  11.110  1.00  1.44
ATOM     47  N   ASP     3       -9.901   5.213   7.821  1.00  0.81
ATOM     48  CA  ASP     3      -10.537   4.606   6.666  1.00  0.70
ATOM     49  C   ASP     3       -9.492   4.188   5.635  1.00  0.47
ATOM     50  O   ASP     3       -9.551   3.094   5.075  1.00  0.44
ATOM     51  CB  ASP     3      -11.475   5.657   6.043  1.00  0.84
ATOM     52  CG  ASP     3      -10.718   6.888   5.535  1.00  2.69
ATOM     53  OD1 ASP     3       -9.785   7.311   6.262  1.00  3.89
ATOM     54  OD2 ASP     3      -11.014   7.332   4.408  1.00  3.84
ATOM     55  H   ASP     3       -9.936   6.228   7.778  1.00  0.79
ATOM     56  HA  ASP     3      -11.099   3.731   6.987  1.00  0.98
ATOM     57 1HB  ASP     3      -11.999   5.202   5.202  1.00  1.88
ATOM     58 2HB  ASP     3      -12.214   5.975   6.779  1.00  1.50
ATOM     59  QB  ASP     3      -12.107   5.589   5.991  1.00  0.99
ATOM     60  N   PHE     4       -8.511   5.055   5.386  1.00  0.50
ATOM     61  CA  PHE     4       -7.520   4.803   4.349  1.00  0.58
ATOM     62  C   PHE     4       -6.738   3.520   4.651  1.00  0.54
ATOM     63  O   PHE     4       -6.261   2.838   3.747  1.00  0.59
ATOM     64  CB  PHE     4       -6.634   6.035   4.141  1.00  0.83
ATOM     65  CG  PHE     4       -5.598   6.301   5.215  1.00  0.99
ATOM     66  CD1 PHE     4       -4.460   5.477   5.302  1.00  1.15
ATOM     67  CD2 PHE     4       -5.738   7.389   6.096  1.00  1.06
ATOM     68  CE1 PHE     4       -3.560   5.631   6.368  1.00  1.36
ATOM     69  CE2 PHE     4       -4.785   7.597   7.109  1.00  1.28
ATOM     70  CZ  PHE     4       -3.733   6.682   7.282  1.00  1.43
```

Appendix 5 : Amino acid atom description table A

(Table extract from)

