SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING
DEPARTMENT OF BIOINFORMATICS

UNIT – 1- SBIA5203 – Biomolecular modeling

MOLECULAR MODELING

The term molecular modelling expanded over the last decade from the tools to visulalize three dimensional structures and to simulate, predict and analyse the properties and the behaviour of the molecules on an atomic level to data mining and platform to organize many compounds and their properties into database and to perform virtual drug screening via 3D database screening for novel drug compounds.

Molecular modelling allow the scients to use computers to visualize molecules means representing molecular structures numerically and simulating their behaviour with the equations of quantum and classical physics to discover new lead compounds for drugs or to refine existing drugs insilico.

Goal:

To develop a sufficient accurate model of the system so that physical experiment may not be necessary

The definition currently accepted of what molecular modeling is, can be stated as this: "Molecular modeling is anything that requires the use of a computer to paint,
describe or evaluate any aspect of the properties of the structure of a molecule" (Pensak, 1989). Methods used in the molecular modeling arena regard automatic structure generation, analysis of three-dimensional (3D) databases, construction of protein models by techniques based on sequence homology, diversity analysis, docking of ligands or continuum methods.

Thus, today molecular modelling is regarded as a field concerned with the use of all sort of different strategies to model and to deduce information of a system at the atomic level. On the other hand, this discipline includes all methodologies used in computational chemistry, like computation of the energy of a molecular system, energy minimization, Monte Carlo methods or molecular dynamics. In other words, it is possible to conclude that computational chemistry is the nucleus of molecular modeling.

Applications

Molecular modelling methods are now routinely used to investigate the structure, dynamics, surface properties and thermodynamics of inorganic, biological and polymeric systems.

The types of biological activity that have been investigated using molecular modelling include protein folding, enzyme catalysis, protein stability, conformational changes associated with biomolecular function, and molecular recognition of proteins, DNA, and membrane complexes.

Why models are used?

- a) to help with analysis and interpretation of experimental data
- b) to uncover new laws and formulate new theories
- c) to help solve problems and hint solutions before doing experiments
- d) to help design new experiments
- e) to predict properties and quantities that are difficult or even impossible to observe experimentally

Simulations and computer "experiments" can be designed to mimic reality, however, are always based on assumptions, approximations and simplifications (i.e. models).

Important characteristics of models are:

- Level of simplification: very simple to very complex
- Generality: general or specific, i.e. relate only to specific systems or problems
- Limitations: one must always be aware of the range of applicability and limits of accuracy of any model.
- Cost and efficiency: CPU time, memory, disk space

Computable quantities:

a) molecular structures: closely tied to energy (best structure - one for which the energy is minimum)

b) energy: potential energy surfaces (PES) - extremely important! PES dictate essentially everything about the molecule or system

c) molecular properties that can be compared to/used to interpret experiments: thermodynamics, kinetics, spectra (IR, UV, NMR)

d) properties that are not experimental observables: bond order, aromaticity, molecular orbitals

Three stages of Molecular Modeling

1. Model is selected to describe the intra and inter mol. Interactions in the system Two common models Quantum mechanics Molecular mechanics
   These models enable the energy of any arrangement if the atoms and mol to be calculated and allow the modeller to determine how the energy of the system varies

as the positions of the atoms and molecular changes

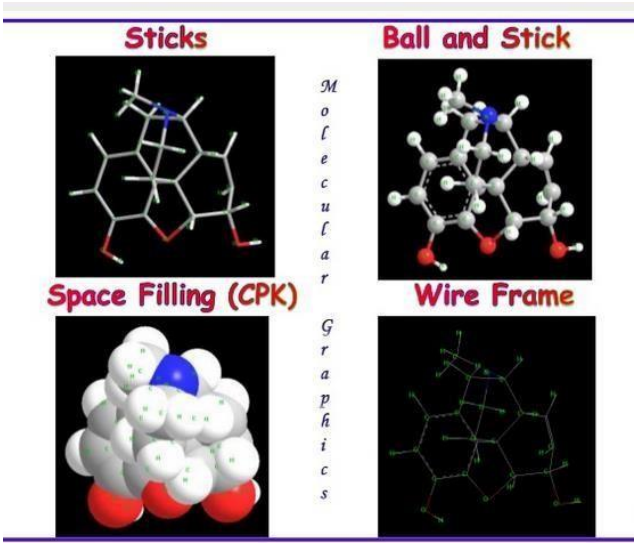
2. Calculation itself such as energy minimization, molecular dynamics or Monte carlo simulations or conformational search

3. Calculation must be analyzed not only to calculate properties but also to check that it has been performed properly

Molecular Visualisation

Once 3D coordinates are available, they can be visualised, an important aid to interpretation of molecular modelling:

- **Wireframe**, **Ball and Stick** and **Spacefill** for small and medium sized molecules
- **Ribbon** for protein, nucleotide and carbohydrate structures to render the tertiary molecular structures, **Polyhedral modes** for eg ionic lattices.
- **Isosurfaces**, which are generated from the sizes of atoms, and onto which can be colour coded further properties such as MOs, charges etc.
- **Animation** to view molecular vibrations and the time dependent properties of molecules such as (intrinsic) reaction coordinates, protein folding dynamics, etc.
- **Integration and Scripting**. Programs such as Jmol or ChemDoodle allow seamless integration of models as part of lecture courses, electronic journals, podcasts, iPads, etc and increasingly elaborate scripting of the models to illustrate scientific points.



1.2 Coordinate Systems

It is obviously important to be able to specify the positions of the atoms and/or molecules in the system to a modelling program*. There are two common ways in which this can be done. The most straightforward approach is to specify the Cartesian (x, y, z) coordinates of all the atoms present. The alternative is to use *internal coordinates*, in which the position of each atom is described relative to other atoms in the system. Internal coordinates are usually written as a Z-matrix. The Z-matrix contains one line for each atom in the system. A sample Z-matrix for the staggered conformation of ethane (see Figure 1.1) is

*For a system containing a large number of independent molecules it is common to use the term 'configuration' to refer to each arrangement; this use of the word 'configuration' is not to be confused with its standard chemical meaning as a different bonding arrangement of the atoms in a molecule

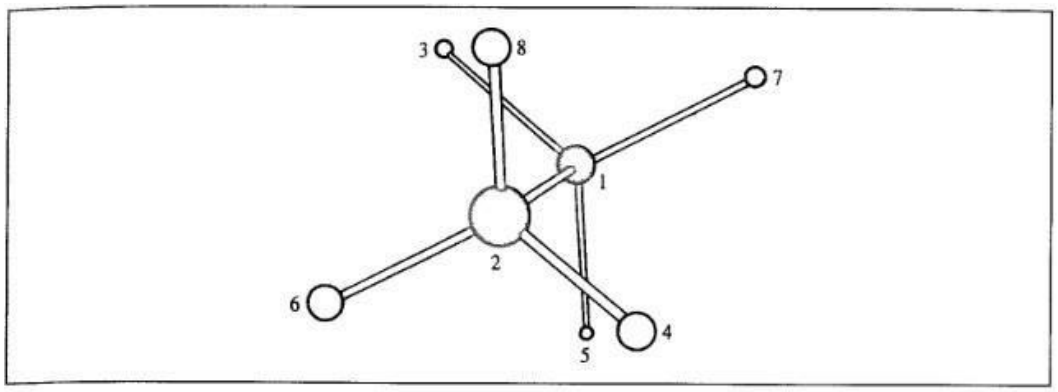


Fig. 1.1 The staggered conformation of ethane

as follows:

1	C						
2	C	1.54	1				
3	H	1.0	1	109.5	2		
4	H	1.0	2	109.5	1	180.0	3
5	H	1.0	1	109.5	2	60.0	4
6	H	1.0	2	109.5	1	−60.0	5
7	H	1.0	1	109.5	2	180.0	6
8	H	1.0	2	109.5	1	60.0	7

In the first line of the Z-matrix we define atom 1, which is a carbon atom. Atom number 2 is also a carbon atom that is a distance of 1.54 Å from atom 1 (columns 3 and 4). Atom 3 is a hydrogen atom that is bonded to atom 1 with a bond length of 1.0 Å. The angle formed by atoms 2–1–3 is 109.5°, information that is specified in columns 5 and 6. The fourth atom is a hydrogen, a distance of 1.0 Å from atom 2, the angle 4–2–1 is 109.5°, and the torsion angle (defined in Figure 1.2) for atoms 4–2–1–3 is 180°. Thus for all except the first three atoms, each atom has three internal coordinates: the distance of the atom from one of the atoms previously defined, the angle formed by the atom and two of the previous atoms, and the torsion angle defined by the atom and three of the previous atoms. Fewer internal coordinates are required for the first three atoms because the first atom can be placed anywhere in space (and so it has no internal coordinates); for the second atom it is only necessary to specify its distance from the first atom and then for the third atom only a distance and an angle are required.

It is always possible to convert internal to Cartesian coordinates and vice versa. However, one coordinate system is usually preferred for a given application. Internal coordinates can usefully describe the relationship between the atoms in a single molecule, but Cartesian coordinates may be more appropriate when describing a collection of discrete molecules. Internal coordinates are commonly used as input to quantum mechanics programs, whereas calculations using molecular mechanics are usually done in Cartesian coordinates. The total number of coordinates that must be specified in the internal coordinate system is six fewer

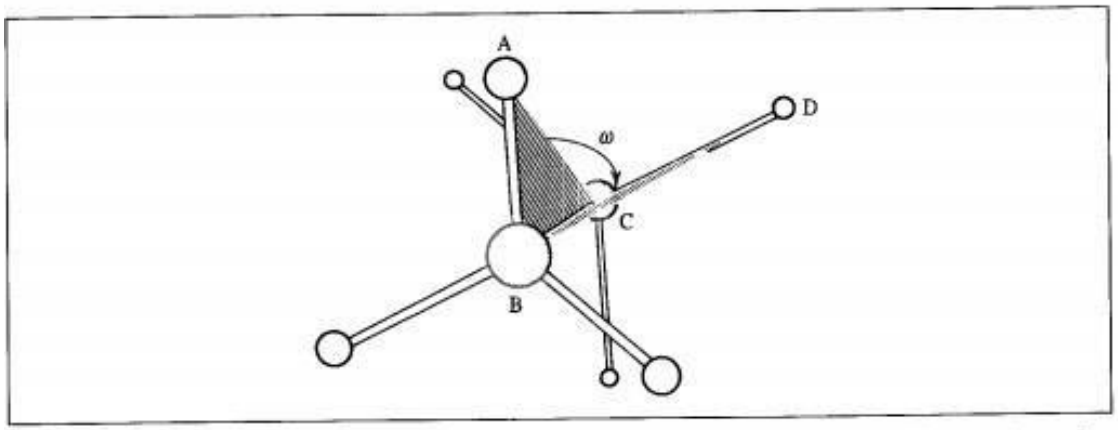


Fig. 1.2 A torsion angle A–B–C–D is defined as the angle between the planes A, B, C and B, C, D A torsion angle can vary through 360° although the range −180° to +180° is most commonly used We shall adopt the IUPAC definition of a torsion angle in which an eclipsed conformation corresponds to a torsion angle of 0° and a trans *or* anti *conformation to a torsion angle of 180°. The reader should note that this may not correspond to some of the definitions used in the literature, where the* trans *arrangement is defined as a torsion angle of 0° If one looks along the bond B–C, then the torsion angle is the angle through which it is necessary to rotate the bond AB in a clockwise sense in order to superimpose the two planes, as shown*

than the number of Cartesian coordinates for a non-linear molecule. This is because we are at liberty to arbitrarily translate and rotate the system within Cartesian space without changing the relative positions of the atoms

POTENTIAL ENERGY SURFACE

A potential energy surface (PES) describes the energy of a system, especially a collection of atoms, in terms of certain parameters, normally the positions of the atoms. The surface might define the energy as a function of one or more coordinates; if there is only one coordinate, the surface is called a *potential energy curve*.

The PES concept finds application in fields such as chemistry and physics, especially in the

theoretical sub-branches of these subjects. It can be used to theoretically explore properties of structures composed of atoms, for example, finding the minimum energy shape of a molecule or computing the rates of a chemical reaction

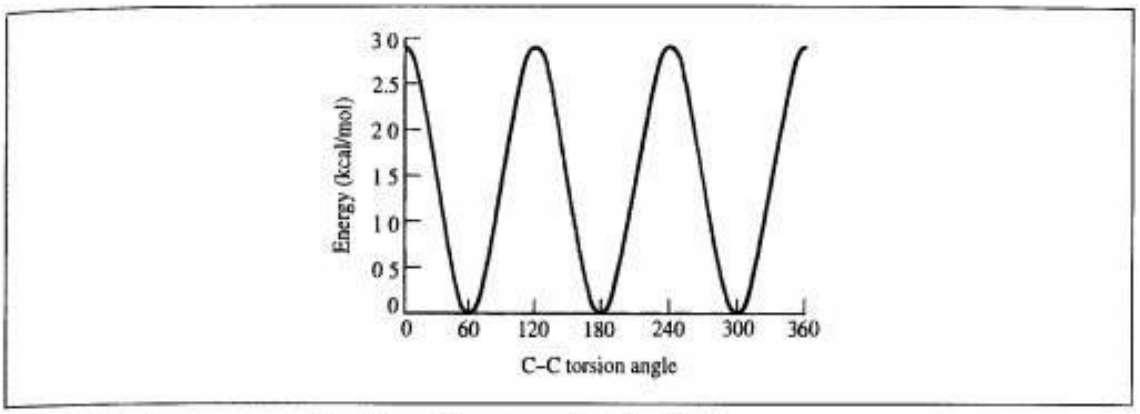


Fig. 1 3 Variation in energy with rotation of the carbon–carbon bond in ethane

Changes in the energy of a system can be considered as movements on a multidimensional 'surface' called the *energy surface*. We shall be particularly interested in stationary points on the energy surface, where the first derivative of the energy is zero with respect to the internal or Cartesian coordinates. At a stationary point the forces on all the atoms are zero. Minimum points are one type of stationary point; these correspond to stable structures. Methods for locating stationary points will be discussed in more detail in Chapter 5, together with a more detailed consideration of the concept of the energy surface.

1.4 Molecular Graphics

Computer graphics has had a dramatic impact upon molecular modelling. It should always be remembered, however, that there is much more to molecular modelling than computer graphics. It is the interaction between molecular graphics and the underlying theoretical methods that has enhanced the accessibility of molecular modelling methods and assisted the analysis and interpretation of such calculations.

Molecular graphics systems have evolved from delicate and temperamental pieces of equipment that cost hundreds of thousands of pounds and occupied entire rooms, to today's inexpensive workstations that fit on or under a desk and yet are hundreds of times more powerful Over the years, two different types of molecular graphics display have been used in molecular modelling. First to be developed were vector devices, which construct pictures using an electron gun to draw lines (or dots) on the screen, in a manner similar to an oscilloscope. Vector devices were the mainstay of molecular modelling for almost two decades but have now been largely superseded by raster devices. These divide the screen into a large number of small 'dots', called pixels. Each pixel can be set to any of a large number of colours, and so by setting each pixel to the appropriate colour it is possible to generate the desired image.

