Computational biophysics

Giacomo Fantoni
telegram: @GiacomoFantoni
Github: https://github.com/giacThePhantom/mathematical-modelling-in-biology

November 10, 2022

Contents

Chapter 1

Introduction and proteins

1.1 Introduction

Biomolecular modelling has seen a recent increase in its use in the recent years, with a field still destined to expand. Most of this models take a top-down approach, starting from the macroscopic rather than to build simulation from the fundamental and quantistic concepts. Examples of systems studied through biomolecular modelling are:

- Channels.
- Photosynthetic systems.
- Viruses.
- DNA/RNA interactions.
- Inorganic systems.

Through biomolecular modelling it is possible to obtain:

- Molecular rationale for biological processes like proteins' function or its misfolding.
- A quantitative evaluation of molecular driving forces.
- A prediction of properties of macromolecular structures and architectures.
- A comparative assessment of molecular affinities through the binding free energy.

1.2 Proteins

Proteins have different functions within a cell:

- Give structure.
- Provide exchange of materials.
- Code for messages.
- Transport ions.
- Catalytic.
- Movement.
- Storage.
- Act as toxins.

Proteins are a polymer of amino-acids and occupy a space-scale of $10nm$. The amino-acids are in the range of $1nm$. They are built through a polymerization reaction as chain of amino-acids coded through a degenerate code of RNA nucleotides. Three bases of RNA code for an amino-acid.

1.2.1 Amino-acids

Amino-acids are the monomers of a protein. They have a general structure with an amino and a carboxyl terminal group for all of them. They are distinguished by a residue on the α-carbon which gives them different chemical and physical properties.

1.2.2 Structure

There are four level of a protein structure.

- Primary structure: the amino-acid sequence.
- Secondary structure: here α-helices and β-sheet can be distinguished.
- Tertiary structure: the spatial, 3D dynamic configuration of a protein which arise during protein folding.
- Quaternary structure: the interaction of multiple correctly-folded proteins.

Chapter 2

Proteins' geometry

2.1 Introduction

The study of the geometry of proteins involve what can be learned from protein coordinates.

2.2 The peptide bond

A protein is a collection of amino acids linked together by a peptide bond. A carboxylic end and an amino end of two amino acid react together losing a water molecule and forming a peptide bond. Beside the α-carbon there is another one bonded to the oxygen in the carboxylic group and a nitrogen bond in the amino group. The α carbon is linked to the nitrogen in the amino group of another amino acid. The carbon and nitrogen display $sp2$ hybridization, the central atom and the 3 that form a bond with it form a plane, so the peptide bond is planar. A plane of the peptide bond is formed and rotation of the plane is allowed only around one axis.

2.2.1 Trans and cis

Looking at the peptide bond the carbon atom of the carboxylic group C' and the nitrogen N are each bonded to a different α-C and a trans or cis conformation can happen. Trying to visualize the atoms that belong to the molecules these repel through the Van der Waals interactions, that can be computed through the Lennard-Jones potential:

$$U_{Lj}(r) = E_0\left[\left(\frac{r_0}{r}\right)^{12} - 2\left(\frac{r_0}{r}\right)^{6}\right]$$

Where:

- r_0 is the distance where the energy is minimum.
- r is the distance between two atoms.
- r_{min} is the distance at which the energy becomes high.

With atoms most of the time the distance between them will be close to r_0. When decreasing the distance a lot of energy is needed and strain is introduced in the molecule. Plotting the values

for the energy, r_0 and r_{min} the expected distance for each couple of atoms can be seen. Focusing on the C-C interaction:

- $r_0 = 3.4$Å.
- $r_{min} = 3.0$Å.

When two carbons atoms are below the minimum value the conformation is strained. Looking back at the conformation of the peptide bond it can be seen that the cis conformation creates a distance of 2.8Å between the two αC, so it is not favourable. So the trans conformation is the least energy-hungry and the most present.

2.3 The Ramachandran angles

The planes formed by the peptide bonds can rotate with respect to each other. So the Ramachandran angles ϕ and ψ can be defined between these planes. For each αC:

- ϕ describes the rotation around its bond with the nitrogen.
- ψ describes the rotation around its bond with the carboxylic group.

These are the angles between the subsequent planes. Some of the angles will require more energy.