Please note that some atoms are in the wrong order. You must check in a correct PDB file the true order. I used this table to fin thryptophian information data (I not sur the hydrogens order in my program, I did like tyrose), and you can notice that the first four atom are not N, C, CA, O as they should be.

Amino acid	IUPAC notation	Stereoisomeric information	PDB notation
ALA	H		H
ALA	HA		HA
ALA	HB1		1HB
ALA	HB2		2HB
ALA	HB3		3HB
ALA	C		C
ALA	CA		CA
ALA	CB		CB
ALA	N		N
ALA	O		O
ARG	H		H
ARG	HA		HA
ARG	HB2	(pro-R)	2HB
ARG	HB3	(pro-S)	3HB
ARG	HG2	(pro-S)	2HG
ARG	HG3	(pro-R)	3HG
ARG	HD2	(pro-S)	2HD
ARG	HD3	(pro-R)	3HD
ARG	HE		HE
ARG	HH11	(Z)	1HH1
ARG	HH12	(E)	2HH1
ARG	HH21	(Z)	1HH2
ARG	HH22	(E)	2HH2
ARG	C		C
ARG	CA		CA
ARG	CB		CB

ARG	CG		CG
ARG	CD		CD
ARG	CZ		CZ
ARG	N		N
ARG	NE		NE
ARG	NH1	(Z)	NH1
ARG	NH2	(E)	NH2
ARG	O		O
ASP	H		H
ASP	HA		HA
ASP	HB2	(pro-S)	2HB
ASP	HB3	(pro-R)	3HB
ASP	HD2		HD2
ASP	C		C
ASP	CA		CA
ASP	CB		CB
ASP	CG		CG
ASP	N		N
ASP	O		O
ASP	OD1		OD1
ASP	OD2		OD2
ASN	H		H
ASN	HA		HA
ASN	HB2	(pro-S)	2HB
ASN	HB3	(pro-R)	3HB
ASN	HD21	(E)	1HD2
ASN	HD22	(Z)	2HD2
ASN	C		C
ASN	CA		CA
ASN	CB		CB
ASN	CG		CG
ASN	N		N
ASN	ND2		ND2
ASN	O		O
ASN	OD1		OD1
CYS	H		H
CYS	HA		HA
CYS	HB2	(pro-S)	2HB
CYS	HB3	(pro-R)	3HB
CYS	HG		HG

CYS	C		C
CYS	CA		CA
CYS	CB		CB
CYS	N		N
CYS	O		O
CYS	SG		SG
GLU	H		H
GLU	HA		HA
GLU	HB2	(pro-R)	2HB
GLU	HB3	(pro-S)	3HB
GLU	HG2	(pro-S)	2HG
GLU	HG3	(pro-R)	3HG
GLU	HE2		HE2
GLU	C		C
GLU	CA		CA
GLU	CB		CB
GLU	CG		CG
GLU	CD		CD
GLU	N		N
GLU	O		O
GLU	OE1		OE1
GLU	OE2		OE2
GLN	H		H
GLN	HA		HA
GLN	HB2	(pro-R)	2HB
GLN	HB3	(pro-S)	3HB
GLN	HG2	(pro-S)	2HG
GLN	HG3	(pro-R)	3HG
GLN	HE21	(E)	1HE2
GLN	HE22	(Z)	2HE2
GLN	C		C
GLN	CA		CA
GLN	CB		CB
GLN	CG		CG
GLN	CD		CD
GLN	N		N
GLN	NE2		NE2
GLN	O		O
GLN	OE1		OE1
GLY	H		H

GLY	HA2	(pro-R)	2HA
GLY	HA3	(pro-S)	3HA