Molecules are most commonly represented on a computer graphics screen using 'stick' or 'space-filling' representations, which are analogous to the Dreiding and Corey-Pauling-Koltun (CPK) mechanical models. Sophisticated variations on these two basic types have

been developed, such as the ability to colour molecules by atomic number and the inclusion of shading and lighting effects, which give 'solid' models a more realistic appearance. Some of the commonly used molecular representations are shown in Figure 1.4 (colour plate section). Computer-generated models do have some advantages when compared with their mechanical counterparts. Of particular importance is the fact that a computer model can be very easily interrogated to provide quantitative information, from simple geometrical measures such as the distance between two atoms to more complex quantities such as the energy or surface area. Quantitative information such as this can be very difficult if not impossible to obtain from a mechanical model. Nevertheless, mechanical models may still be preferred in certain types of situation due to the ease with which they can be manipulated and viewed in three dimensions. A computer screen is inherently two-dimensional, whereas molecules are three-dimensional objects. Nevertheless, some impression of the three-dimensional nature of an object can be represented on a computer screen using techniques such as depth cueing (in which those parts of the object that are further away from the viewer are made less bright) and through the use of perspective. Specialised hardware enables more realistic three-dimensional stereo images to be viewed. In the future 'virtual reality' systems may enable a scientist to interact with a computer-generated molecular model in much the same way that a mechanical model can be manipulated.

Even the most basic computer graphics program provides some standard facilities for the manipulation of models, including the ability to translate, rotate and 'zoom' the model towards and away from the viewer. More sophisticated packages can provide the scientist with quantitative feedback on the effect of altering the structure. For example, as a bond is rotated then the energy of each structure could be calculated and displayed interactively.

For large molecular systems it may not always be desirable to include every single atom in the computer image; the sheer number of atoms can result in a very confusing and cluttered picture. A clearer picture may be achieved by omitting certain atoms (e.g. hydrogen atoms) or by representing groups of atoms as single 'pseudo-atoms' The techniques that have been developed for displaying protein structures nicely illustrate the range of computer graphics representation possible (the use of computational techniques to investigate the structures of proteins is considered in Chapter 10). Proteins are polymers constructed from amino acids, and even a small protein may contain several thousand atoms. One way to produce a clearer picture is to dispense with the explicit representation of any atoms and to represent the protein using a 'ribbon'. Proteins are also commonly represented using the cartoon drawings developed by J Richardson, an example of which is shown in Figure 1.5 (colour plate section). The cylinders in this figure represent an arrangement of amino acids called an α-helix, and the flat arrows an alternative type of regular structure called a β-strand. The regions between the cylinders and the strands have no such regular structure and are represented as 'tubes'.

1.5 Surfaces

Many of the problems that are studied using molecular modelling involve the non-covalent interaction between two or more molecules. The study of such interactions is often facilitated by examining the van der Waals, molecular or accessible surfaces of the molecule. The *van der Waals surface* is simply constructed from the overlapping van der Waals spheres of the atoms, Figure 1.6. It corresponds to a CPK or space-filling model. Let us now consider the approach of a small 'probe' molecule, represented as a single van der Waals sphere, up to the van der Waals surface of a larger molecule. The finite size of the probe sphere means that there will be regions of 'dead space', crevices that are not accessible to the probe as it rolls about on the larger molecule. This is illustrated in Figure 1.6. The amount of dead space increases with the size of the probe; conversely, a probe of zero size would be able to access all of the crevices. The *molecular surface* [Richards 1977] is traced out by the inward-facing part of the probe sphere as it rolls on the van der Waals surface of the molecule. The molecular surface contains two different types of surface element. The *contact surface* corresponds to those regions where the probe is actually in contact with the van der Waals surface of the 'target'. The *re-entrant* surface regions occur where there are crevices that are too narrow for the probe molecule to penetrate. The molecular surface is usually defined using a water molecule as the probe, represented as a sphere of radius 1.4 Å.

The *accessible surface* is also widely used. As originally defined by Lee and Richards [Lee and Richards 1971] this is the surface that is traced by the centre of the probe molecule as it rolls on the van der Waals surface of the molecule (Figure 1.6). The centre of the probe molecule can thus be placed at any point on the accessible surface and not penetrate the van der Waals spheres of any of the atoms in the molecule.

Widely used algorithms for calculating the molecular and accessible surfaces were developed by Connolly [Connolly 1983a, b], and others [e.g. Richmond 1984] have described formulae for the calculation of exact or approximate values of the surface area. There are many ways to represent surfaces, some of which are illustrated in Figure 1.7 (colour plate section). As shown, it may also be possible to endow a surface with a translucent quality, which enables the molecule inside the surface to be displayed. Clipping can also be used to cut through the surface to enable the 'inside' to be viewed. In addition, properties such as the electrostatic potential can be calculated on the surface and represented using an appropriate colour scheme. Useful though these representations are, it is important to remember that the electronic distribution in a molecule formally extends to infinity. The 'hard sphere' representation is often very convenient and has certainly proved very valuable, but it may not be appropriate in all cases [Rouvray 1997, 1999, 2000].

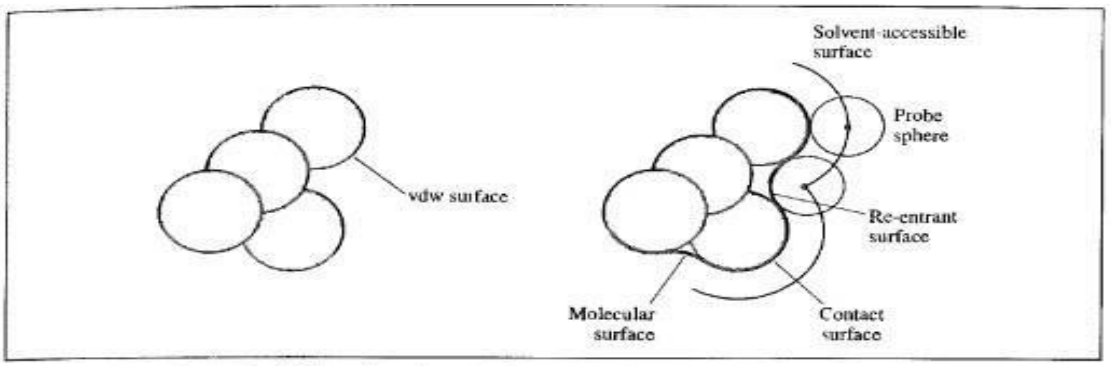


Fig 1.6· The van der Waals (vdw) surface of a molecule corresponds to the outward-facing surfaces of the van der Waals spheres of the atoms. The molecular surface is generated by rolling a spherical probe (usually of radius 1.4 Å to represent a water molecule) on the van der Waals surface. The molecular surface is constructed from contact and re-entrant surface elements. The centre of the probe traces out the accessible surface

Chapter 3 we then build upon this chapter and consider more advanced concepts. Quantum mechanics does, of course, predate the first computers by many years, and it is a tribute to the pioneers in the field that so many of the methods in common use today are based upon their efforts. The early applications were restricted to atomic, diatomic or highly symmetrical systems which could be solved by hand. The development of quantum mechanical techniques that are more generally applicable and that can be implemented on a computer (thereby eliminating the need for much laborious hand calculation) means that quantum mechanics can now be used to perform calculations on molecular systems of real, practical interest. Quantum mechanics explicitly represents the electrons in a calculation, and so it is possible to derive properties that depend upon the electronic distribution and, in particular, to investigate chemical reactions in which bonds are broken and formed. These qualities,

Molecular Mechanics Methods

Molecules modeled as spheres (atoms) connected by springs (bonds) Fast, $>10^6$ atoms
Limited flexibility due to lack of electron treatment Typical applications Simulating biomolecules in explicit solvent/membrane

- Geometry optimization
- Conformational search

Quantum Mechanical Methods

Molecules represented using electron structure (Schrödinger equation) Computationally expensive , <10-100 atoms, depending on method
Highly flexible – any property can in principle be calculated Typical applications Chemical reactions
Spectra
Accurate (gas phase) structures, energies

QUANTUM MECHANICS

Fundamentals of Quantum mechanics

Light- energy- photons/quanta- wave –particle-duality

Schrodinger -Every quantum particle is characterized by wave function Developed a differential equation which describes the evolution of Predicts analytically and precisely the probability of events/outcome (TIME)

QUANTUM MECHANICS

Fundamentals of Quantum mechanics

Light- energy- photons/quanta- wave –particle-duality

Schrodinger -Every quantum particle is characterized by wave function Developed a differential equation which describes the evolution of Predicts analytically and precisely the probability of events/outcome (TIME)

- Represents electrons in a calculation
  - Derive the properties that depend on electronic distribution – particularly the chemical reactions in which bonds are broken and formed

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial x^2}+V(x)\psi=E\psi \qquad \text{or} \qquad \hat{H}\psi=E\psi$$

- H – Hamiltonian operator
- E – energy of the system
- ψ - wave function

- But SE can be used only for very small mol such as H and He
- So approximations must be used in order to extend the utility of the method to polyatomic systems

The starting point for any discussion of quantum mechanics is, of course, the Schrödinger equation. The full, time-dependent form of this equation is

$$\left\{-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)+\mathscr{V}\right\}\Psi(\mathbf{r},t)=i\hbar\frac{\partial\Psi(\mathbf{r},t)}{\partial t} \tag{2.1}$$

Equation (2.1) refers to a single particle (e.g. an electron) of mass m which is moving through space (given by a position vector $\mathbf{r}=x\mathbf{i}+y\mathbf{j}+z\mathbf{k}$) and time ($t$) under the influence of an external field $\mathscr{V}$ (which might be the electrostatic potential due to the nuclei of a molecule). $\hbar$ is Planck's constant divided by 2π and i is the square root of -1. Ψ is the *wavefunction* which characterises the particle's motion; it is from the wavefunction that we can derive various properties of the particle. When the external potential $\mathscr{V}$ is independent of time then the wavefunction can be written as the product of a spatial part and a time part: $\Psi(\mathbf{r},t)=\psi(\mathrm{r})T(t)$. We shall only consider situations where the potential is independent of time, which enables the time-dependent Schrödinger equation to be written in the more familiar, time-independent form:

$$\left\{-\frac{\hbar^2}{2m}\nabla^2+\mathscr{V}\right\}\Psi(\mathbf{r})=E\Psi(\mathbf{r}) \tag{2.2}$$

Here, E is the energy of the particle and we have used the abbreviation ∇^2 (pronounced 'del-squared').

$$\nabla^2=\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2} \tag{2.3}$$

It is usual to abbreviate the left-hand side of Equation (2.1) to $\mathscr{H}\Psi$, where $\mathscr{H}$ is the *Hamiltonian operator*:

$$\mathscr{H}=-\frac{\hbar^2}{2m}\nabla^2+\mathscr{V} \tag{2.4}$$

This reduces the Schrödinger equation to $\mathscr{H}\Psi=E\Psi$. To solve the Schrödinger equation it is necessary to find values of E and functions Ψ such that, when the wavefunction is operated upon by the Hamiltonian, it returns the wavefunction multiplied by the energy. The Schrödinger equation falls into the category of equations known as partial differential eigenvalue equations in which an operator acts on a function (the eigenfunction) and returns the

function multiplied by a scalar (the eigenvalue). A simple example of an eigenvalue equation is:

$$\frac{d}{dx}(y) = ry \tag{2.5}$$

The operator here is d/dx. One eigenfunction of this equation is $y = e^{ax}$ with the eigenvalue r being equal to a. Equation (2.5) is a first-order differential equation. The Schrödinger equation is a second-order differential equation as it involves the second derivative of Ψ. A simple example of an equation of this type is

$$\frac{d^2y}{dx^2} = ry \tag{2.6}$$

The solutions of Equation (2.6) have the form $y = A\cos kx + B\sin kx$, where A, B and k are constants. In the Schrödinger equation Ψ is the eigenfunction and E the eigenvalue.

Time-Dependent Schrodinger Wave Equation

$$i\hbar\frac{\partial}{\partial t}\Psi(x,t) = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\Psi(x,t) + V(x)\Psi(x,t)$$

PHYSICS NOTATION

Total E term

K.E. term

P.E. term

$$\Psi(x,t) = e^{-iEt/\hbar}\psi(x)$$

Time-Independent Schrodinger Wave Equation

$$E\psi(x) = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2}\psi(x) + V(x)\psi(x)$$

Eigen value equation

- Operator (H) acts on function (eigen function) (□)
- Returns the function (□) multiplied by a scalar value (eigen value)(E)
- Hamiltonian operator

2.1.1 Operators

or is thus:

The concept of an operator is an important one in quantum mechanics. The *expectation value* (which we can consider to be the average value) of a quantity such as the energy, position or linear momentum can be determined using an appropriate operator. The most commonly used operator is that for the energy, which is the Hamiltonian operator itself, $\mathscr{H}$. The energy can be determined by calculating the following integral:

$$E = \frac{\int \Psi^* \mathscr{H} \Psi \, d\tau}{\int \Psi^* \Psi \, d\tau} \qquad (2.7)$$

(2.10)

(2.11)

The two integrals in Equation (2.7) are performed over all space (i.e. from $-\infty$ to $+\infty$ in the x, y and z directions). Note the use of the complex conjugate notation (Ψ^*), which reminds us that the wavefunction may be a complex number. This equation can be derived by pre-multiplying both sides of the Schrödinger equation, $\mathscr{H}\Psi = E\Psi$, by the complex conjugate of the wavefunction, Ψ^*, and integrating both sides over all space. Thus:

following

$$\int \Psi^* \mathscr{H} \Psi \, d\tau = \int \Psi^* E \Psi \, d\tau \qquad (2.8)$$

(2.12)

E is a scalar and so can be taken outside the integral, thus leading to Equation (2.7). If the wavefunction is normalised then the denominator in Equation (2.7) will equal 1.

The Hamiltonian operator is composed of two parts that reflect the contributions of kinetic and potential energies to the total energy. The kinetic energy operator is

$$-\frac{\hbar^2}{2m}\nabla^2 \qquad (2.9)$$

and the operator for the potential energy simply involves multiplication by the appropriate expression for the potential energy. For an electron in an isolated atom or molecule the potential energy operator comprises the electrostatic interactions between the electron and the nucleus and the interactions between the electron and the other electrons For a

2.3.1 The Born–Oppenheimer Approximation

It was stated above that the Schrödinger equation cannot be solved exactly for any molecular systems However, it is possible to solve the equation exactly for the simplest molecular

the kinetic and potential energy of the electrons moving in the electrostatic field of the nuclei, together with electron–electron repulsion: $E_{tot} = E(\text{electrons}) + E(\text{nuclei})$.