2.3.1 Difficulty of rotation

It can be seen how a rotation of the ϕ angle could cause the two C' to come at a distance of 2.9Å (where $r_{min} = 3.0$Å. On the other hand a rotation of the ψ angle could cause the two N to come at the same distance, but in this case $r_{min}(N \dots N) = 2.7$Å. In the case of carbon atoms the distance is less than the minimum distance, while in the case of nitrogen it is greater than the minimum allowed value. Looking at this it can be seen how the ψ rotation is easier.

2.3.2 Ramachandran plot

A Ramachandran plot is a map with the ϕ angle on the x axis and the ψ angle on the y axis. Because a rotation along the ϕ angle is highly disfavoured the angle 0 is strongly disfavoured and is represented like a black stripe (disallowed region). If the amino acids where composed only by carbon and nitrogen atom the Ramachandran map would be **??**, where:

- A forbidden region in the middle.
- Some strained region like for $\psi = 0$.

Looking at a real protein the complexity is increased and the other oxygen and nitrogen atoms are included **??** and other regions become disallowed due to steady clashes. It can be seen how the regions are quite complex. Looking at a glycine and alanine complex it can be seen in **??** the space becomes even more complex. In this case the white regions is very small and a strained region can be seen and the black one. Including other residues the allowed region reduces **??**. This is due to the presence of larger residues.

2.3.2.1 Observed Ramachandran plot

Trying to plot for each amino acid its angles an amino acid is represented as a dot. Most of the points fall inside of the allowed regions but there are some outliers. In some conformation the protein forces the amino acid to assume strange conformations. This is done to check if the structure places the amino acids in a proper way.

2.4 Contact map of proteins

Starting from the coordinates a contact can be built. It is a matrix that map all the contact between the amino acids. A primary structure can be represented as a collection of beads which will be in contact in the 3D structure. A square matrix can be built such that each entry in the matrix will determine whether there is a contact or not. This matrix will be symmetric with diagonal elements with value 1 and two parallel diagonals for the neighbouring amino acids. Secondary structures will have specific signatures:

- α-helices: is usually represented by a line parallel to the diagonal. This is because the amino acids i is interacting with $i + 4$.
- β-strands: the situation is complicated. For parallel β sheets can be parallel to the diagonal. For anti-parallel it can be anti-parallel to the diagonal.

2.4.1 Defining a contact

The contact between two amino acids needs to be defined. To do so the distance between α-C or the distance between the tail of the residue and an α-C. There is also the need to make a trade-off between computational speed and cost. Also the dimension of the protein need to be considered when choosing the distance.

2.5 Topology diagram

Having found the secondary structures with a contact map a topology diagram help to understand how those interact with each other. In a topology diagram the start is the N terminus and the end the C terminus. β-strands are represented as arrows. If the strands always change direction they will form an anti-parallel β-sheet. α-helices are represented as small cylinder. Usually color codes represent the nature of the structure. This helps with numbering of the secondary structures.

2.6 Coordinates

The coordinates of all atoms in a protein are described in a PDB file. This is a tabulated file containing different columns:

- Atom record: ATOM.
- Atom number: a unique identifier for the atom.
- Atom identifier: an identifier for the type of atom.
- Amino type: the amino acid from which the atom is from.
- Chain identifier: identifier for the chain.
- Residue sequence number: the number of the residue in the chain.
- x, y, z: the coordinates in angstrom.
- Occupancy: the probability of an atom to be in that space.
- B-factor: how mobile that atom is in the crystal, it represent the noise in the x-ray diffraction map.
- Element symbol: the symbol of the element of the atom.

Once the coordinates of a protein is obtained, some geometrical properties can be directly computed.

2.6.1 Protein centre of mass

The protein centre of mass is the average position for the protein centre. It is an average weighted by the mass of the atom.

$$\vec{R}_{cm} = \frac{\sum_{i=1}^{N} m_i \vec{r}_i}{\sum_{i=1}^{N} m_i}$$

2.6.2 Radius of gyration

Once the centre of mass is known the radius of gyration can be computed. This measures the size of the protein as if it was a sphere. It is a good indication of the globular size of a protein. The distance of each atom and the centre of mass is computed and the square is taken, weighted with the mass of the atom.

$$R_g = \sqrt{\frac{\sum_{i=1}^{N} m_i (\vec{r}_i - \vec{R}_{cm})^2}{\sum_{i=1}^{N} m_i}}$$

2.6.3 Comparing protein structures

Proteins have structures that loop in a similar way, with similar regions within each other. To quantify the similarity between the protein structure a procedure needs to be followed:

- Select common regions: a 1-1 correspondence between amino acid need to be found: the parts present only in one protein are not considered. A correspondence is built between the common regions on the single amino-acids. These can be different, usually the coordinates are confronted between the α-C atom and the residue is not considered. One of the things that can be done is to look at the secondary structures and add loops only when they look similar.