GLY	C		C
GLY	CA		CA
GLY	N		N
GLY	O		O
HIS	H		H
HIS	HA		HA
HIS	HB2	(pro-S)	2HB
HIS	HB3	(pro-R)	3HB
HIS	HD1		HD1
HIS	HD2		HD2
HIS	HE1		HE1
HIS	HE2		HE2
HIS	C		C
HIS	CA		CA
HIS	CB		CB
HIS	CG		CG
HIS	CD2		CD2
HIS	CE1		CE1
HIS	N		N
HIS	ND1		ND1
HIS	NE2		NE2
HIS	O		O
ILE	H		H
ILE	HA		HA
ILE	HB		HB
ILE	HG12	(pro-R)	2HG1
ILE	HG13	(pro-S)	3HG1
ILE	HG21		1HG2
ILE	HG22		2HG2
ILE	HG23		3HG2
ILE	HD11		1HD1
ILE	HD12		2HD1
ILE	HD13		3HD1
ILE	C		C
ILE	CA		CA
ILE	CB		CB
ILE	CG1		CG1
ILE	CG2		CG2

ILE	CD1		CD1
ILE	N		N
ILE	O		O
LEU	H		H
LEU	HA		HA
LEU	HB2	(pro-R)	2HB
LEU	HB3	(pro-S)	3HB
LEU	HG		HG
LEU	HD11		1HD1
LEU	HD12		2HD1
LEU	HD13		3HD1
LEU	HD21		1HD2
LEU	HD22		2HD2
LEU	HD23		3HD2
LEU	C		C
LEU	CA		CA
LEU	CB		CB
LEU	CG		CG
LEU	CD1	(pro-R)	CD1
LEU	CD2	(pro-S)	CD2
LEU	N		N
LEU	O		O
LYS	H		H
LYS	HA		HA
LYS	HB2	(pro-R)	2HB
LYS	HB3	(pro-S)	3HB
LYS	HG2	(pro-R)	2HG
LYS	HG3	(pro-S)	3HG
LYS	HD2	(pro-S)	2HD
LYS	HD3	(pro-R)	3HD
LYS	HE2	(pro-S)	2HE
LYS	HE3	(pro-R)	3HE
LYS	HZ1		1HZ
LYS	HZ2		2HZ
LYS	HZ3		3HZ
LYS	C		C
LYS	CA		CA
LYS	CB		CB
LYS	CG		CG
LYS	CD		CD

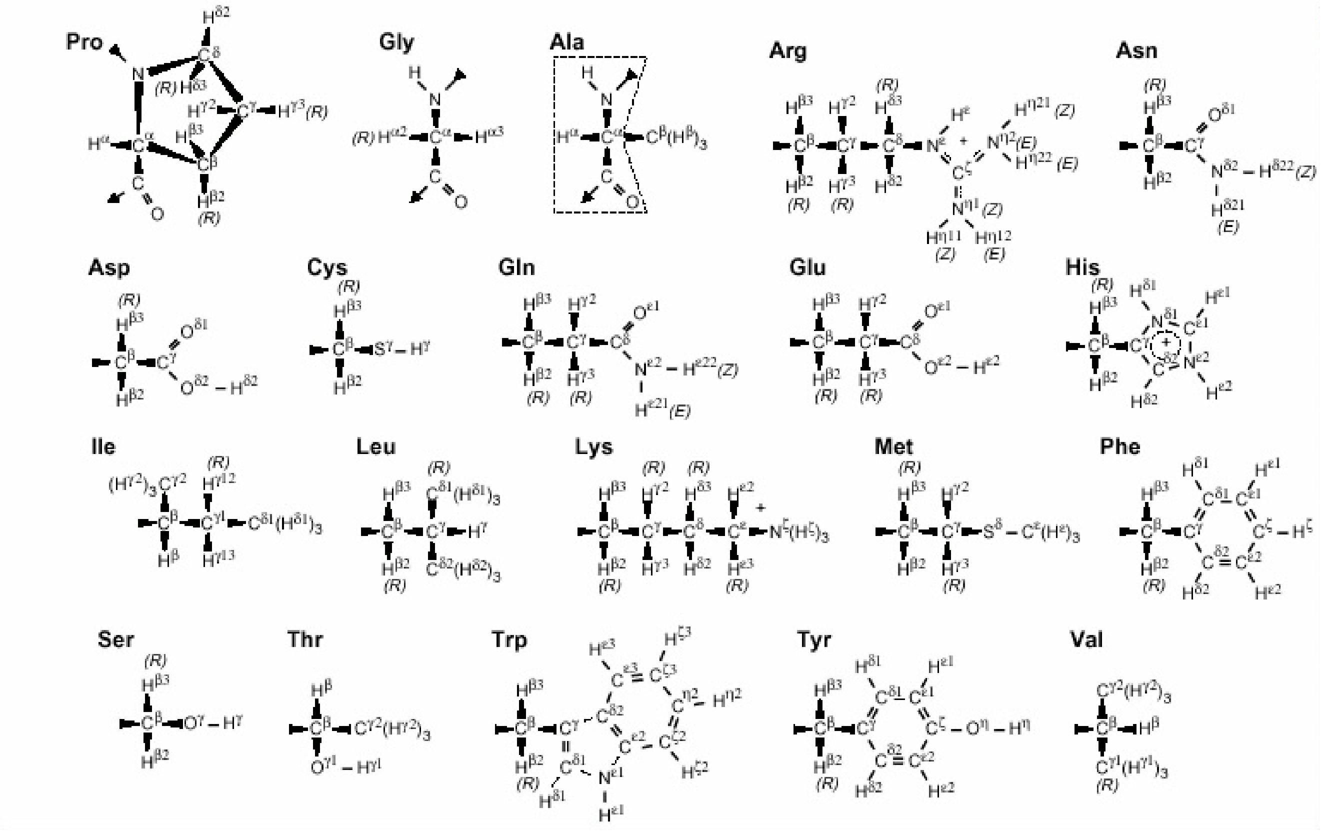
LYS	CE		CE
LYS	N		N
LYS	NZ		NZ
LYS	O		O
MET	H		H
MET	HA		HA
MET	HB2	(pro-S)	2HB
MET	HB3	(pro-R)	3HB
MET	HG2	(pro-S)	2HG
MET	HG3	(pro-R)	3HG
MET	HE1		1HE
MET	HE2		2HE
MET	HE3		3HE
MET	C		C
MET	CA		CA
MET	CB		CB
MET	CG		CG
MET	CE		CE
MET	N		N
MET	O		O
MET	SD		SD
PHE	H		H
PHE	HA		HA
PHE	HB2	(pro-R)	1HB
PHE	HB3	(pro-S)	2HB
PHE	HD1		HD1
PHE	HD2		HD2
PHE	HE1		HE1
PHE	HE2		HE2
PHE	HZ		HZ
PHE	C		C
PHE	CA		CA
PHE	CB		CB
PHE	CG		CG
PHE	CD1		CD1
PHE	CD2		CD2
PHE	CE1		CE1
PHE	CE2		CE2
PHE	CZ		CZ
PHE	N		N

PHE	O		O
PRO	H2	(pro-R)	H2
PRO	H3	(pro-S)	H3
PRO	HA		HA
PRO	HB2	(pro-R)	2HB
PRO	HB3	(pro-S)	3HB
PRO	HG2	(pro-S)	2HG
PRO	HG3	(pro-R)	3HG
PRO	HD2	(pro-S)	2HD
PRO	HD3	(pro-R)	3HD