When the Born–Oppenheimer approximation is used we concentrate on the electronic motions; the nuclei are considered to be fixed. For each arrangement of the nuclei the Schrödinger equation is solved for the electrons alone in the field of the nuclei. If it is desired to change the nuclear positions then it is necessary to add the nuclear repulsion to the electronic energy in order to calculate the total energy of the configuration.

$$\Psi_{tot}(\text{nuclei, electrons}) = \Psi(\text{electrons})\Psi(\text{nuclei}) \tag{2.31}$$

The total energy equals the sum of the nuclear energy (the electrostatic repulsion between the positively charged nuclei) and the electronic energy. The electronic energy comprises

Helium atom

$$\left\{-\frac{\hbar^2}{2m}\nabla_1^2-\frac{Ze^2}{4\pi\varepsilon_0 r_1}-\frac{\hbar^2}{2m}\nabla_2^2-\frac{Ze^2}{4\pi\varepsilon_0 r_2}\right\}\Psi(\mathbf{r}_1,\mathbf{r}_2)=E\Psi(\mathbf{r}_1,\mathbf{r}_2) \tag{2.32}$$

Or, in atomic units,

$$\left\{-\frac{1}{2}\nabla_1^2-\frac{Z}{r_1}-\frac{1}{2}\nabla_2^2-\frac{Z}{r_2}\right\}\Psi(\mathbf{r}_1,\mathbf{r}_2)=E\Psi(\mathbf{r}_1,\mathbf{r}_2) \tag{2.33}$$

We can abbreviate this equation to

$$\{\mathscr{H}_1+\mathscr{H}_2\}\Psi(\mathbf{r}_1,\mathbf{r}_2)=E\Psi(\mathbf{r}_1,\mathbf{r}_2) \tag{2.34}$$

$\mathscr{H}_1$ and $\mathscr{H}_2$ are the individual Hamiltonians for electrons 1 and 2. Let us assume that the

Thermodynamics

Thermodynamics, science of the relationship between heat, work, temperature, and energy. In broad terms, thermodynamics deals with the transfer of energy from one place to another and from one form to another. The key concept is that heat is a form of energy corresponding to a definite amount of mechanical work.

The most important laws of thermodynamics are:

- *The zeroth law of thermodynamics.* When two systems are each in thermal equilibrium with a third system, the first two systems are in thermal equilibrium with each other. This property makes it meaningful to use thermometers as the "third system" and to define a temperature scale.
- *The first law of thermodynamics, or the law of conservation of energy.* The change in a system's internal energy is equal to the difference between heat added to the system from its surroundings and work done by the system on its surroundings.
- *The second law of thermodynamics.* Heat does not flow spontaneously from a colder region to a hotter region, or, equivalently, heat at a given temperature cannot be converted entirely into work. Consequently, the entropy of a closed system, or heat energy per unit temperature, increases over time toward some maximum value. Thus, all closed systems tend toward an equilibrium state in which entropy is at a maximum and no energy is available to do useful work.
- *The third law of thermodynamics.* The entropy of a perfect crystal of an element in its most stable form tends to zero as the temperature approaches absolute zero. This allows an absolute scale for entropy to be established that, from a statistical point of view, determines the degree of randomness or disorder in a system.

Thermodynamics:
→ Describes macroscopic properties of equilibrium systems
→ Entirely Empirical
→ Built on 4 Laws and "simple" mathematics

0th Law ⇒ Defines Temperature (T)
1 st Law ⇒ Defines Energy (U)
2nd Law ⇒ Defines Entropy (S)
3rd Law ⇒ Gives Numerical Value to Entropy

Comparison of first and second law of thermodynamics.

	First Law	Second Law
1.	According to the first law heat and work are of same quality indicating 100% efficiency of a cyclic engine	Work is considered to be a high grade energy where as heat as a low grade energy
2.	Results in the definition of the extensive property; internal energy	Results in the definition of the extensive property; entropy
3.	States that energy of an isolated system can neither be created nor destroyed	States that entropy of an isolated system cannot be destroyed but it can be created
4.	Energy of the universe is constant	The entropy of the universe increases towards a maximum
5.	Energy is conserved in every real process	Energy is degraded in every real process

Definitions:
• System: The part of the Universe that we choose to study
• Surroundings: The rest of the Universe
• Boundary: The surface dividing the System from the Surroundings

Systems can be:
• Open: Mass and Energy can transfer between the System and the Surroundings
• Closed: Energy can transfer between the System and the Surroundings, but NOT mass
• Isolated: Neither Mass nor Energy can transfer between the System and the Surroundings

Closed and Open Systems

A *system* is defined as a quantity of matter or a region in space chosen for study. The mass or region outside the system is called the *surroundings*.

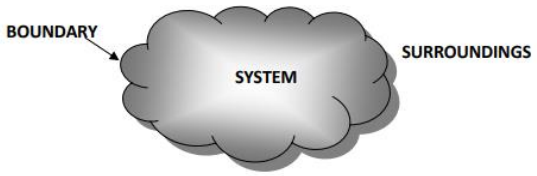


Fig. 1: System, surroundings, and boundary

Describing systems requires:
• A few macroscopic properties: p, T, V, n, m, …
• Knowledge if System is Homogeneous or Heterogeneous
• Knowledge if System is in Equilibrium State
• Knowledge of the number of components

Two classes of Properties:
• Extensive: Depend on the size of the system (n,m,V,…)
• Intensive: Independent of the size of the system (T,p, n V V = ,…)

Adiabatic system: A closed or open system that does not exchange energy with the surroundings by heat.

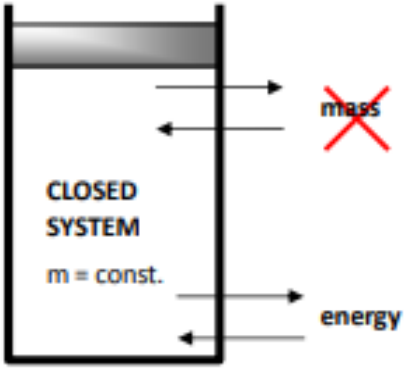


Fig. 2: Closed system, mass cannot cross the boundaries, but energy can.

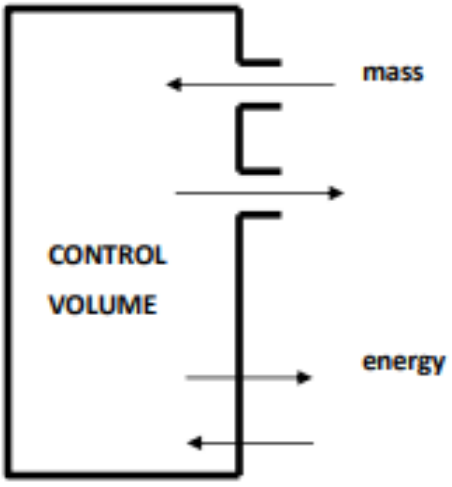


Fig. 3: Control volume, both mass and energy can cross the boundaries.

Energy

In thermodynamics, we deal with change of the total energy only. Thus, the total energy of a system can be assigned a value of zero at some reference point. Total energy of a system has two groups: *macroscopic* and *microscopic*.

Macroscopic forms of energy: forms of energy that a system posses as a whole with respect to some outside reference frame, such as kinetic and potential energy. The macroscopic energy of a system is related to motion and the influence of some external effects such as gravity, magnetism, electricity, and surface tension.

- Kinetic energy: energy that a system posses as a result of its relative motion relative to some reference frame, KE

$$KE = \frac{mV^2}{2} \qquad (kJ)$$

where V is the velocity of the system in (m/s).

- Potential energy: is the energy that a system posses as a result of its elevation in a gravitational field, PE

$$PE = mgz \qquad (kJ)$$

where g is the gravitational acceleration and z is the elevation of the center of gravity of the system relative to some arbitrary reference plane.

Microscopic forms of energy: are those related to molecular structure of a system. They are independent of outside reference frames. The sum of microscopic energy is called the *internal energy, U*.

The total energy of a system consists of the kinetic, potential, and internal energies:

$$E = U + KE + PE = U + \frac{mV^2}{2} + mgz \qquad (kJ)$$

where the contributions of magnetic, electric, nuclear energy are neglected. Internal energy is related to the molecular structure and the degree of molecular activity and it may be viewed as the sum of the kinetic and potential energies of molecules.

- The sum of translational, vibrational, and rotational energies of molecules is the kinetic energy of molecules, and it is also called the *sensible energy*. At higher temperatures, system will have higher sensible energy.
- Internal energy associated with the phase of a system is called *latent heat*. The intermolecular forces are strongest in solids and weakest in gases.
- The internal energy associated with the atomic bonds in a molecule is called *chemical* or *bond energy*. The tremendous amount of energy associated with the bonds within the nucleolus of atom itself is called *atomic energy*.

Energy interactions with a closed system can occur via *heat transfer* and *work*.

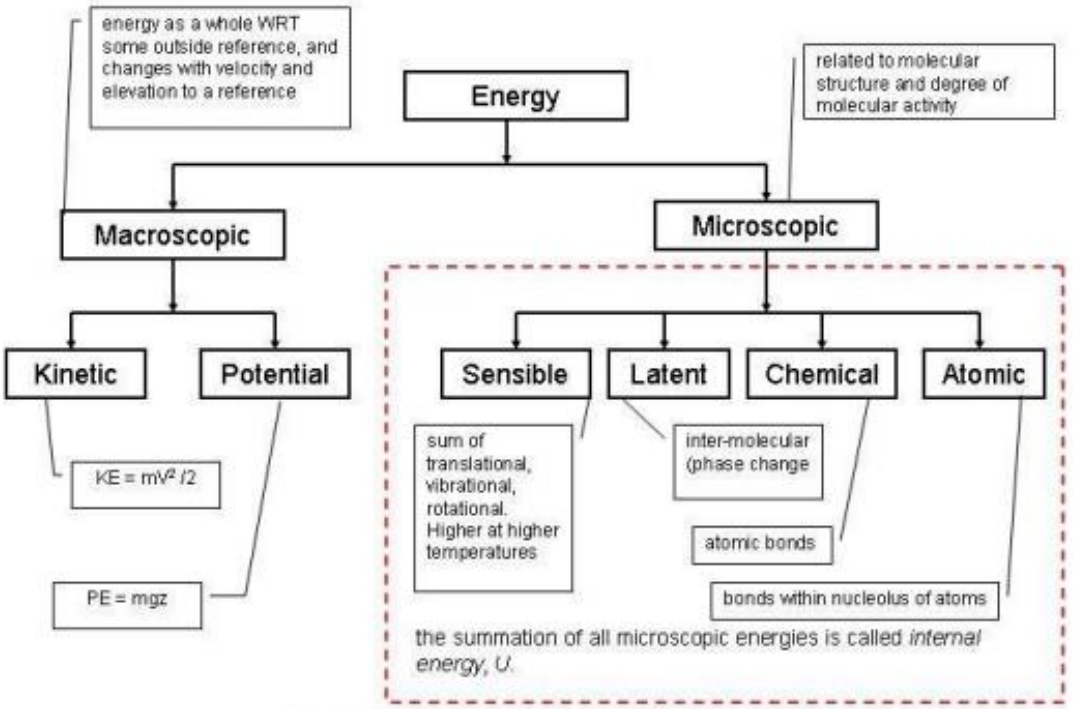


Fig. 1-4: Forms of energy.

$$E = U + KE + PE \qquad (kJ)$$

Properties of a System

Any characteristic of a system is called a *property*. In classical thermodynamics, the substance is assumed to be a *continuum*, homogenous matter with no microscopic holes. This assumption holds as long as the volumes, and length scales are large with respect to the intermolecular spacing.

Intensive properties: are those that are independent of the size (mass) of a system, such as temperature, pressure, and density. They are not additive.

Extensive properties: values that are dependant on size of the system such as mass, volume, and total energy U. They are additive.

- Generally, uppercase letters are used to denote extensive properties (except mass m), and lower case letters are used for intensive properties (except pressure P, temperature T).
- Extensive properties per unit mass are called specific properties, e.g. specific volume (v=V/m).

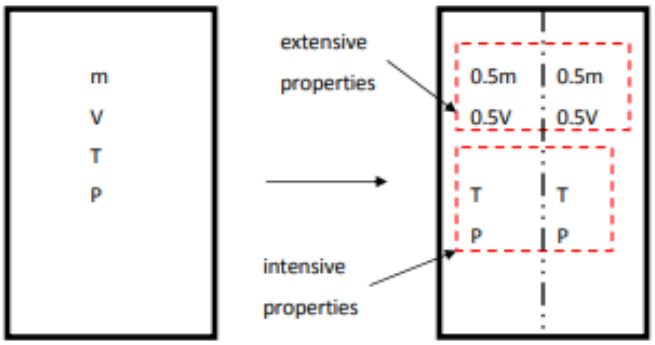


Fig. 1-5: Intensive and extensive properties of a system.

State and Equilibrium

At a given *state*, all the properties of a system have fixed values. Thus, if the value of even one property changes, the state will change to different one.

In an equilibrium state, there are no unbalanced potentials (or driving forces) within the system. A system in equilibrium experiences no changes when it is isolated from its surroundings.

- Thermal equilibrium: when the temperature is the same throughout the entire system.

- Mechanical equilibrium: when there is no change in pressure at any point of the system. However, the pressure may vary within the system due to gravitational effects.
- Phase equilibrium: in a two phase system, when the mass of each phase reaches an equilibrium level.
- Chemical equilibrium: when the chemical composition of a system does not change with time, i.e., no chemical reactions occur.

Processes and Cycles

Any change a system undergoes from one equilibrium state to another is called a *process*, and the series of states through which a system passes during a process is called a *path*.

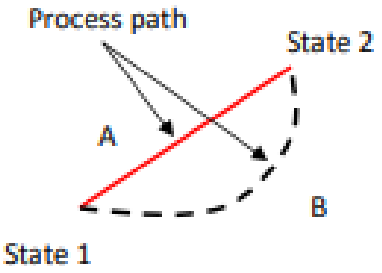


Fig. 6: To specify a process, initial and final states and path must be specified.

Quasi-equilibrium process: can be viewed as a sufficiently slow process that allows the system to adjust itself internally and remains infinitesimally close to an equilibrium state at all times. Quasi-equilibrium process is an idealized process and is not a true representation of the actual process. We model actual processes with quasi-equilibrium ones. Moreover, they serve as standards to which actual processes can be compared.

Process diagrams are used to visualize processes. Note that the process path indicates a series of equilibrium states, and we are not able to specify the states for a non-quasi-equilibrium process.

Prefix *iso-* is used to designate a process for which a particular property is constant.

- Isothermal: is a process during which the temperature remains constant
- Isobaric: is a process during which the pressure remains constant
- Isometric: is process during which the specific volume remains constant.

A system is said to have undergone a *cycle* if it returns to its initial state at the end of the process.

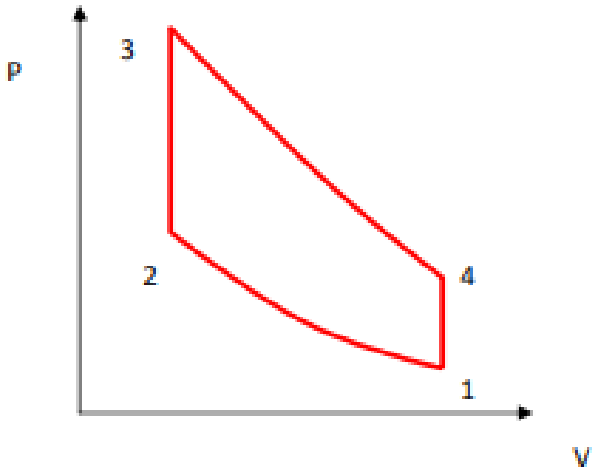


Fig. 1-7: A four-process cycle in a P-V diagram.