- Align the two structures: compute the centre of mass of the two proteins and translate the proteins so the centre of masses coincide.

- Finding the optimal rotation: (add algorithm) the principal axes are computed and the proteins are rotated so that they superimpose. Once the optimal rotation is obtained the difference can be quantified.

- Compute RMSD: root mean square deviation: take the coordinates of the amino-acid i in protein A and A, their squared difference is computed and an average over all amino acid is computed and squared:

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\vec{r}_{Ai} - \vec{r}_{Bi})^2}$$

  This can be done for two proteins or for the protein taken at two different time step in a molecular dynamics simulation:

$$RMSD(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\vec{r}_i(t) - \vec{r}_i(0))^2}$$

Now the $RMSD$ can be plotted with respect to time. It can be seen how at $t = 0$ $RMSD = 0$ and after the value will increase. When the number reaches a plateau the protein should be in equilibrium. The plateau can jump to another value, meaning that the state is a meta-stable state of the protein, or the protein has more stable states or a loop is making something. This value is assigned to very complicated structures and different structures can have the same $RMSD$. The $RMSD$ is an indicator of equilibrium: it is a necessary but not sufficient condition.

2.6.4 Native state

The native state is the functional state of a protein. It is not the state found for a crystallized protein, but only closely related to it. This is due to proximity effect and the fact that the protein is not a static system. Proteins are extremely flexible and are moving a lot because the temperature corresponds to a constant movement of water molecule around causes movement in the protein. The native state is an ensemble of closely related states, all compatible with the conditions of the situation studied. Proteins need to be studied in the isothermal-isobaric ensemble. All the calculation need to be done at constant temperature and pressure. The native state is so a collection of functional state.

In the case of the unfolded state the possibilities are too many to sample all of them.

2.6.5 RMSF

The flexibility of each amino acid can be computed. With flexibility is intended the movement of amino acid with respect to one another. In the α-helix, for example, less fluctuation is expected, while in loops more fluctuation is expected. This quantity is computed in the root mean squared fluctuation, which will be computed for each amino acid in the protein. With f referring to the frame, let:

- $\langle\vec{r}_i\rangle = \frac{1}{M}\sum_{f=1}^{M}\vec{r}_{i,f}$, the average position of atom i.
- $\Delta\vec{r}_{i,f} = \vec{r}_{i,f} - \langle\vec{r}_i\rangle$, the displacement of each atom in each frame with respect to its average.
- $\langle\Delta\vec{r}_o^{\,w}\rangle = \frac{1}{M}\sum_{f=1}^{M}(\vec{r}_{i,f} - \langle\vec{r}_i\rangle)^2$, the average squared distance over the frames.

So, the root mean squared fluctuation is:

$$RMSF_i = \sqrt{\langle\Delta\vec{r}_i\rangle^2}$$

Plotting the $RMSF$ with respect to residue number and the more mobile residue can be identified. This can be mapped onto the sequence so loops, helices and strands can be recognized. Usually the fluctuating part correspond to loops. The terminus have the highest $RMSF$.

2.6.5.1 B-factors

$RMSF$ can be translated into B-factors. They are the Debye-Waller factors and are a scaled version of the $RMSF$ squared. So the result of a simulation can be compared with the B-factor and a strong correspondence can be seen. The differences are due to the fact that the crystal is a different environment with respect to the normal one and packing effect can happen (some regions of the protein can interact with the image of the protein in the crystal).

$$B_i = \frac{8\pi^2}{3}\langle \Delta \vec{r}_i^2 \rangle = \frac{8\pi^2}{3} RMSF_i^2$$

Chapter 3

Semi-empirical force fields

3.1 Introduction

A protein system can be modelled using a force field. Whenever a simulation is started the interactions between atoms need to be determined. To do so a topology file is built. This file will contain the connecting information. This is important because the kind of force field is specified here. Semi-empirical force field will be used because a very complicated process is modelled using formula that can be easily computed. These formulae are not rigorous and what it is done is to compare the result of the simulation with an experiment. Because of these the used force fields will be called semi-empirical.