PRO	C		C
PRO	CA		CA
PRO	CB		CB
PRO	CG		CG
PRO	CD		CD
PRO	N		N
PRO	O		O
SER	H		H
SER	HA		HA
SER	HB2	(pro-S)	2HB
SER	HB3	(pro-R)	3HB
SER	HG		HG
SER	C		C
SER	CA		CA
SER	CB		CB
SER	N		N
SER	O		O
SER	OG		OG
THR	H		H
THR	HA		HA
THR	HB		HB
THR	HG1		HG1
THR	HG21		1HG2
THR	HG22		2HG2
THR	HG23		3HG2
THR	C		C
THR	CA		CA
THR	CB		CB
THR	CG2		CG2
THR	N		N

THR	O		O
THR	OG1		OG1
TRP	H		H
TRP	HA		HA
TRP	HB2	(pro-R)	2HB
TRP	HB3	(pro-S)	3HB
TRP	HD1		HD1
TRP	HE1		HE1
TRP	HE3		HE3
TRP	HZ2		HZ2
TRP	HZ3		HZ3
TRP	HH2		HH2
TRP	C		C
TRP	CA		CA
TRP	CB		CB
TRP	CG		CG
TRP	CD1		CD1
TRP	CD2		CD2
TRP	CE2		CE2
TRP	CE3		CE3
TRP	CZ2		CZ2
TRP	CZ3		CZ3
TRP	CH2		CH2
TRP	N		N
TRP	NE1		NE1
TRP	O		O
TYR	H		H
TYR	HA		HA
TYR	HB2	(pro-R)	2HB
TYR	HB3	(pro-S)	3HB
TYR	HD1		HD1
TYR	HD2		HD2
TYR	HE1		HE1
TYR	HE2		HE2
TYR	HH		HH
TYR	C		C
TYR	CA		CA
TYR	CB		CB
TYR	CG		CG
TYR	CD1		CD1

TYR	CD2		CD2
TYR	CE1		CE1
TYR	CE2		CE2
TYR	CZ		CZ
TYR	N		N
TYR	O		O
TYR	OH		OH
VAL	H		H
VAL	HA		HA
VAL	HB		HB
VAL	HG11		1HG1
VAL	HG12		2HG1
VAL	HG13		3HG1
VAL	HG21		1HG2
VAL	HG22		2HG2
VAL	HG23		3HG2
VAL	C		C
VAL	CA		CA
VAL	CB		CB
VAL	CG1	(pro-R)	CG1
VAL	CG2	(pro-S)	CG2
VAL	N		N
VAL	O		O

Appendix 6 : Amino acid IUPAC notation



Appendix 7 : Amino acid atom description table B

(Table extract from 1pit.pdb file)

This file was supplied with Molmol. As it is Molmol example file, Molmol reads it.

So I used this table until I found another better. It's in fact part of the file, it's a big proteine description, the differents amino acids occurs often, I extracted the information when they appear for the first time in the file.

Note there is a difference in proline description order with Molmol PDB file. I assumed that the first carbon (before the nitrogen) was the CD but I'm not sure of this.