The state of a system is described by its properties. The state of a *simple compressible system* is completely specified by two *independent*, intensive properties.

A system is called simple compressible system in the absence of electrical, magnetic, gravitational, motion, and surface tension effects (external force fields).

Independent properties: two properties are independent if one property can be varied while the other one is held constant.

Thermodynamic equilibrium

A particularly important concept is thermodynamic equilibrium, in which there is no tendency for the state of a system to change spontaneously.

For example, the gas in a cylinder with a movable piston will be at equilibrium if the temperature and pressure inside are uniform and if the restraining force on the piston is just sufficient to keep it from moving.

The system can then be made to change to a new state only by an externally imposed change in one of the state functions, such as the temperature by adding heat or the volume by moving the piston.

A sequence of one or more such steps connecting different states of the system is called a process. – reversible and irreversible

Many of the results of thermodynamics are derived from the properties of reversible processes.

Temperature

The concept of temperature is fundamental to any discussion of thermodynamics, but its precise definition is not a simple matter.

In general, when two objects are brought into thermal contact, heat will flow between them until they come into equilibrium with each other. When the flow of heat stops, they are said to be at the same temperature. The zeroth law of thermodynamics formalizes this by asserting that if an object *A* is in simultaneous thermal equilibrium with two other objects *B* and *C*, then *B* and *C* will be in thermal equilibrium with each other if brought into thermal contact.

Object *A* can then play the role of a thermometer through some change in its physical properties with temperature, such as its volume or its electrical resistance.

With the definition of equality of temperature in hand, it is possible to establish a temperature scale by assigning numerical values to certain easily reproducible fixed points. For example, in the Celsius (°C) temperature scale, the freezing point of pure water is arbitrarily assigned a temperature of 0 °C and the boiling point of water the value of 100 °C (in both cases at 1 standard atmosphere; *see* atmospheric pressure). In the Fahrenheit (°F) temperature scale, these same two points are assigned the values 32 °F and 212 °F, respectively. There are absolute temperature scales related to the second law of thermodynamics. The absolute scale related to the Celsius scale is called the Kelvin (K) scale, and that related to the Fahrenheit scale is called the Rankine (°R) scale. These scales are related by the equations K = °C + 273.15, °R = °F + 459.67, and °R = 1.8 K. Zero in both the Kelvin and Rankine scales is at absolute zero.

Thermodynamic potentials are different quantitative measures of the stored energy in a system. Potentials are used to measure the energy changes in systems as they evolve from an initial state to a final state. The potential used depends on the constraints of the system, such as constant temperature or pressure. For example, the Helmholtz and Gibbs energies are the energies available in a system to do useful work when the temperature and volume or the pressure and temperature are fixed, respectively.

The five most well known potentials are:

Name	Symbol	Formula	Natural variables
Internal energy	U	$\int (T\mathrm{d}S - p\mathrm{d}V + \sum_i \mu_i \mathrm{d}N_i)$	$S, V, \{N_i\}$
Helmholtz free energy	F	$U - TS$	$T, V, \{N_i\}$
Enthalpy	H	$U + pV$	$S, p, \{N_i\}$
Gibbs free energy	G	$U + pV - TS$	$T, p, \{N_i\}$
Landau potential, or grand potential	Ω, Φ_G	$U - TS - \sum_i \mu_i N_i$	$T, V, \{\mu_i\}$

where T is the temperature, S the entropy, p the pressure, V the volume, μ the chemical potential, N the number of particles in the system, and i is the count of particles types in the system.

Thermodynamic potentials can be derived from the energy balance equation applied to a thermodynamic system. Other thermodynamic potentials can also be obtained through Legendre transformation.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING
DEPARTMENT OF BIOINFORMATICS

UNIT – 2- SBIA5203 – Biomolecular modeling

Molecular Mechanics Force Field

The "mechanical" molecular model was developed out of a need to describe molecular structures and properties in as practical a manner as possible. The range of applicability of molecular mechanics includes:

- Molecules containing thousands of atoms.
- Organics, oligonucleotides, peptides, and saccharides (metallo-organics and inorganics in some cases).
- Vacuum, implicit, or explicit solvent environments.
- Ground state only.
- Thermodynamic and kinetic (via molecular dynamics) properties.

The great computational speed of molecular mechanics allows for its use in procedures such as molecular dynamics, conformational energy searching, and docking. All the procedures require large numbers of energy evaluations.

Molecular mechanics methods are based on the following principles:

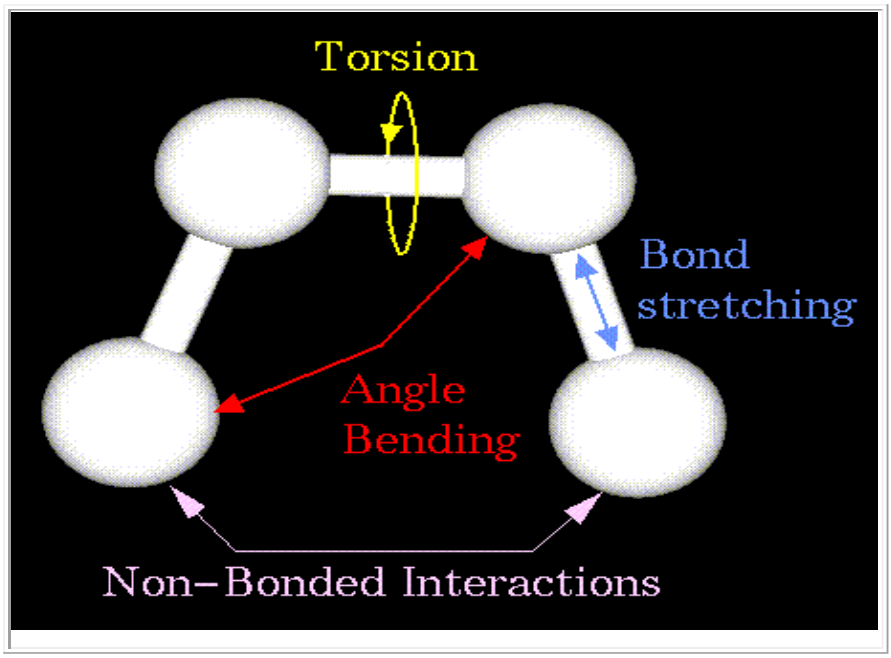
- Nuclei and electrons are lumped into atom-like particles.
- Atom-like particles are spherical (radii obtained from measurements or theory) and have a net charge (obtained from theory).
- Interactions are based on springs and classical potentials.
- Interactions must be preassigned to specific sets of atoms.

Interactions determine the spatial distribution of atom-like particles and their energies

To define a force field one must specify not only the functional form but also the parameters (i.e.the various constants). Two force fields may use an identical functional form yet have very different parameters. A force field should be considered as a single entity; it is not strictly correct to divide the energy into its individual components, let alone to take some of the parameters from one forcefield and mix them with parameters from another force field. The forcefields used in molecular modelling are primarily designed to reproduce structural properties but they can also be used to predict other properties, such as molecular spectra. However, molecular mechanics force fields can rarely predict spectra with great accuracy (although the more recent molecular. mechanics force fields are much better in this regard). A force field is generally designed to predict certain properties and will be

parametrised accordingly. While it is useful to try to predict other quantities which have not been included in the parametrisation process it is not necessarily a failing if a force field is unable to do so. Transferability of the functional form and parameters is an important feature of a forcefield. Transferability means that the same set of parameters can be used to model a series of related molecules, rather than having to define a new set of parameters for each individual molecule. A concept that is common to most force fields is that of an atom type. When preparing the input for a quantum mechanics calculation it is usually necessary to specify the atomic numbers of the nuclei present, together with the geometry of the system and the overall charge and spin multiplicity. For a force field the overall charge and spin multiplicity are not explicitly required, but it is usually necessary to assign an atom type to each atom in the system. The atom type is more than just the atomic number of an atom; it usually con• tains information about its hybridisation state and sometimes the local environment. For example, it is necessary in most force fields to distinguish between sp3 - hybridised carbon atoms (which adopt a tetrahedral geometry), sp2-hybridised carbons (which are trigonal) and sp-hybridised carbons (which are linear).

The mechanical molecular model considers atoms as spheres and bonds as springs. The mathematics of spring deformation can be used to describe the ability of bonds to stretch, bend, and twist:



Non-bonded atoms (greater than two bonds apart) interact through van der Waals attraction, steric repulsion, and electrostatic attraction/repulsion. These

properties are easiest to describe mathematically when atoms are considered as spheres of characteristic radii.

The object of molecular mechanics is to predict the energy associated with a given conformation of a molecule. However, molecular mechanics energies have no meaning as absolute quantities. Only differences in energy between two or more conformations have meaning. A simple molecular mechanics energy equation is given by:

Energy = Stretching Energy + Bending Energy + Torsion Energy + Non-Bonded Interaction Energy

- A **force field** refers to the form and parameters of mathematical functions used to describe the potential energy of a system of particles (typically molecules and atoms).
- calculates the molecular system's potential energy (E) in a given conformation as a sum of individual energy terms.
- where the components of the covalent and noncovalent contributions are given by the following summations:

$$E_{\text{noncovalent}} = E_{\text{electrostatic}} + E_{\text{van der Waals}}$$

- where the components of the covalent and noncovalent contributions are given by the following summations

$$E_{\text{covalent}} = E_{\text{bond}} + E_{\text{angle}} + E_{\text{dihedral}}$$

$$E_{\text{noncovalent}} = E_{\text{electrostatic}} + E_{\text{van der Waals}}$$

- FF is a mathematical function which returns the energy of the system as a function of the conformation of the system.

$$\mathcal{V}(\mathbf{r}^N) = \sum_{\text{bonds}} \frac{k_i}{2}(l_i - l_{i,0})^2 + \sum_{\text{angles}} \frac{k_i}{2}(\theta_i - \theta_{i,0})^2 + \sum_{\text{torsions}} \frac{V_n}{2}(1 + \cos(n\omega - \gamma))$$
$$+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left(4\varepsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \right)$$

$\mathcal{V}(\mathbf{r}^N)$ Potential energy as a function of position r of N particles

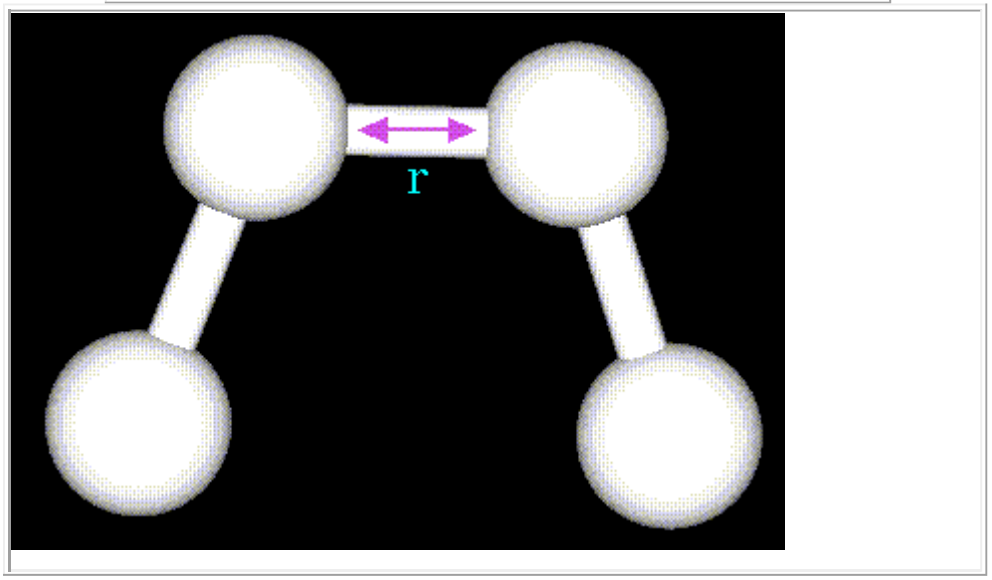
- Reproduce the structural properties such as molecular spectra
- Transferability

These equations together with the data (parameters) required to describe the behavior of different kinds of atoms and bonds, is called a force-field. Many different kinds of force- fields have been developed over the years. Some include additional energy terms that describe other kinds of deformations. Some force-fields account for coupling between bending and stretching in adjacent bonds in order to improve the accuracy of the mechanical model.

The mathematical form of the energy terms varies from force-field to force-field. The more common forms will be described.