3.2 Potential energy surface

Any system will correspond to a given potential energy surface. There are states which correspond to minima of this energy where the system will spend most of its time. In the global minimum the system is expected to be in equilibrium. There are also local minima where the system can spend some time and the system will go from the global to some other local minima and the time spent there depends on how deep that minimum is. There will be transitions from one minimum to another exploring the potential energy surface passing through a set of points. In principle the potential energy surface is known because it is assumed that there is a potential energy for all the interaction that there are in the system. This is not that simple because of temperature: at 0 temperature the potential energy surface will tell everything: the system goes through the global minimum and stay there. At finite and high temperatures the system will jump with a certain frequency, which depends on temperature, from one state to another. In this case statistical mechanics is necessary not the potential energy surface is necessary but also the free energy surface and the entropy have to be taken into account. The entropy part is the most difficult. In order to compute entropy the conformational space need to be explored: so the compatible conformation in certain condition need to be found. The phase space, the possibilities of a molecule need to be explored. These are so many in the case of a protein that estimating entropy will be the more costly process. In order to estimate entropy long simulation are needed as the conformational space will be explored as much as possible. In this chapter entropy will not be considered.

The potential energy surface or PES is the landscape of what values the potential energy of an atom can assume. Different points can be recognized like:

- Saddle point.
- Local maximums.
- Local minimum.

3.2.1 Bond stretching

In figure **??** the typical potential energy for a chemical bond can be seen. Let r_{eq} the distance for which the energy is minimal, where the equilibrium is. Moving from it the energy increases. There are transitions between different states and in that case quantum physics should be used. Everything will be assumed to be described using classical mechanics. This is valid when dealing with conformational transitions. SO all variables are continuous. Using this assumption light absorption and a chemical reactions cannot be described. The potential energy surface to describe an unbreakable bond can be seen in figure **??**. When close to the minimum the well can be assume symmetric, but some asymmetry later: the left part is steeper. An approximation for these potential energy is built into the force field. The potential energy surface can be reconstructed using vibrational spectroscopy and then it is reconstructed using a mathematical model. Using a Taylor expansion the potential energy is approximated at point r by taking the value at the equilibrium and then constructing all the corrections. The first one is the first derivative at the value at equilibrium multiplied by the distance. Then for the second correction the second derivative and so on. If the potential is symmetric the first derivative is 0, as it happens near the equilibrium. So the first correction can be neglected. Also the third order term is either 0 because the third derivative is 0 if the minimum is really shallow. This term is really small if r is really close to the equilibrium so it can neglected.

$$U(r) = U(r_{eq}) + \frac{dU}{dr}|_{r=r_{eq}}(r - r_{eq}) + \frac{1}{2!}\frac{d^2U}{dr^2}|_{r=r_{eq}}(r - r_{eq}^2) + \frac{1}{3!}\frac{d^3U}{dr^3}|_{r=r_{eq}}(r - r_{eq}^3) + \cdots$$

$$U(r) = U(r_{eq}) + \cancel{\frac{dU}{dr}|_{r=r_{eq}}(r - r_{eq})} + \frac{1}{2!}\frac{d^2U}{dr^2}|_{r=r_{eq}}(r - r_{eq}^2) + \cancel{\frac{1}{3!}\frac{d^3U}{dr^3}|_{r=r_{eq}}(r - r_{eq}^3)} + \cdots$$

So that in the end the typical harmonic potential is obtained.

$$U(r_{AB}) = \frac{1}{2}k_{AB}(r_{AB} - r_{AB,eq})^2$$

The constant k is related to the second derivative with respect to the distance. The distances and the k constants are labelled with A and B, which stand for the fact that this interaction has to be described for each couple of atom. Transferability of these parameter is an issue: using a particular force field then the parameters cannot be used for another. So each force field will come with its own set of parameters.

3.2.1.1 Anharmonic force constant

Considering that the shape is not completely symmetric the third order term can be inserted because it can be important. This introduces an asymmetry in the system: an anharmonic force constant.

$$U(r_{AB}) = \frac{1}{2}[k_{AB} + k_{AB}^{(3)}(r_{AB} - r_{AB,eq})](r_{AB} - r_{AB,eq})^2$$

3.2.1.2 Quartic correction

Also the fourth order term can be inserted.

$$U(r_{AB}) = \frac{1}{2}[k_{AB} + k_{AB}^{(3)}(r_{AB} - r_{AB,eq}) + k_{AB}^{(4)}(r_{AB} - r_{AB,eq})^2](r_{AB} - r_{AB,eq})^2$$

3.2.1.3 Morse potential

The Morse potential is used in implicit solvent simulation and uses the exponential because it can describe screen interaction that happens with implicit solvent. The exponential is quite expensive for a computer to compute. It is useful also for soft-system or coarse grained system.