```
ATOM     30  N   PRO     2       -7.785   4.939  10.781  1.00  1.36
ATOM     31  CA  PRO     2       -8.423   5.636   9.681  1.00  1.20
ATOM     32  C   PRO     2       -9.076   4.653   8.705  1.00  0.99
ATOM     33  O   PRO     2       -8.809   3.451   8.734  1.00  1.08
ATOM     34  CB  PRO     2       -7.326   6.451   8.997  1.00  1.19
ATOM     35  CG  PRO     2       -6.002   5.839   9.465  1.00  1.26
ATOM     36  CD  PRO     2       -6.338   5.014  10.708  1.00  1.37
ATOM     37  HA  PRO     2       -9.189   6.315  10.062  1.00  1.32
ATOM     38 1HB  PRO     2       -7.420   6.411   7.912  1.00  1.04
ATOM     39 2HB  PRO     2       -7.394   7.491   9.312  1.00  1.37
ATOM     40 1HG  PRO     2       -5.595   5.206   8.677  1.00  1.16
ATOM     41 2HG  PRO     2       -5.276   6.613   9.713  1.00  1.40
ATOM     42 1HD  PRO     2       -5.903   4.019  10.633  1.00  1.29
ATOM     43 2HD  PRO     2       -5.944   5.525  11.587  1.00  1.60
ATOM     44  QB  PRO     2       -7.407   6.951   8.612  1.00  1.19
ATOM     45  QG  PRO     2       -5.436   5.910   9.195  1.00  1.28
ATOM     46  QD  PRO     2       -5.924   4.772  11.110  1.00  1.44
ATOM     60  N   PHE     4       -8.511   5.055   5.386  1.00  0.50
ATOM     61  CA  PHE     4       -7.520   4.803   4.349  1.00  0.58
ATOM     62  C   PHE     4       -6.738   3.520   4.651  1.00  0.54
ATOM     63  O   PHE     4       -6.261   2.838   3.747  1.00  0.59
ATOM     64  CB  PHE     4       -6.634   6.035   4.141  1.00  0.83
ATOM     65  CG  PHE     4       -5.598   6.301   5.215  1.00  0.99
ATOM     66  CD1 PHE     4       -4.460   5.477   5.302  1.00  1.15
ATOM     67  CD2 PHE     4       -5.738   7.389   6.096  1.00  1.06
ATOM     68  CE1 PHE     4       -3.560   5.631   6.368  1.00  1.36
ATOM     69  CE2 PHE     4       -4.785   7.597   7.109  1.00  1.28
ATOM     70  CZ  PHE     4       -3.733   6.682   7.282  1.00  1.43
ATOM     71  H   PHE     4       -8.582   5.967   5.843  1.00  0.59
ATOM     72  HA  PHE     4       -8.061   4.644   3.413  1.00  0.63
ATOM     73 1HB  PHE     4       -6.115   5.888   3.195  1.00  0.97
ATOM     74 2HB  PHE     4       -7.291   6.899   4.040  1.00  0.85
ATOM     75  HD1 PHE     4       -4.299   4.684   4.585  1.00  1.15
ATOM     76  HD2 PHE     4      -6.606   8.034   6.044  1.00  0.98
ATOM     77  HE1 PHE     4      -2.758   4.922   6.499  1.00  1.53
ATOM     78  HE2 PHE     4      -4.915   8.405   7.815  1.00  1.36
ATOM     79  HZ  PHE     4      -3.104   6.739   8.159  1.00  1.63
ATOM     80  QB  PHE     4      -6.703   6.394   3.618  1.00  0.90
ATOM     81  QR  PHE     4      -4.645   6.512   6.236  1.00  1.21
ATOM     93  N   LEU     6      -7.870   0.824   5.047  1.00  0.38
ATOM     94  CA  LEU     6      -8.700  -0.238   4.492  1.00  0.47
ATOM     95  C   LEU     6      -8.865  -0.089   2.962  1.00  0.51
ATOM     96  O   LEU     6      -9.617  -0.859   2.365  1.00  0.66
ATOM     97  CB  LEU     6     -10.067  -0.223   5.211  1.00  0.56
ATOM     98  CG  LEU     6      -9.972  -0.264   6.755  1.00  0.57
ATOM     99  CD1 LEU     6     -10.554   0.995   7.396  1.00  0.68
ATOM    100  CD2 LEU     6     -10.651  -1.499   7.345  1.00  0.65
ATOM    101  H   LEU     6      -8.215   1.764   4.879  1.00  0.48
ATOM    102  HA  LEU     6      -8.234  -1.209   4.666  1.00  0.55
ATOM    103 1HB  LEU     6     -10.606   0.675   4.909  1.00  0.56