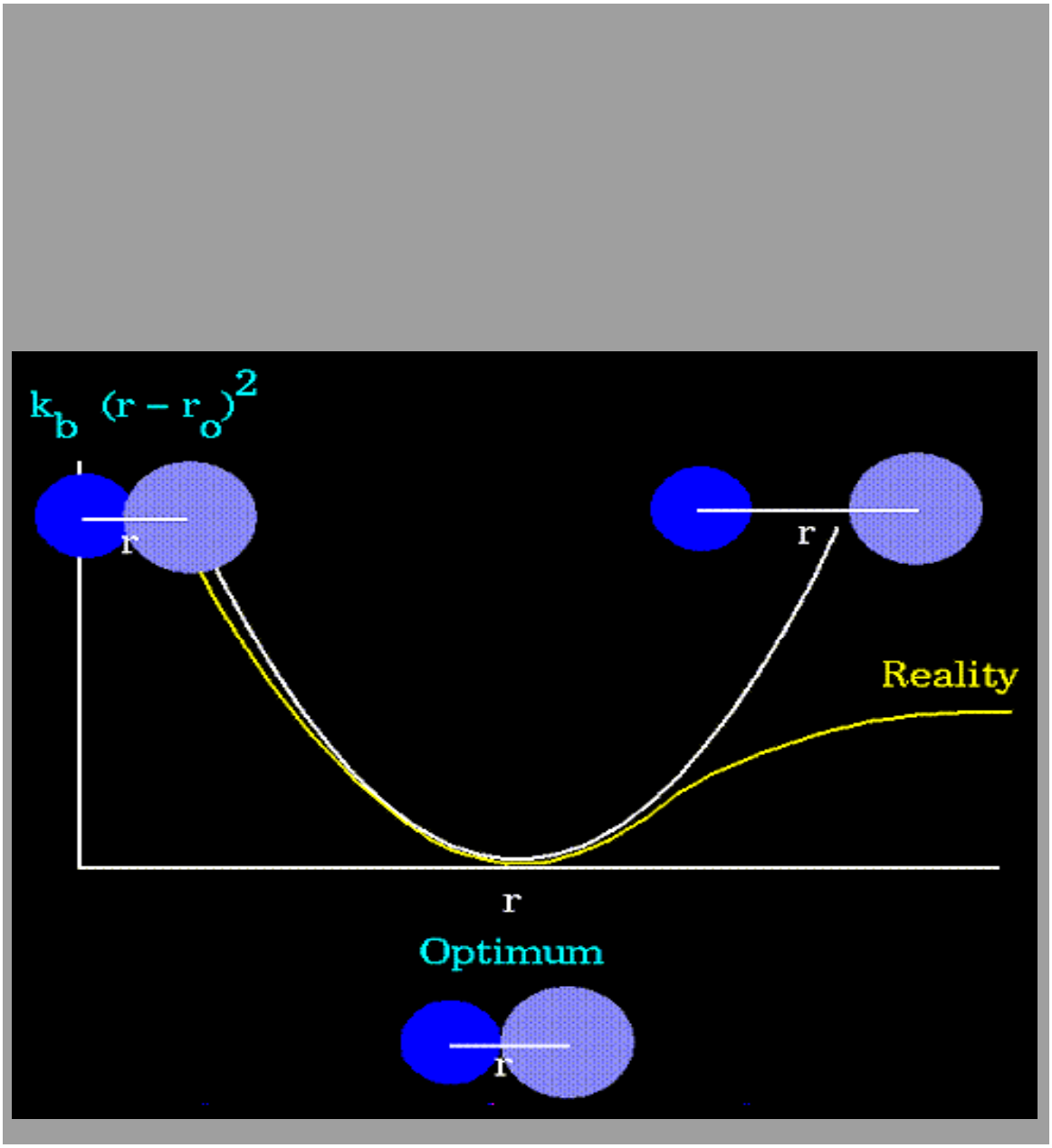
Stretching Energy

$$E = \sum_{bonds} k_b \, (r - r_o)^2$$



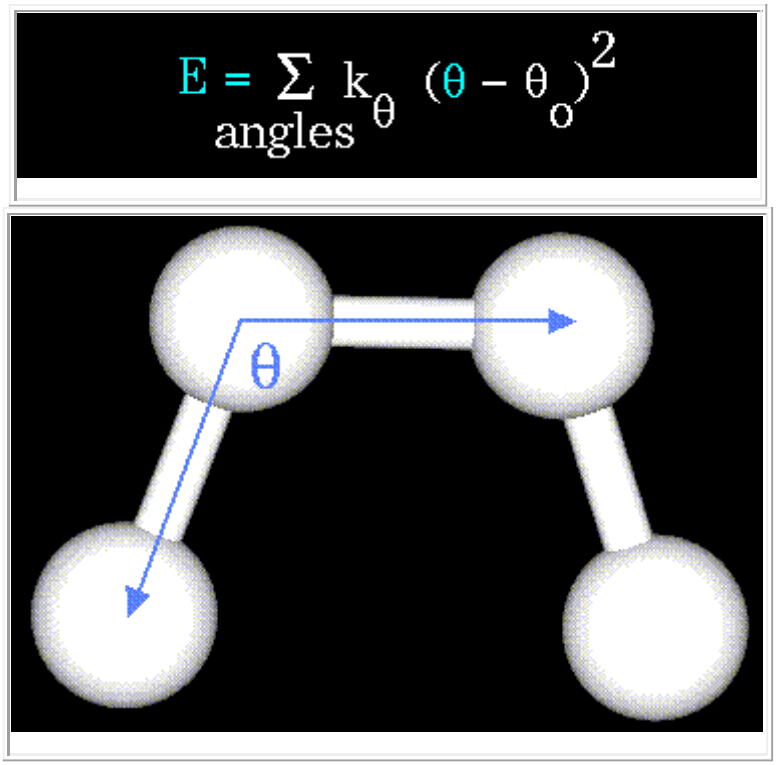
The stretching energy equation is based on Hooke's law. The "k*b*" parameter controls

the stiffness of the bond spring, while "r*o*" defines its equilibrium length. Unique "k*b*" and "r*o*" parameters are assigned to each pair of bonded atoms based on their types (e.g. C-C, C-H, O-C, etc.). This equation estimates the energy associated with vibration about the equilibrium bond length. This is the equation of a parabola, as can be seen in the following plot

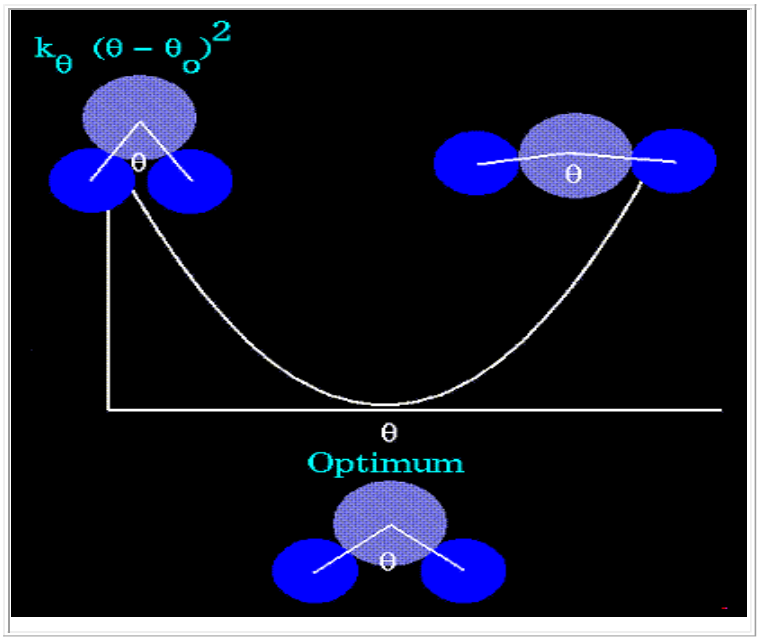


Notice that the model tends to break down as a bond is stretched toward the point of dissociation.

Bending Energy

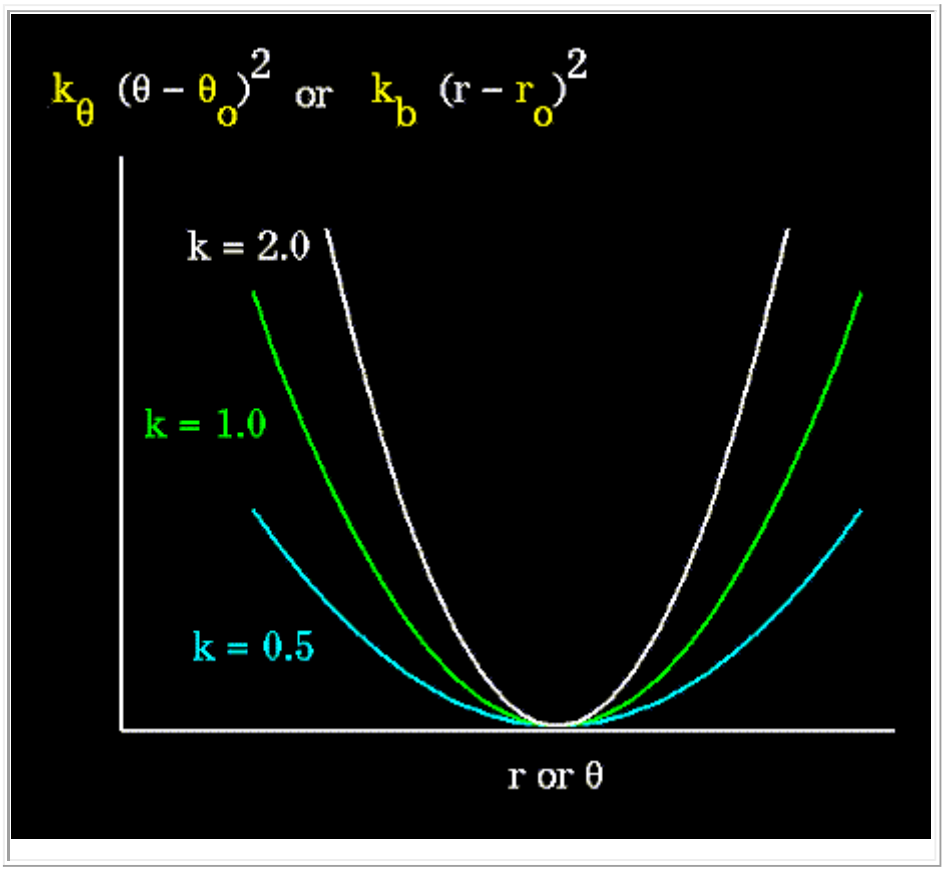


The bending energy equation is also based on Hooke's law. The "k*theta*" parameter controls the stiffness of the angle spring, while "theta*o*" defines its equilibrium angle. This equation estimates the energy associated with vibration about the equilibrium bond angle:

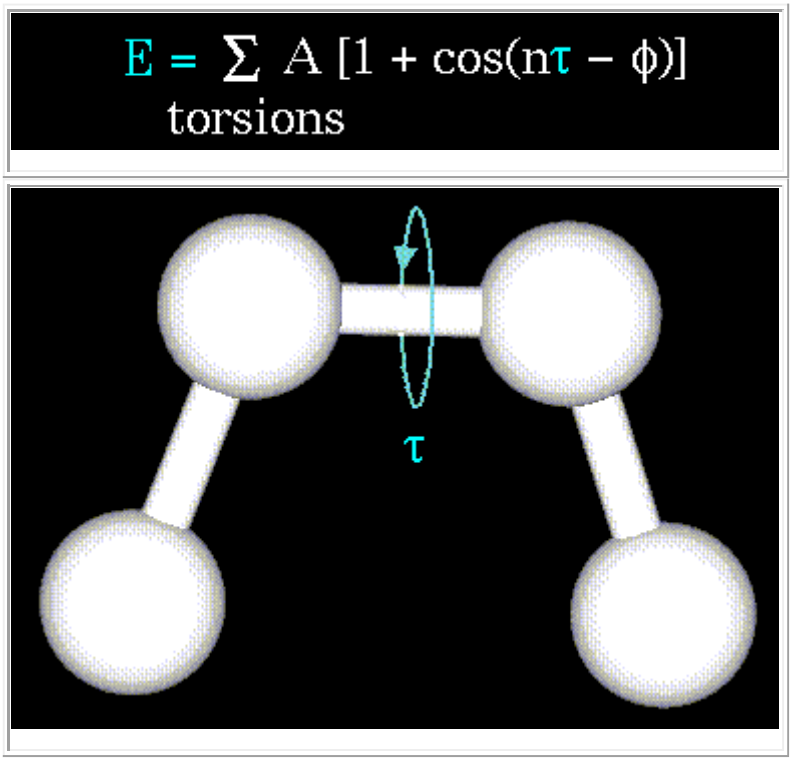


Unique parameters for angle bending are assigned to each bonded triplet of

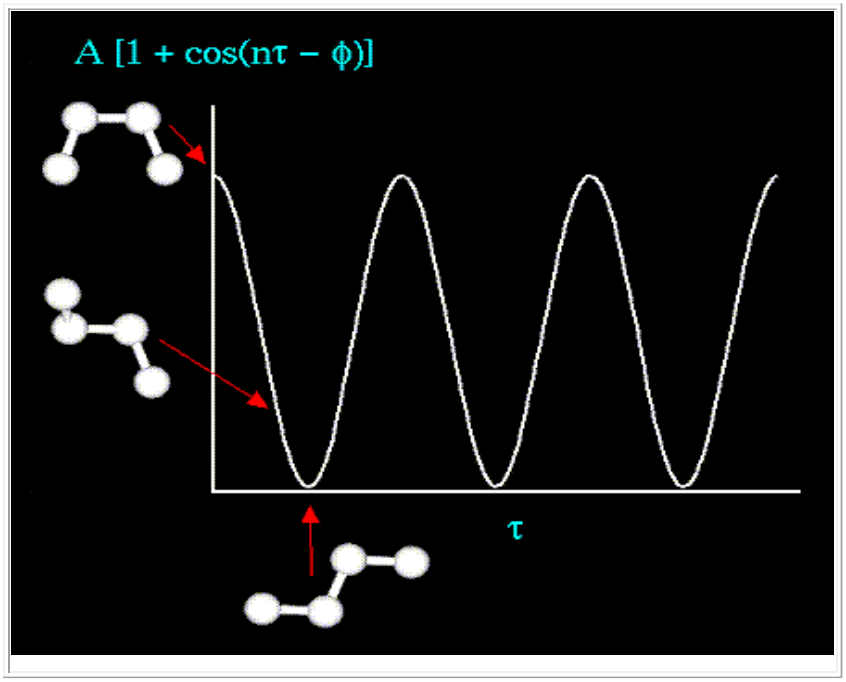
atoms based on their types (e.g. C-C-C, C-O-C, C-C-H, etc.). The effect of the "k*b*" and "k*theta*" parameters is to broaden or steepen the slope of the parabola. The larger the value of "k", the more energy is required to deform an angle (or bond) from its equilibrium value. Shallow potentials are achieved for "k" values between 0.0 and 1.0. The Hookeian potential is shown in the following plot for three values of "k":



Torsion Energy



The torsion energy is modeled by a simple periodic function, as can be seen in the following plot:



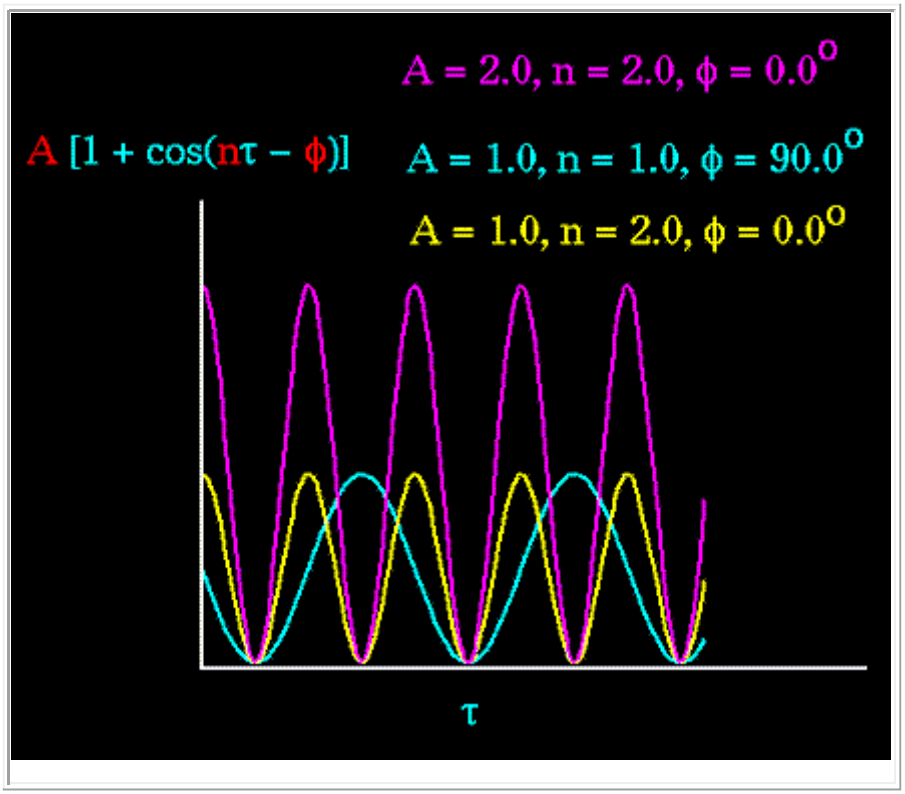
The torsion energy in molecular mechanics is primarily used to correct the remaining

energy terms rather than to represent a physical process. The torsional energy represents the amount of energy that must be added to or subtracted from the Stretching Energy + Bending Energy

+ Non-Bonded Interaction Energy terms to make the total energy agree with experiment or rigorous quantum mechanical calculation for a model dihedral angle (ethane, for example

might be used a a model for any H-C-C-H bond).