$$U(r_{AB}) = D_{AB}[1 - e^{-\alpha_{AB}(r_AB - r_{AB,eq})^2}]$$

3.3 Valence angle bending

Chemical bonds are not described only by strings and beads but also bending of the bonds can happen. These are valence angle bending. These bonds cannot be broken during a classical molecular dynamic simulation. In order to describe them a Taylor expansion introducing and equilibrium angle is built. The first order term is not considered as it will be equal to 0. Then the second introduces the harmonic and the third for the anharmonic one, in principle also the quartic correction could be added. This formula is similar to the previous one but all the constant have to be described between each triplet of atoms.

$$U(\theta_{ABC}) = \frac{1}{2}[k_{ABC} + k_{ABC}^{(3)}(\theta_{ABC} - \theta_{ABC,eq}) + k_{ABC}^{(4)}(\theta_{ABC} - \theta_{ABC,eq})^2 + \cdots](\theta_{ABC} - \theta_{ABC,eq})^2$$

3.3.1 Multiple minima

There is another problem with valence angle bending: angles are not varying continuously: the Taylor expansion is good if the angles are not varying to much. If the angles invert the Taylor expansion cannot take track of it. In order to take this into account the potential energy is built using a Fourier expansion. A Fourier expansion introduces a periodic function like cos that contains oscillations in a period, allowing to model any possible periodic potential. Since the angles are being considered the potential is periodic. So in the parametrization of the angle bending interaction the Fourier term is introduced, labelling terms with the value j. So the amplitude, the value that multiplies each Fourier component is decreasing with j and a cut-off of the given value of j is given. In this case only low-frequency component are considered.

$$U(\theta_{ABC}) = \sum_{\{j\}_{ABC}} k_{j,ABC}^{fourier}[1 + \cos(j\theta_{ABC} + \psi_j)]$$

Where the amplitude:

$$k_{j,ABC}^{fourier} = \frac{2k_{ABC}^{harmonic}}{j^2}$$

Dividing by j^2 makes the contribution smaller for higher j.

3.4 Torsions

In order to describe torsion 4 atoms bonded by subsequent bonds are considered $A->B->C->D$. In this case two planes ABC and BCD are built and the angle ω between two plane is built. If they are on the same plane the angle is 0. So an angle ω is computed to describe the torsion that is around the bond $B->C$. In order to compute the angle the perpendicular vectors between the planes is computed and the angle between these two vector is ω. In this case the harmonic oscillation and the Taylor expansion are not used but the Fourier expansion and the periodic potential is used. This is because the value of the energy will be the same after a rotation. In principle torsions will be explored during a simulation. All the possible groups with permutations and repetitions of 4 atoms need to be considered to parametrize torsions.

$$U(\omega_{ABCD}) = \frac{1}{2} \sum_{\{j\}_{ABCD}} V_{j,ABCD}[1 + (-1)^{j+1} \cos(j\omega_{ABCD} + \psi_{j,ABCD})]$$

3.4.1 Improper torsions

Improper torsions happen when the bonds are $A->B->C$ and $B->D$. B can be in the same plane of A, C, D or it can pop out. In this case two planes are built, usually ABC and BCD and the angle between them will be built. The potential will describe the energy of the angle. The angle can be described as the out of plane angle OOP, obtaining the equation for the plane ACD and how much B is out of that plane. Here all possibilities of an atoms connected with other 3 atoms need to be considered. In the case of sp^2 the four atoms need to be in a plane.

$$U(\omega_{ABCD}) = \frac{1}{2} \sum_{\{j\}_{ABCD}} V_{j,ABCD}[1 + (-1)^{j+1} \cos(j\omega_{ABCD} + \psi_{j,ABCD})]$$

3.5 Van der Waals interactions

Van der Waals interactions happen whenever two atoms come close to one another withouth any chemical bond connecting them. This interaction is called dispersion interaction and depends on the interaction between the electron clouds that become correlated. Even if electrons are not shared the fluctuation of the electron clouds of an atom will interact with the one of an atom close to it. This can be computed and described by an attraction that goes with the 6th power of the distance. This interaction is the Van der Waals interaction. This attraction is an attractive force and will decrease with the 6th power of the distance. Also a repulsion happen. The repulsion is due to the occupancy of the possible orbitals. The problem is then Pauli's exclusion principle: whenever two atoms come close the energy level becomes occupied and the atoms cannot come to close to one another. This is difficult to compute but it can be modelled assuming that is a repulsion with an hard limit. The energy becomes very high when the distance becomes less than the equilibrium. If the attraction is modelled with the 6th power this interaction can be modelled by the 12th power. In this case if the distance is too small the energy becomes very high.