ATOM    104 2HB  LEU     6     -10.647  -1.079   4.865  1.00  0.74
ATOM    105  HG  LEU     6      -8.939  -0.299   7.072  1.00  0.60
ATOM    106 1HD1 LEU     6     -11.560   1.187   7.024  1.00  1.94
ATOM    107 2HD1 LEU     6     -10.578   0.902   8.482  1.00  1.65
ATOM    108 3HD1 LEU     6      -9.903   1.826   7.143  1.00  1.41
ATOM    109 1HD2 LEU     6     -10.292  -2.387   6.833  1.00  1.67
ATOM    110 2HD2 LEU     6     -10.410  -1.579   8.406  1.00  1.07
ATOM    111 3HD2 LEU     6     -11.729  -1.424   7.221  1.00  1.67
ATOM    112  QB  LEU     6     -10.627  -0.202   4.887  1.00  0.63
ATOM    113  QD1 LEU     6     -10.680   1.305   7.550  1.00  0.72
ATOM    114  QD2 LEU     6     -10.810  -1.796   7.487  1.00  0.69
ATOM    115  QQD LEU     6     -10.745  -0.245   7.518  1.00  0.64
ATOM    116  N   GLU     7      -8.182   0.873   2.316  1.00  0.47
ATOM    117  CA  GLU     7      -8.200   1.028   0.853  1.00  0.52
ATOM    118  C   GLU     7      -7.653  -0.242   0.164  1.00  0.48
ATOM    119  O   GLU     7      -6.879  -0.979   0.778  1.00  0.66
ATOM    120  CB  GLU     7      -7.355   2.257   0.447  1.00  0.69
ATOM    121  CG  GLU     7      -8.159   3.487   0.003  1.00  1.58
ATOM    122  CD  GLU     7      -7.277   4.706  -0.280  1.00  1.66
ATOM    123  OE1 GLU     7      -6.692   5.235   0.691  1.00  2.58
ATOM    124  OE2 GLU     7      -7.211   5.114  -1.462  1.00  2.65
ATOM    125  H   GLU     7      -7.556   1.471   2.843  1.00  0.44
ATOM    126  HA  GLU     7      -9.236   1.183   0.547  1.00  0.74
ATOM    127 1HB  GLU     7      -6.777   2.589   1.302  1.00  1.99
ATOM    128 2HB  GLU     7      -6.661   1.976  -0.346  1.00  1.99
ATOM    129 1HG  GLU     7      -8.745   3.259  -0.886  1.00  2.88
ATOM    130 2HG  GLU     7      -8.829   3.749   0.814  1.00  2.83
ATOM    131  QB  GLU     7      -6.719   2.283   0.478  1.00  1.45
ATOM    132  QG  GLU     7      -8.787   3.504  -0.036  1.00  2.47
ATOM    167  N   TYR    10      -4.346  -0.035  -5.128  1.00  0.56
ATOM    168  CA  TYR    10      -4.479   0.631  -6.417  1.00  0.60
ATOM    169  C   TYR    10      -3.121   0.838  -7.111  1.00  0.64
ATOM    170  O   TYR    10      -2.390   1.782  -6.823  1.00  0.80
ATOM    171  CB  TYR    10      -5.226   1.951  -6.225  1.00  0.64
ATOM    172  CG  TYR    10      -5.704   2.573  -7.520  1.00  0.99
ATOM    173  CD1 TYR    10      -6.790   1.997  -8.206  1.00  1.67
ATOM    174  CD2 TYR    10      -5.072   3.715  -8.043  1.00  1.06
ATOM    175  CE1 TYR    10      -7.242   2.559  -9.411  1.00  2.21
ATOM    176  CE2 TYR    10      -5.564   4.310  -9.218  1.00  1.54
ATOM    177  CZ  TYR    10      -6.629   3.718  -9.917  1.00  2.08
ATOM    178  OH  TYR    10      -7.059   4.274 -11.084  1.00  2.65
ATOM    179  H   TYR    10      -4.411   0.520  -4.282  1.00  0.64
ATOM    180  HA  TYR    10      -5.105   0.001  -7.052  1.00  0.64
ATOM    181 1HB  TYR    10      -6.101   1.767  -5.599  1.00  0.73
ATOM    182 2HB  TYR    10      -4.572   2.641  -5.695  1.00  0.76
ATOM    183  HD1 TYR    10      -7.275   1.115  -7.811  1.00  1.89
ATOM    184  HD2 TYR    10      -4.208   4.138  -7.550  1.00  1.08
ATOM    185  HE1 TYR    10      -8.065   2.096  -9.937  1.00  2.81
ATOM    186  HE2 TYR    10      -5.089   5.194  -9.614  1.00  1.69
ATOM    187  HH  TYR    10      -7.786   3.791 -11.484  1.00  3.19