The "A" parameter controls the amplitude of the curve, the n parameter controls its periodicity, and "phi" shifts the entire curve along the rotation angle axis (tau). The parameters are determined from curve fitting. Unique parameters for torsional rotation are assigned to each bonded quartet of atoms based on their types (e.g. C-C-C-C, C-O-C-N, H-C- C-H, etc.). Torsion potentials with three combinations of "A", "n", and "phi" are shown in the following plot:



Notice that "n" reflects the type symmetry in the dihedral angle. A CH3-CH3 bond, for example, ought to repeat its energy every 120 degrees. The *cis* conformation of a dihedral angle is assumed to be the zero torsional angle by convention. The parameter phi can be used to synchronize the torsional potential to the initial rotameric state of the molecule whose energy is being

computed.

Cross terms

The presence of cross terms in a forcefield reflects coupling between the internal coordinates. For example, as a bond angle is decreased it is found that the adjacent bonds stretch to reduce the interaction between the 1,3 atoms, as illustrated in Figure.

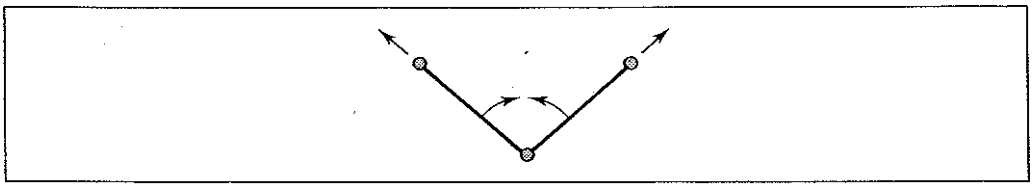
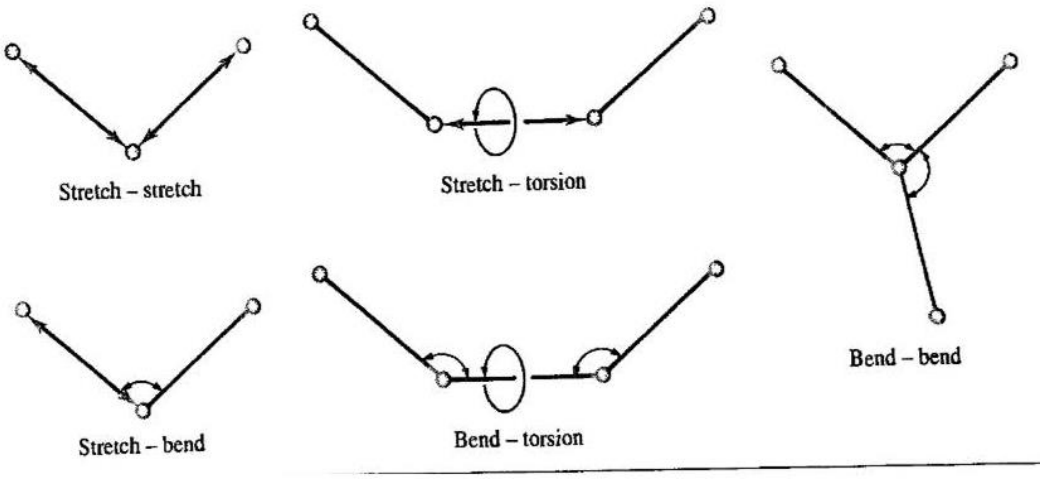


Fig. 4.12: Coupling between the stretching of the bonds as an angle closes.

One should in principle include cross terms between all contributions to a force field. However, only a few cross terms are generally found to be necessary in order to reproduce structural properties accurately; more may be needed to reproduce other properties such as vibrational frequencies, which are more sensitive to the presence of such terms. In general, any interactions involving motions that are far apart in a molecule can usually be set to zero. Most cross terms are functions of two internal coordinates, such as stretch-stretch, stretch-bend and stretch-torsion terms, but cross terms involving more than two internal coordinates such as the bend- bend- torsion have also been used.

Cross terms



Various functional forms are possible for the cross terms. For example,the stretch-stretch cross term between two bonds 1 and 2 can be modelled as:

$$v(l_1, l_2) = \frac{k_{l_1,l_2}}{2}[(l_1 - l_{1,0})(l_2 - l_{2,0})] \quad (4.13)$$

The stretching of the two bonds adjoining an angle could be modelled using an equation of the following form (as in MM2, MM3 and MM4):

$$v(l_1, l_2, \theta) = \frac{k_{l_1,l_2,\theta}}{2}[(l_1 - l_{1,0}) + (l_2 - l_{2,0})](\theta - \theta_0) \quad (4.14)$$

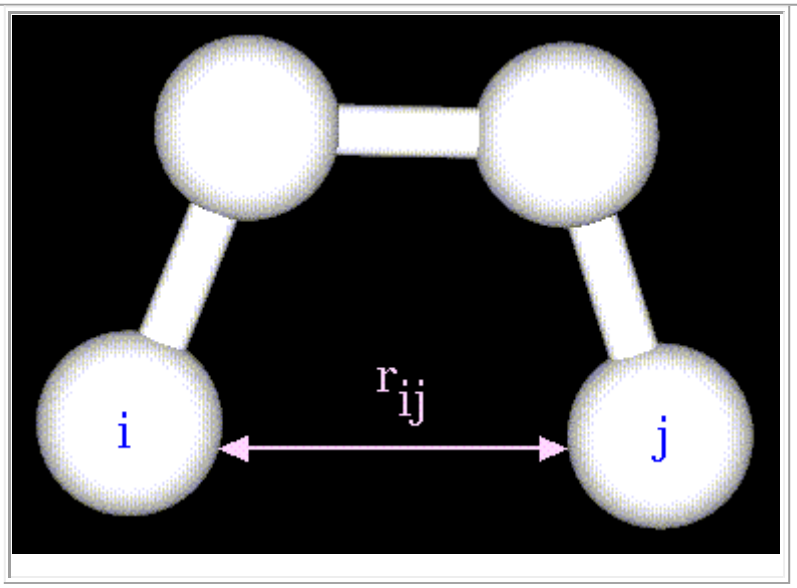
Non-Bonded Energy

Independent molecules and atoms interact through non-bonded forces, which also play an important role in determining the structure of individual molecular species. The non-bonded interactions do not depend upon a specific bonding relationship between atoms. They are 'through-space' interactions and are usually modelled as a function of some inversepower of the distance. The non-bonded terms in a forcefield are usually considered in two groups, one comprising electrostatic interactions and the other van der Waals interactions.

The non-bonded energy represents the pair-wise sum of the energies of all possible interacting non-bonded atoms i and j:

$$E = \sum_i \sum_j \frac{-A_{ij}}{r_{ij}^6} + \frac{B_{ij}}{r_{ij}^{12}} + \sum_i \sum_j \frac{q_i q_j}{r_{ij}}$$

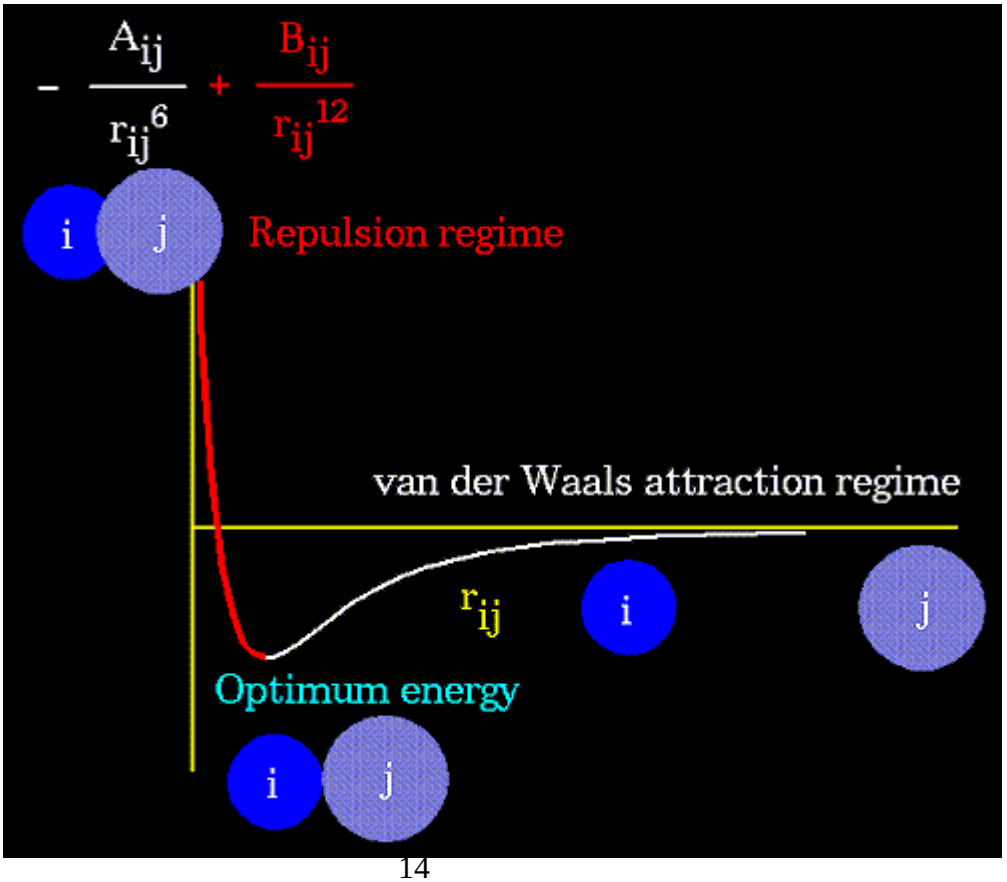
van der Waals term Electrostatic term



The non-bonded energy accounts for repulsion, van der Waals attraction, and electrostatic interactions.

Van der Waals attraction occurs at short range, and rapidly dies off as the interacting atoms move apart by a few Angstroms. Repulsion occurs when the distance between interacting atoms becomes even slightly less than the sum of their contact radii. Repulsion is modeled by an equation that is designed to rapidly blow up at close distances. The energy term that describes attraction/repulsion provides for a smooth transition between these two regimes. These effects are often modeled using a 6-12 equation, as shown in the following plot:

The "A" and "B" parameters control the depth and position (interatomic distance) of the potential energy well for a given pair of non-bonded interacting atoms (e.g. C:C, O:C, O:H, etc.). In effect, "A" determines the degree of "stickiness" of the van der Waals attraction and "B" determines the degree of "hardness" of the atoms (e.g marshmallow-like, billiard ball-like, etc.).


$-\frac{A_{ij}}{r_{ij}^6}+\frac{B_{ij}}{r_{ij}^{12}}$
i
j
Repulsion regime
van der Waals attraction regime
r_{ij}
i
j
Optimum energy
i
j

Vanderwaals interaction

- Dispersive interactions- long range attractive forces
- Due to instantaneous dipoles which arise due to fluctuation in electron clouds
- This can induce a dipole in neighboring atoms – giving rise to an attractive inductive effect

A simple model to explain the dispersive interaction was proposed by Drude. This model consists of 'molecules' with two charges, $+q$ and $-q$, separated by a distance r. The negative charge performs simple harmonic motion with angular frequency ω along the z axis about the stationary positive charge (Figure 4.33). If the force constant for the oscillator is k and if the mass of the oscillating charge is m, then the potential energy of an isolated Drude molecule is $\frac{1}{2}kz^2$, where z is the separation of the two charges. ω is related to the force constant by $\omega = \sqrt{k/m}$. The Schrödinger equation for a Drude molecule is:

$$-\frac{\hbar^2}{2m}\frac{\partial^2\psi}{\partial z^2} + \frac{1}{2}kz^2\psi = E\psi \qquad (4.59)$$

This is the Schrödinger equation for a simple harmonic oscillator. The energies of the system are given by $E_\nu = (\nu + \frac{1}{2}) \times \hbar\omega$ and the zero-point energy is $\frac{1}{2}\hbar\omega$.

Electrostatic interactions

Electrostatic interactions also arise from changes in the charge distribution of a molecule or atom caused by an external field, a process called polarisation. The primary effect of the external electric field (which in our case will be caused by neighbouring molecules) is to induce a dipole in the molecule. The magnitude of the induced dipole moment μind is proportional to the electric field E, with the constant of proportionality being the polarisability a:

$$\boldsymbol{\mu}_{\text{ind}} = \alpha\mathbf{E} \qquad (4.51)$$

The energy of interaction between a dipole $\boldsymbol{\mu}_{\text{ind}}$ and an electric field $\mathbf{E}$ (the induction energy) is determined by calculating the work done in charging the field from zero to E, using the following integral:

$$v(\alpha, E) = -\int_0^E d\mathbf{E}\,\boldsymbol{\mu}_{\text{ind}} = -\int_0^E d\mathbf{E}\,\alpha\mathbf{E} = -\tfrac{1}{2}\alpha E^2 \qquad (4.52)$$

In strong electric fields contributions to the induced dipole moment that are proportional to E^2 or E^3 can also be important, and higher-order moments such as quadrupoles can also be induced. We will not be concerned with such contributions.

The electrostatic contribution is modeled using a Coulombic potential. The electrostatic energy is a function of the charge on the non-bonded atoms, their interatomic distance, and a molecular dielectric expression that accounts for the attenuation of electrostatic interaction by the environment (e.g. solvent or the molecule itself). Often, the molecular dielectric is set to a constant value between 1.0

and 5.0. A linearly varying distance-dependent dielectric (i.e. 1/r) is sometimes used to account for the increase in environmental bulk as the separation distance between interacting atoms increases.

- Central multipole expansion
  - Electronegative elements attract electrons
  - Unequal charge distribution – fractional point charges through out the mol
  - Charges – produce the electrostatic potential
  - Charges restricted to nuclear centres – partial atomic charges

often referred to as *partial atomic charges* or *net atomic charges*. The electrostatic interaction between two molecules (or between different parts of the same molecule) is then calculated as a sum of interactions between pairs of point charges, using Coulomb's law:

$$\mathscr{V} = \sum_{i=1}^{N_A} \sum_{j=1}^{N_B} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \qquad (4.19)$$

N_A and N_B are the numbers of point charges in the two molecules. This approach to the

Conformational analysis

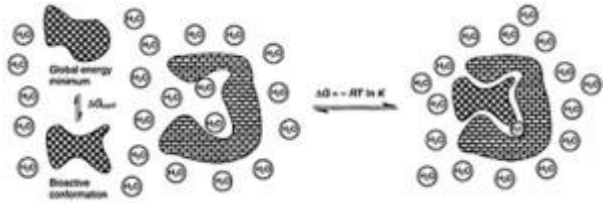
The most important concerns in Medicinal chemistry and pharmaceutical research are structure elucidation, conformational analysis, physicochemical characterization and biological activity determination. The determination of molecular structure is essential as the structure of the molecule predicts the physical, chemical, and biological properties of the molecule.