3.5.1 Lennard-Jones potential

The discussion done above lead to the introduction of the Lennard-Jones potential. This models Van der Waals interaction using the 6th power for the dispersion and the repulsion due to Pauli's

exclusion principle is modelled by the 12th power. This power is used to make the computation cheap and has no physical basis. In this way this formula is obtained:

$$U_(r_{AB}) = \frac{a_{AB}}{r_{AB}^{12}} - \frac{b_{AB}}{r_{AB}^{6}} = \\ = 4\epsilon_{AB}\left[\left(\frac{\sigma_{AB}}{r_{AB}}\right)^{12} - \left(\frac{\sigma_{AB}}{r_{AB}}\right)^{6}\right]$$

And the distance with minimum energy or equilibrium.

$$r_{AB}^{*} = 2^{\frac{1}{6}}\sigma_{AB}$$

And it can be seen how it is a multiple of σ_{AB} or the Van der Waals radius, the distance where the value of the energy if exactly equal to 0. $-\epsilon$ is the minimum value of the energy (equilibrium). In this way dispersion and repulsion of any couple of atoms can be described. A value of σ_{AB} have to be introduced for any couple of atom type.

3.5.2 Morse potential

The Morse potential and the Hill potential can be used with coarse grained system. Dealing only with all-atom simulation these will be rarely used. Working in soft-matter simulation these could be used. These have the same features of the Lennard-Jones potential.

$$U(r_{AB}) = D_{AB}[1 - e^{-a_{AB}(r_{AB}-r_{AB,eq})^2}]$$

3.5.3 Hill potential

$$U(r_{AB}) = \epsilon_{AB}\left[\frac{6}{\beta_{AB}-6}e^{\beta_{AB}\frac{1-r_{AB}}{r_{AB}^{*}}} - \frac{\beta_{AB}}{\beta_{AB}-6}\left(\frac{r_{AB}^{*}}{r_{AB}}\right)^{6}\right]$$

In some force fields 1-4 interactions (successive bonds atoms can be numbered) are reduced by a scaled factor. Atom 1 and 4 are cose to each other in a proper torsion but attraction and repulsion has been already described by the torsion, so in some force fields, the torsions takes into account the proximity of those atoms. So the Van der Waals interactions are excluded for 1-4 atoms as they are already been described by the torsion interaction.

3.6 Electrostatic interactions

Electrostatic interactions, together with Van der Waals one are two kind of non-bonded interactions. Moreover interactions between molecules that might be charged or that can be polar. The distribution of charges need to be described. In principle a multiple expansion should be performer: the cloud of positive and negative charges should be described by using given shapes of the clouds. And each molecule should be described and all the molecules should be described by these multiple expansion and they should be multiplied as matrix.

$$U_{AB} = \vec{M}^{(A)}V^{(B)}$$

Now, summing over all molecules:

$$U_{AB} = \sum_{A} \sum_{B>A} \vec{M}^{(A)} \vec{V}^{(B)}$$

This is a very costly procedure but accurate.

3.6.1 Point like charges

To make the computation easier molecules are represented as point-like partial charges. In a chemical molecule partial charges can be assumed that are point-like and are located on the atom. The electron cloud is displaced toward the negative cloud and away from the positive one. In this way the distribution of charges can be represented by number placed on each atom. Once this is done an electric charge is associated with each bead and the Coulomb interaction is used:

$$U_{AB} = \frac{q_A q_B}{\epsilon_{AB} r_{AB}}$$

The dielectric constant is 1 in the case of explicit solvent, but in the case of the implicit one it will be the one of the solvent.