ATOM    188  QB  TYR    10      -5.337   2.204  -5.647  1.00  0.64
ATOM    189  QR  TYR    10      -6.159   3.136  -8.728  1.00  1.51
ATOM    308  N   ILE    18       6.444  -2.052  -9.552  1.00  0.61
ATOM    309  CA  ILE    18       7.301  -1.612  -8.466  1.00  0.62
ATOM    310  C   ILE    18       6.768  -2.279  -7.202  1.00  0.52
ATOM    311  O   ILE    18       5.615  -2.063  -6.845  1.00  0.46
ATOM    312  CB  ILE    18       7.222  -0.075  -8.340  1.00  0.63
ATOM    313  CG1 ILE    18       7.582   0.606  -9.671  1.00  0.71
ATOM    314  CG2 ILE    18       8.134   0.430  -7.214  1.00  0.66
ATOM    315  CD1 ILE    18       7.572   2.136  -9.579  1.00  0.72
ATOM    316  H   ILE    18       5.449  -1.921  -9.408  1.00  0.72
ATOM    317  HA  ILE    18       8.335  -1.914  -8.642  1.00  0.69
ATOM    318  HB  ILE    18       6.196   0.200  -8.091  1.00  0.57
ATOM    319 1HG1 ILE    18       8.563   0.268 -10.002  1.00  0.80
ATOM    320 2HG1 ILE    18       6.845   0.324 -10.421  1.00  0.82
ATOM    321 1HG2 ILE    18       7.974  -0.136  -6.298  1.00  1.56
ATOM    322 2HG2 ILE    18       9.178   0.339  -7.514  1.00  1.86
ATOM    323 3HG2 ILE    18       7.904   1.474  -7.001  1.00  1.31
ATOM    324 1HD1 ILE    18       6.641   2.472  -9.122  1.00  1.57
ATOM    325 2HD1 ILE    18       8.419   2.490  -8.992  1.00  1.49
ATOM    326 3HD1 ILE    18       7.646   2.557 -10.582  1.00  1.76
ATOM    327  QG1 ILE    18       7.704   0.296 -10.211  1.00  0.78
ATOM    328  QG2 ILE    18       8.352   0.559  -6.937  1.00  0.67
ATOM    329  QD1 ILE    18       7.569   2.507  -9.565  1.00  0.74
ATOM    330  N   ILE    19       7.577  -3.085  -6.516  1.00  0.51
ATOM    331  CA  ILE    19       7.174  -3.654  -5.240  1.00  0.42
ATOM    332  C   ILE    19       6.969  -2.504  -4.243  1.00  0.41
ATOM    333  O   ILE    19       7.883  -1.711  -4.011  1.00  0.52
ATOM    334  CB  ILE    19       8.215  -4.680  -4.749  1.00  0.46
ATOM    335  CG1 ILE    19       8.571  -5.744  -5.809  1.00  0.50
ATOM    336  CG2 ILE    19       7.736  -5.355  -3.455  1.00  0.46
ATOM    337  CD1 ILE    19       7.423  -6.688  -6.188  1.00  0.50
ATOM    338  H   ILE    19       8.515  -3.248  -6.839  1.00  0.57
ATOM    339  HA  ILE    19       6.225  -4.167  -5.383  1.00  0.39
ATOM    340  HB  ILE    19       9.134  -4.137  -4.525  1.00  0.49
ATOM    341 1HG1 ILE    19       8.926  -5.256  -6.717  1.00  0.57
ATOM    342 2HG1 ILE    19       9.393  -6.353  -5.430  1.00  0.56
ATOM    343 1HG2 ILE    19       7.574  -4.611  -2.674  1.00  1.35
ATOM    344 2HG2 ILE    19       6.802  -5.890  -3.627  1.00  1.76
ATOM    345 3HG2 ILE    19       8.494  -6.058  -3.106  1.00  1.71
ATOM    346 1HD1 ILE    19       6.493  -6.144  -6.339  1.00  1.33
ATOM    347 2HD1 ILE    19       7.678  -7.200  -7.116  1.00  1.70
ATOM    348 3HD1 ILE    19       7.282  -7.437  -5.408  1.00  1.64
ATOM    349  QG1 ILE    19       9.159  -5.804  -6.073  1.00  0.55
ATOM    350  QG2 ILE    19       7.623  -5.519  -3.135  1.00  0.46
ATOM    351  QD1 ILE    19       7.151  -6.927  -6.287  1.00  0.52
```

Appendix 8 : Basic instruction syntax

CLOSE

Closes one or more open files or devices.

CLOSE [[#]filenumber%[,[#]filenumber%]...]

- **filenumber%** The number of an open file or device.

- CLOSE with no arguments closes all open files and devices.

DATA

DATA specifies values to be read by subsequent READ statements.