Conformational search methods find applications in the design of targeted chemical hosts and drug discovery [2]. Conformations are different 3D spatial arrangements of the atoms in a molecule are interconvertible by free rotation of single bonds [3].

The major objective of conformational analysis is to gain insight on conformational characteristic of flexible biomolecules and drugs but to also identify the relation between the role of conformational flexibility and their activity. Therefore, it plays a significant role in computer aided design as well. The significance of conformational analysis not just extends to computational docking and screening but also for lead optimization

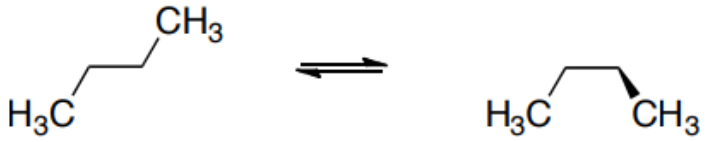
Conformational Analysis: DHR Barton is considered the most important contributor to modern conformational analysis. In 1950, he showed how various substituents at the equatorial and axial positions affect the rate of reactivity of substituted cyclohexanes. Identification of all possible minimum-energy structures (conformations) of a molecule is the goal of conformational analysis [5].

Conformational analysis is a computational method in which restraints are used such that the molecule presumes a conformation similar to the rigid template molecule. Conformational analysis is a difficult problem because even simple molecules may have a large number of conformational isomers. The usual strategy in conformational analysis is to use a search algorithm to generate a series of initial conformations.

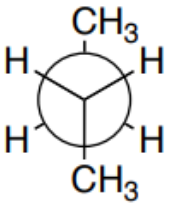


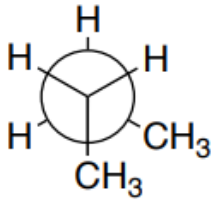
Each of these in turn is then subjected to energy minimization in order to derive the associated minimum energy structure. Global minimum-energy conformation is the conformation with the lowest energy. It is not imperative that the global energy minimum conformation is the bioactive conformation of the drug. Most drugs being flexible molecules can by means of distortions and rotations about rotatable bonds adopt large number of conformations. Pharmacophore is a collection of steric and electronic features that are essential to ensure optimal communication between the specific biological targets (Receptor/Enzyme) so as to illict a biological response.

An important aspect of organic compounds is that the compound is not static, but rather has conformational freedom by rotating, stretching and bending about bonds. Each different arrangement in space of the atoms is called a "Conformer" (a less used term is a "Rotamer" if change is caused by a bond rotation) Different conformers can have vastly different energies and the relative proportion of each conformer is related to the energy difference between them.



Lower Energy



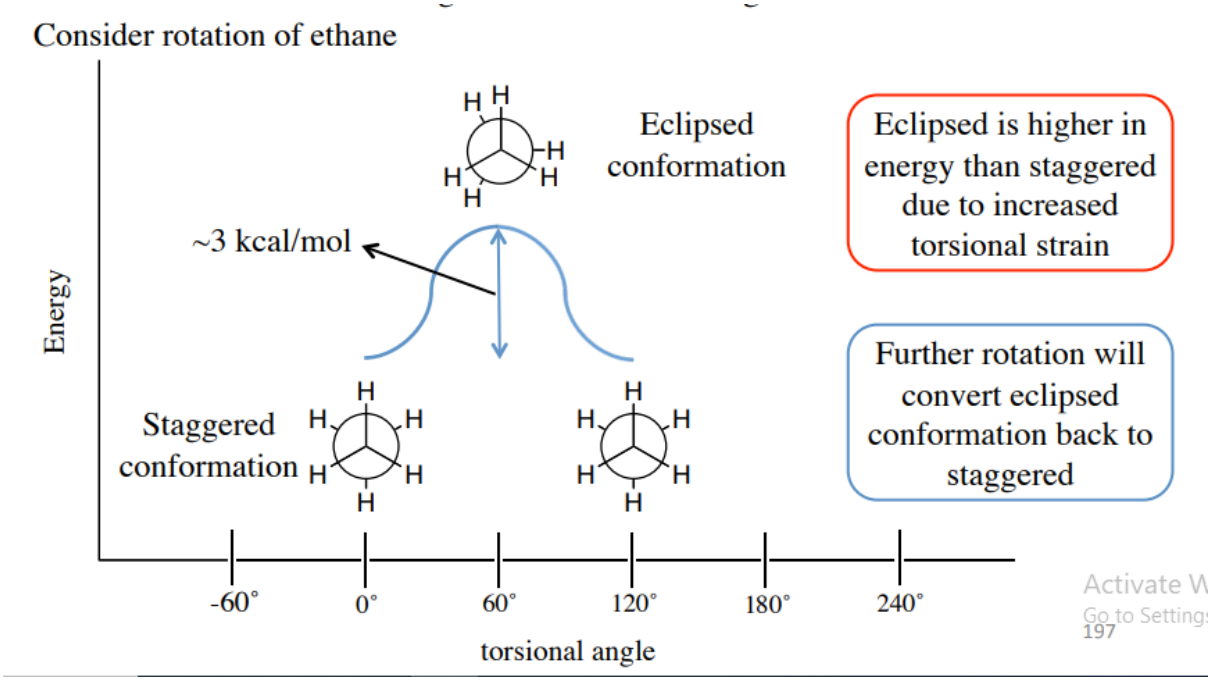


To observe the spatial arrangement of atoms in space, often draw conformers in a Newman projection

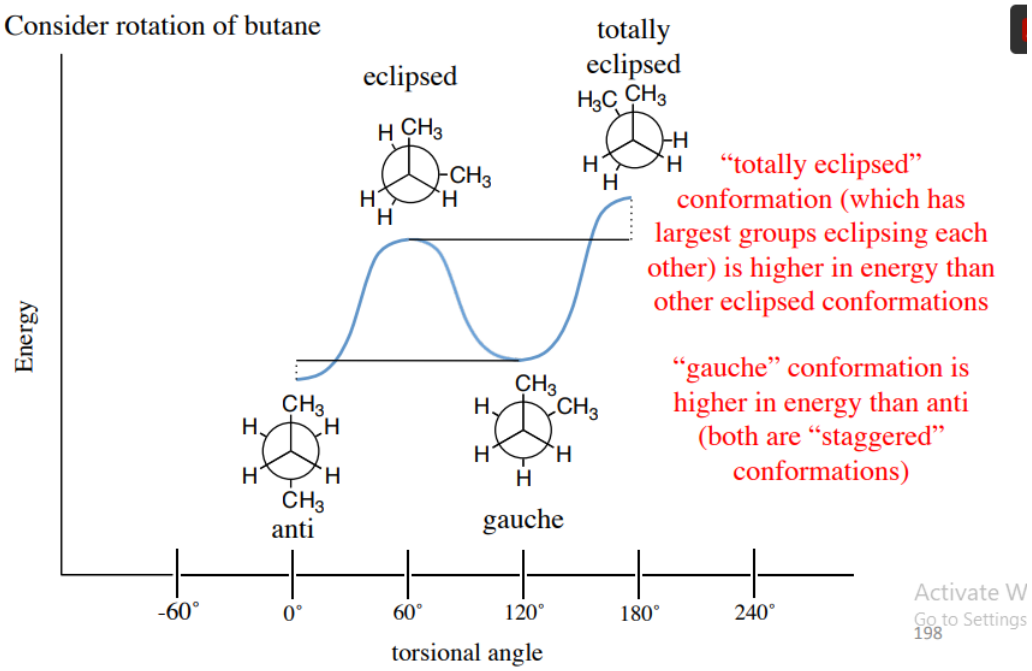
- **Conformations** - Different spatial arrangements that a molecule can adopt due to rotation about sigma bonds.
- **Staggered** - A low energy conformation where the bonds on adjacent atoms bisect each other (60^o dihedral angle), maximizing the separation.
- **Eclipsed** - A high energy conformation where the bonds on adjacent atoms are aligned with each other (0^o dihedral angle).

Conformational Analysis Conformers will be of different energy due to strain
Sources of strain are generally categorized in one of three types:

1) **Torsional strain** E n e r g y torsional angle -60˚ 0˚ 60˚ 120˚ 180˚ 240˚ Torsional strain is due to interactions as groups change relative position with a change in torsional bond angle



2) **van der Waals strain** Another source of strain is when groups are placed in positions closer than the sum of their van der Waals radii

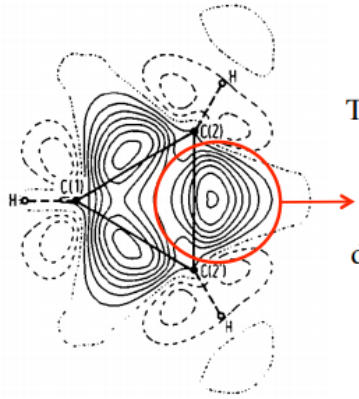


The difference in energy thus affects the amount of each conformer present

3) **Angle strain** A third source of strain is due to angle strain (molecules that are forced to have a bond angle far from ideal [~109.5˚ for sp3])

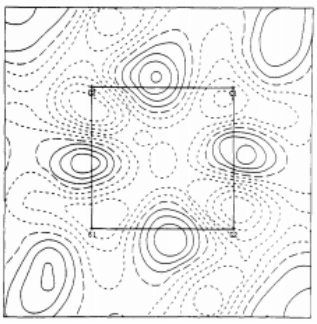
This angle strain is due to forcing the electron density in bonds at angles that are not ideal

Have observed this effect with cyclopropane where the 3 carbons are forced to be coplanar



The electron density truly does form "bent" bonds, bonding electron density is not along internuclear axis

Observe same "bent" bond effect in cyclobutane, here the 4 carbons need not be in the same plane but the angle strain would still be large



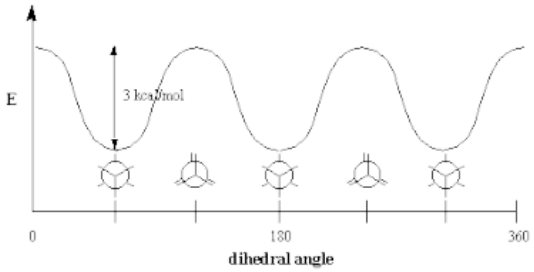
As rings become larger, however, would not expect this type of "bent" bonds due to lower angle strain

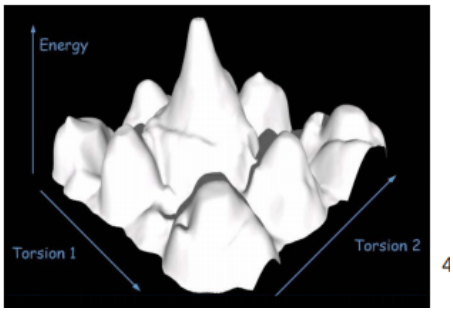
Conformational Analysis

- Properties of molecules depend on their three-dimensional structures (i.e. conformations)
- Conformational analysis is the study of the conformations of a molecule and their influence on its properties
- Conformational analysis is used in drug design to search conformations of small molecules (putative drugs)
- In protein folding this is used to find protein 3D structure with minimal energy that usually corresponds to biologically active structure
- A key component of conformational analysis is the conformational search, the objective of which is to identify 'preferred' conformations, i.e. conformations with low energies

Conformations of Small Molecules

- The conformations of a molecule are those arrangements of its atoms that can be interconverted purely by rotation about single bonds
- Different conformations have different energy because of electronic changes, steric clashes, non-bonding interactions, etc.
- For molecules with one rotatable bond the “conformational potential surface” consists of the curve representing the molecular energy as a function of the dihedral (torsion) angle. The minima of the curve correspond to low energy conformations.
- For molecule with two rotatable bonds the total energy is represented as a function of two variable torsion angles





Conformation Searching Methods

- The following algorithms are used to find global minima on conformational potential surface:
  - Systematic search (also called 'grid search', 'exhaustive search' or 'brute force search')
  - Random search (Monte Carlo search)
  - Simulated annealing
  - Genetic algorithms
  - Distance-geometry algorithms
  - The fragment approach
  - Chain growth
  - Rule-based systems

Conformational Analysis

- Conformation generally means structural arrangement
- Conformational analysis is needed to identify the ideal conformation of a molecule

$$N = \left(\frac{360}{\delta}\right)^{nbonds}$$

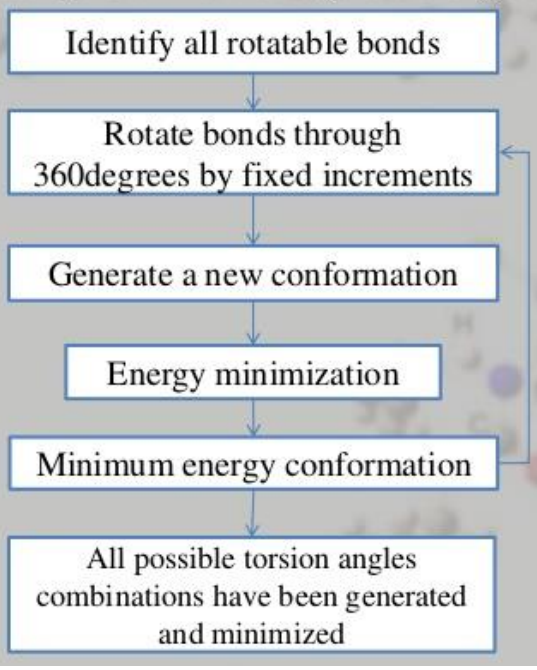
N = # conformations
δ = rotation increment in degrees
nbonds = # of rotatable bonds (degrees of freedom)

- The biological activity of molecules is strongly dependent on their conformation
- Done by exploring the energy surface of a molecule and determining the conformation with minimum energy
- Needed:
  - Conformational space
  - Search method
  - An energy determination method

Systematic Method

- Simplest
- Deterministic
- *Grid search*

Assumption: All bond lengths and angles remain fixed throughout the calculation

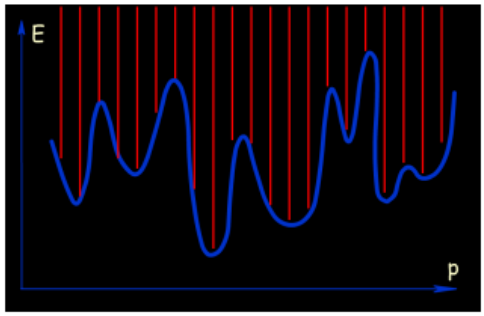


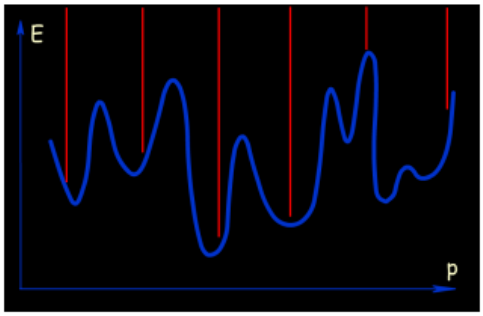
Advantages

- A deterministic approach, hence a systematic exploration of conformational space
- Can be used for small molecules or systems with 10-15 bonds