3.6.2 Dipolar interactions

Bipolar interaction can be included when considering electrostatic forces. These formulae are mostly used in corse grained models. Because in that case the approximation of single charges on a bead cannot be made and dipoles are considered. In this case a functional form and the parameters need to be included. μ is the dipole of two molecules. Then there is the dielectric constant, the cube of the distance and the cosine term represent the orientation of the two dipoles.

$$U_{AB/CD} = \frac{\mu_{AB}\mu_{CD}}{\epsilon_{AB/CD} r^3_{AB/CD}} (\cos \chi_{AB/CD} - 3 \cos \alpha_{AB} \cos \alpha_{CD})$$

3.6.3 Dielectric constants

$$U_{AB} = \frac{q_A q_B}{\epsilon_{AB} r_{AB}}$$

The dielectric constant can assume different values. If they are connected by a chemical bond or if there is a valence angle between the atoms the dielectric constant if ∞ and the interaction is not computed as it is already taken into account by other terms. In several force field a factor of 2 is added for 1-4 interaction. In the last case the value depend on the force field.

$$\epsilon_{AB} = \begin{cases} \infty & \text{if } A \wedge B \text{are 1,2- or 1,3-related} \\ 3.0 & \text{if } A \wedge B \text{are 1,4-related} \\ 1.5 & \text{otherwise} \end{cases}$$

3.7 Cross terms

There are also cross terms: angle bending and bond stretching and other movement are not independent: all the possible degrees of freedom should be considered together. This is done only by some force fields.

$$
\begin{aligned}
U(\vec{q}) =& U(\vec{q}_{eq}) + \sum_{i=1}^{3N-6} (q_i - q_{i,eq}) \frac{\partial U}{\partial q_i}|_{\vec{q}=\vec{q}_{eq}} + \\
&+ \frac{1}{2!} \sum_{i=1}^{3N-6} \sum_{j=1}^{3N-6} (q_i - q_{i.eq})(q_j - q_{j,eq}) \frac{\partial^2 U}{\partial q_i \partial q_j}|_{\vec{q}=\vec{q}_{eq}} + \\
&= \frac{1}{3!} \sum_{i=1}^{3N-6} \sum_{j=1}^{3N-6} \sum_{k=1}^{3N-6} (q_i - q_{i,eq})(q_j - q_{j.eq})(q_k - q_{k,eq}) \frac{\partial^3 U}{\partial q_i \partial q_j \partial q_k}|_{\vec{q}=\vec{q}_{eq}} + \cdots
\end{aligned}
$$

This is done by starting from a Lagrangian describing all the interaction and compute all the terms. In practice everything is done without considering them.

$$U(r_{AB}, \theta_{ABC}) = \frac{1}{2} k_{AB,ACB}(r_{AB} - r_{AB,eq})(\theta_{ABC} - \theta_{ABC,eq})$$

3.8 Parametrization

All the interactions need a large set of parameters. The number of parameters increases with the number of atoms N:

$$p = N + (N-1) + (N-2) + \cdots = N\frac{N+1}{2}$$

In order to obtain these parameters data from experiments are used. This is the reason why the force-field are semi-empirical. To do so the results of the experiments are compared from the results of simulations. Then the penalty function:

$$Z = \left[\sum_{i}^{observables} \sum_{j}^{occurrences} \frac{(calc_{i,j} - expt_{i,j})^2}{w_i^2} \right]^{\frac{1}{2}}$$

Sums over all the data from all the observables and all possibilities. Then all the observable are measured and are compared with the simulation result. Then the square is taken to not compensate the error and it is divided by a weight. This penalty function is minimized changing the parameters until some reasonable compromise is obtained. The more experimental data is obtained the more accurate the description. This is the reason for the constant update of the force fields. A way to reduce the number of parameters is to employ a strategy so that a parameter to each atom is computed and then it is computed the interaction between two atoms:

$$\sigma_{AB} = \sigma_A + \sigma_B \epsilon_{AB} \qquad\qquad = (\epsilon_A \epsilon_B)^{\frac{1}{2}}$$

3.9 Force field energies

In molecules the ground state of one molecule is different from the ground state of another. Molecules can have different states and when they come in contact the ground state have to be taken into account. So parametrizing the force field is very complicated.

3.9.1 Geometry optimization

Once the force fields are obtained some atoms can be missing from the model and have to be included by hand. Sometimes it may happen that two atoms are too close to each other and by using the potential energy it can be high leading to high forces, causing the atom that experience them to be kicked out. So the simulation will explode. In order to avoid this a geometry optimization is done to obtain the parameters or to minimize so to avoid high forces values at the start of a simulation. When starting a simulation the system will be at an energy value which be greater than the equilibrium. The gradient of the energy is computed and the system is moved toward the minimum energy. The derivative need to be computed of U with respect to each component in the system.

$$\vec{g}(\vec{q}) = \begin{bmatrix} \frac{\partial U}{\partial q_1} \\ \frac{\partial U}{\partial q_2} \\ \vdots \\ \frac{\partial U}{\partial q_n} \end{bmatrix}$$

Such that the cost reaches a global minimum $J_{min}(\vec{w})$.