DATA constant[,constant]...

- constant One or more numeric or string constants specifying the data to be read. String constants containing commas, colons, or leading or trailing spaces are enclosed in quotation marks (" ").

DIM

DIM declares an array or specifies a data type for a nonarray variable.

DIM variable[(subscripts)] [AS type]
[,variable[(subscripts)] [AS type]]...

- **variable** The name of an array or variable.
- **subscripts** Dimensions of the array, expressed as follows:

[lower TO] upper [,[lower TO] upper]...

• **lower**	The lower bound of the array's subscripts. The default lower bound is zero.
• **upper**	The upper bound.
• **AS type**	Declares the data type of the array or variable (INTEGER, LONG, SINGLE, DOUBLE, STRING, or a user-defined data type).

DO

Repeats a block of statements while a condition is true or until a condition becomes true.

```
DO [{WHILE | UNTIL} condition]
   [statementblock]
LOOP

DO
   [statementblock]
LOOP [{WHILE | UNTIL} condition]
```

• **condition**	A numeric expression that Basic evaluates as true (nonzero) or false (zero).

END IF

(see IF)

FOR

Repeats a block of statements a specified number of times.
(The loop begins by FOR and ends by NEXT)

```
FOR counter = start TO end [STEP increment]
   [statementblock]
NEXT [counter [,counter]...]
```

• **counter**	A numeric variable used as the loop counter.
• **start and end**	The initial and final values of the counter.

• **increment**	The amount the counter is changed each time through the loop.

IF

Executes a statement or statement block depending on specified conditions.

IF condition1 **THEN**
   [statementblock]
END IF

IF condition **THEN** statements [**ELSE** statements]

• **condition**	Any expression that can be evaluated as
• **statementblock**	One or more statements on one or more lines.

LINE INPUT

LINE INPUT reads a line of up to 255 characters from a file.
LINE INPUT reads all characters up to a carriage return.

LINE INPUT #filenumber%, variable$

• **variable$**	Holds a line of characters or read from a file.
• **filenumber%**	The number of an open file.

LOOP

(see DO)

MID$

MID$ replaces part of a string variable with another string.

**MID$**(stringvariable$, start%) = stringexpression$

- **stringexpression$** The string from which the MID$ function returns a substring, or the replacement string used by the MID$ statement. It can be any string expression.
- **start%** The position of the first character in the substring being returned or replaced.
- **stringvariable$** The string variable being modified by the MID$ statement.

MID$ (function)

MID$ function returns part of a string (a substring).

**MID$**(stringexpression$,start%[,length%])

- **stringexpression$** The string from which the MID$ function returns a substring, or the replacement string used by the MID$ statement. It can be any string expression.
- **start%** The position of the first character in the substring being returned or replaced.
- **length%** The number of characters in the substring. If the length is omitted, MID$ returns or replaces all characters to the right of the start position.

NEXT

(see FOR)

OPEN

OPEN mode2$,[#]filenum%,file$[,reclen%]

- **mode2$** A string expression that begins with one of the following characters and specifies the file mode:

  - **O** Sequential output mode.
  - **I** Sequential input mode.

- **filenum%** A number in the range 1 through 255 that identifies the file while it is open.
- **file$** The name of the file (may include drive and path).

PRINT

PRINT writes data to a file.

PRINT #filenumber%, expressionlist

- **filenumber%** The number of an open file. If you don't specify a file number, PRINT writes to the screen.
- **expressionlist** A list of one or more numeric or string expressions to print.

READ

READ reads those values and assigns them to variables.

READ variablelist

- **variablelist** One or more variables, separated by commas, that are assigned data values.

REM

Allows explanatory remarks to be inserted in a program.

REM remark
' remark

RIGHT$ (function)

Return a specified number of rightmost characters in a string.

**RIGHT$**(stringexpression$,n%)

• **stringexpression$**	Any string expression.
• **n%**	The number of characters to return, beginning with the rightmost string character.

STR$ (function)

STR$ returns a string representation of a number.

STR$(numeric-expression)

• **numeric-expression**	Any numeric expression.

THEN

(see IF)

VAL (function)

VAL converts a string representation of a number to a number.

VAL(stringexpression$)

• **stringexpression$**	A string representation of a number. (return 0 if the string isn't a represation of a number)