3.9.2 Derivative of the potential function

First the derivatives need to be computed. All the potentials are given in term of the mutual distance of two atoms. So, considering x the coordinates of a point:

$$\frac{\partial U}{\partial x_A} = \sum_{i \in A} \frac{\partial U}{\partial r_{Ai}} \frac{\partial r_{Ai}}{\partial x_A}$$

Taking the derivative of this:

$$U(r_{AB}) = \frac{1}{2}[k_{AB} + k_{AB}^{(3)}(r_{AB} - r_{AB,eq}) + k_{AB}^{(4)}(r_{AB} - r_{AB,eq})^2](r_{AB} - r_{AB,eq})^2$$

With the respect on the distance:

$$\frac{\partial U}{\partial r_{Ai}} = \frac{1}{2}[2k_{Ai} + 3k_{Ai}^{(3)} r_{Ai} - r_{Ai,eq}) + 4k_{Ai}^{(4)}(r_{Ai} - r_{Ai,eq})^2](r_{Ai} - r_{Ai,eq})$$

In order to take the derivative with respect to r with respect to x need to be known:

$$\frac{\partial r_{Ai}}{\partial x_A} = \frac{x_A - x_i}{\sqrt{(x_A - x_i)^2 + (y_A - y_i)^2 + (z_A - z_i)^2}}$$

Then this formula can plugged in according to the chain rule.

3.9.2.1 Newton-Raphson

To perform the minimization an iterative procedure, the Newton-Raphson to find a local minimum is employed. Here (n) refers to the iteration. This allow to obtain iteration $k+1$ starting from iteration k and is a Taylor expansion of coordinates at iteration k:

$$U(\vec{q}^{(k+1)}) = U(\vec{q}^{(k)}) + (\vec{q}^{(k+1)} - \vec{q}^{(k)})\vec{g}^{(k)} + \frac{1}{2}(\vec{q}^{(k+1)} - \vec{q}^{(k)})H^{(k)}(\vec{q}^{(k+1)} - \vec{q}^{(k)})$$

This is done to find eventually the coordinates at iteration $k+1$ which are unknown. Where H is the Hessian matrix built:

$$H_{ij}^{(k)} = \frac{\partial^2 U}{\partial q_i \partial q_j}|_{\vec{q}=\vec{q}^{(k)}}$$

Everything is computed from a Taylor expansion from coordinates at iteration k. And:

$$\frac{\partial U(\vec{q}^{(k+1)})}{\partial q_i^{(k+1)}} = \frac{\partial \vec{q}^{(k+1)}}{\partial q_i^{(k+1)}}\vec{g}^{(k)} + \frac{1}{2}\frac{\partial \vec{q}^{(k+1)}}{\partial q_i^{(k+1)}}H^{(k)}(\vec{q}^{(k+1)} - \vec{q}^{(k)}) + \frac{1}{2}(\vec{q}^{(k+1)} - \vec{q}^{(k)})H^{(k)}\frac{\partial \vec{q}^{(k+1)}}{\partial q_i^{(k+1)}}$$

This formula tells that:

$$\vec{g}_i^{(k+1)} = \vec{g}_i^{(k)} + [H^{(k)}(\vec{q}^{(k+1)} - \vec{q}^{(k)})]_i$$

Where at the end it should be obtained a stationary condition, where the system does not change after an iteration:

$$\vec{0} = \vec{g}^{(k)} + H^{(k)}(\vec{q}^{(k+1)} - \vec{q}^{(k)}) \Rightarrow \vec{q}^{(k+1)} = \vec{q}^{(k)} - [H^{(k)}]^{-1}\vec{g}^{(k)}$$

Because the starting point was a Taylor expansion the new coordinates will not be exact, and so other iterations will be needed. This process is repeated until convergence is reached, choosing a given tolerance on $\vec{0}$.

3.9.3 Types of force fields

Force fields can be categorized as:

- All atoms: one atom corresponds to one bead.
- More atoms: more atoms correspond to one bead.
- Corse grained: groups of atoms correspond to one bead.
- Polarizable force fields: point charges are variables.

As a golden rule parameters from different force fields should never be mixed.

Chapter 4

Classical mechanics

Chapter 5

Foundations of statistical mechanics

Chapter 6

Microcanonical ensemble

Chapter 7

Molecular dynamics

Chapter 8

Direct translation

Chapter 9

Evaluation of energies and forces

Chapter 10

Canonical ensemble

Chapter 11

Thermostats