Systematic Search

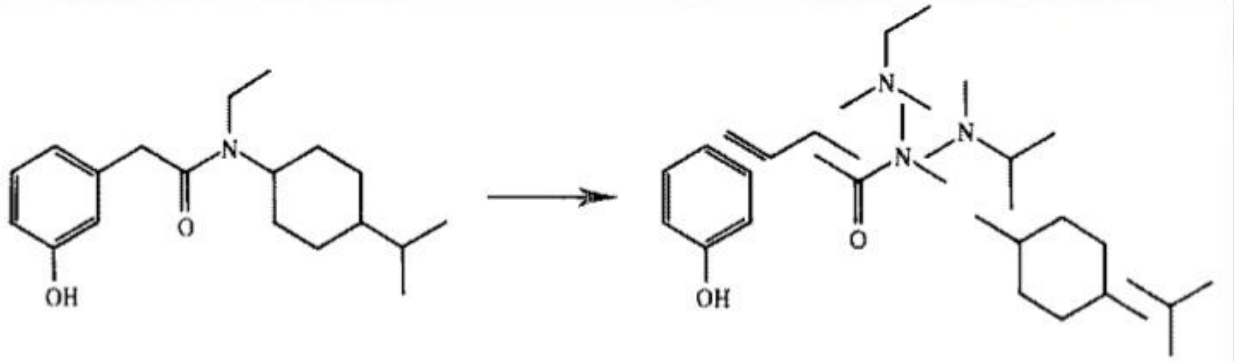
- A simple method for exploring the conformational potential surface is systematic scanning of all geometries of the molecule
- Within this systematic searches dihedral angles are changed systematically by specified increment value
- Too big increment values can result in missing the global minima while small increment values significantly increase algorithm complexity
- Complexity of systematic search grows rapidly with number of rotatable bonds (~$(360/m)^n$ where n is number of rotatable bonds and m is increment value)





Model-building method

- Construct conformations of molecules by joining together three-dimensional structures of molecular fragments



- Substructure search algorithm
  - Substructure searching is finding a mapping for a query to a target molecule.
  - In other words, no bonds are broken and no new bonds are formed.

Genetic Algorithms

- Genetic algorithms are inspired by biological evolution, they stem from the observation that the evolution process tends to produce increasingly well-adapted populations
- At the start, the algorithm creates random population of individuals (individual = conformer). Each individual is encoded by genes (gen = value of one dihedral angle).
- Fitness of individuals is evaluated (i.e. energy is calculated) and individuals with high fitness (low energy) are reproduced to make new generation of individuals
- Some genes are mutated (some their dihedral angle values are randomly changed) and crossover is performed (values of dihedral angle are switched between some pairs of individuals)
- The process is repeated until it converges (i.e. individuals of new generation have almost the same energy as several previous generations)
- A genetic algorithm is an effective way of generating large number of low-energy conformers. However, there is no guarantee that a global minimum will be found
- Practical test have shown genetic algorithms to be superior to simulated annealing and random search

Other Algorithms

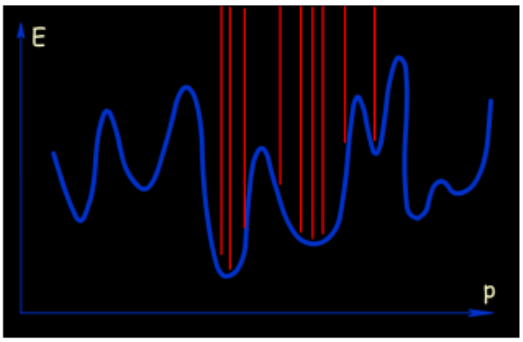
- Distance-geometry algorithms are used if some atom distances in the structure are known (typically from NMR experiment). These distances are used as constraints within a conformation search.
- The fragment approach is based on optimization of one part of a molecule at a time. For instance, protein side chains are individually optimized and subsequently the backbone is optimized while keeping the side chains fixed.
- Within the chain growth algorithm, the full molecule is built up one unit at a time. As each unit is added, its conformation is searched without changing the rest of the chain.
- Rule-based systems try to identify certain subsequences of amino acids that tend to have a particular secondary structure (α-helices, β-threads, etc.). These sections can be held rigid while the conformations of the connecting fragments are searched.

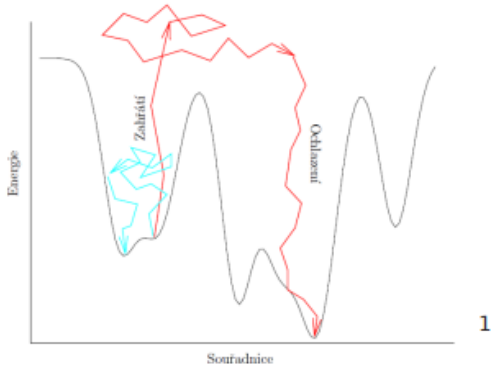
Molecular Dynamics

- Each atom of a protein has a potential energy and therefore feels a force exerted on it
- This force can be used to simulate motion of protein atoms using equations of motion from classical mechanics
- Position of each atom can be calculated along a series of extremely small time steps and the resulting series of snapshots of structures over time is called a trajectory
- Molecular dynamics simulation should mimic behaviour of real molecule
- Within molecular dynamics simulation the molecule adopts different conformations thus it searches conformational space

Simulated Annealing

- A simulated annealing algorithm is a molecular dynamics simulation, in which the amount of kinetic energy in the molecule (the simulation temperature) is high at the beginning and it slowly decreases over the course of the simulation
- At the beginning of the simulation, many high-energy structures are being examined and high-energy barriers can be crossed. At the end of calculation, only structures that are close to the best-known low-energy structures are examined
- This algorithm is most effective for finding low-energy conformers that are similar in shape to the starting geometry





1

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING
DEPARTMENT OF BIOINFORMATICS

UNIT – 3- SBIA5203 – Biomolecular modeling

Energy minimization

In the field of computational chemistry, energy minimization (also called energy optimization, geometry minimization, or geometry optimization) is the process of finding an arrangement in space of a collection of atoms where, according to some computational model of chemical bonding, the net inter-atomic force on each atom is acceptably close to zero and the position on the potential energy surface (PES) is a stationary point. The collection of atoms might be a single molecule, an ion, a condensed phase, a transition state or even a collection of any of these. The computational model of chemical bonding might, for example, be quantum mechanics.

The motivation for performing a geometry optimization is the physical significance of the obtained structure: optimized structures often correspond to a substance as it is found in nature and the geometry of such a structure can be used in a variety of experimental and theoretical investigations in the fields of chemical structure, thermodynamics, chemical kinetics, spectroscopy and others.

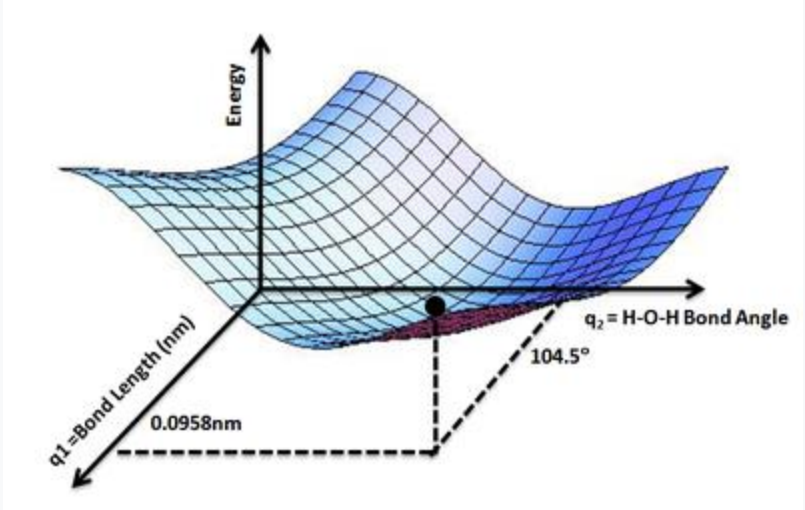
Typically, but not always, the process seeks to find the geometry of a particular arrangement of the atoms that represents a local or global energy minimum. Instead of searching for global energy minimum, it might be desirable to optimize to a transition state, that is, a saddle point on the potential energy surface. Additionally, certain coordinates (such as a chemical bond length) might be fixed during the optimization.

- Energy minimization methods can precisely locate minimum energy conformations by mathematically "homing in" on the energy function minima (one at a time).
- The goal of energy minimization is to find a route (consisting of variation of the intramolecular degrees of freedom) from an initial conformation to the nearest minimum energy conformation using the smallest number of calculations possible.
- The way in which the energy varies with the coordinates is usually referred to as PES or hyper surface
- Energy of any conformation is a function of its internal or cartesian coordinates
- N atoms – energy is a function of 3N-6 internal coordinates or 3N cartesian coordinates
- Changes in the energy are a function of its nuclear coordinates.

Potential energy Surface

A **potential energy surface** (**PES**) describes the energy of a system, especially a collection of atoms, in terms of certain parameters, normally the positions of the atoms. The surface might define the energy as a function of one or more coordinates; if there is only one coordinate, the surface is called a *potential energy curve* or energy profile. An example is the Morse/Long-range potential.

It is helpful to use the analogy of a landscape: for a system with two degrees of freedom (e.g. two bond lengths), the value of the energy (analogy: the height of the land) is a function of two bond lengths (analogy: the coordinates of the position on the ground).[1]



PES for water molecule: Shows the energy minimum corresponding to optimized molecular structure for water- O-H bond length of 0.0958nm and H-O-H bond angle of 104.5°

The PES concept finds application in fields such as chemistry and physics, especially in the theoretical sub-branches of these subjects. It can be used to theoretically explore properties of structures composed of atoms, for example, finding the minimum energy shape of a molecule or computing the rates of a chemical reaction.

The geometry of a set of atoms can be described by a vector, **r**, whose elements represent the atom positions. The vector **r** could be the set of the Cartesian coordinates of the atoms, or could also be a set of inter-atomic distances and angles.

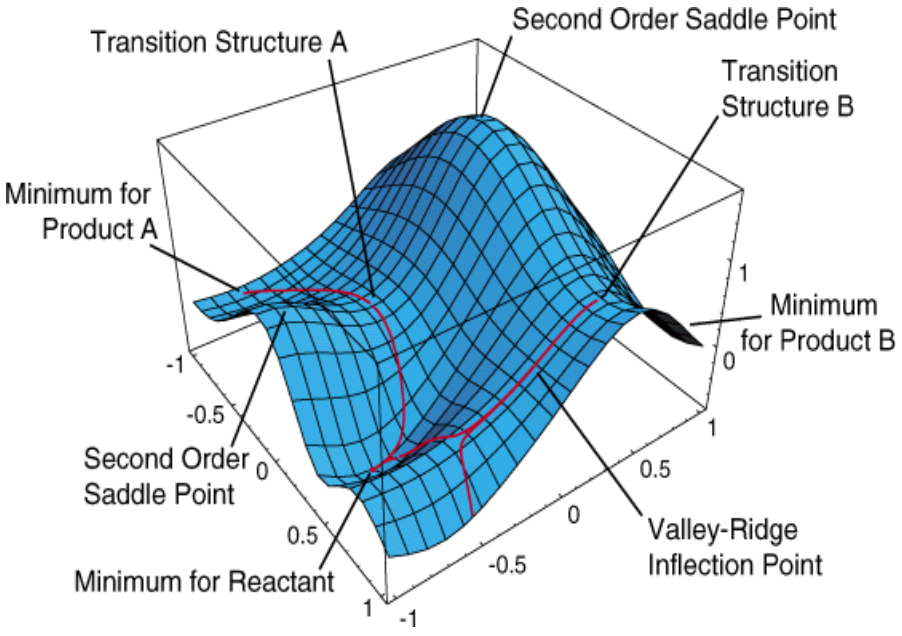
Given **r**, the energy as a function of the positions, $E(\mathbf{r})$, is the value of $E(\mathbf{r})$ for all **r** of interest. Using the landscape analogy from the introduction, *E* gives the height on the "energy landscape" so that the concept of a potential energy *surface* arises.

To study a chemical reaction using the PES as a function of atomic positions, it is necessary to calculate the energy for every atomic arrangement of interest. Methods of calculating the energy of a particular atomic arrangement of atoms are well described in the computational chemistry article, and the emphasis here will be on finding approximations of $E(\mathbf{r})$ to yield fine-grained energy-position information.

For very simple chemical systems or when simplifying approximations are made about inter-atomic interactions, it is sometimes possible to use an analytically derived expression for the energy as a function of the atomic positions.

- Changes in the energy of a system can be considered as movements on a multidimensional surface called energy surface.
- Changes in the energy funtion of its nuclear coordinates.
- Movement of the nuclei influences change in energy

- Mathematical function that gives the energy of a molecule as a function of its geometry
- Energy is plotted on the vertical axis, geometric coordinates (e.g bond lengths, valence angles, etc.) are plotted on the horizontal axes
- A PES can be thought of it as a hilly landscape, with valleys, mountain passes and peaks
- Real PES have many dimensions, but key feature can be represented by a 3 dimensional PES



- Equilibrium molecular structures correspond to the positions of the minima in the valleys on a PES
- Energetics of reactions can be calculated from the energies or altitudes of the minima for reactants and products
- A reaction path connects reactants and products through a mountain pass
- A transition structure is the highest point on the lowest energy path
- Reaction rates can be obtained from the height and profile of the potential energy surface around the transition structure
- The shape of the valley around a minimum determines the vibrational spectrum
- Each electronic state of a molecule has a separate potential energy surface, and the separation between these surfaces yields the electronic spectrum
- Properties of molecules such as dipole moment, polarizability, NMR shielding, etc. depend on the response of the energy to applied electric and magnetic fields

- Minima, lowest – global energy minima
- Minimization algorithms
- Highest point in the pathway between 2 minima is saddle point represents the transition state
- Minima and saddle points are stationary states on PES where the first derivative of energy function is 0
- E = f(x)
- E is a function of coordinates either cartesian or internal
- At minimum the first derivatives are zero and the second derivatives are all positive

5.1.1 Energy Minimisation: Statement of the Problem

The minimisation problem can be formally stated as follows: given a function f which depends on one or more independent variables $x_1, x_2, \ldots, x_i$, find the values of those variables where f has a minimum value. At a minimum point the first derivative of the function with respect to each of the variables is zero and the second derivatives are all positive:

$$\frac{\partial f}{\partial x_i} = 0; \qquad \frac{\partial^2 f}{\partial x_i^2} > 0 \tag{5.1}$$

The functions of most interest to us will be the quantum mechanics or molecular mechanics energy with the variables x_i being the Cartesian or the internal coordinates of the atoms.

- Minimization algorithm can go down hill on the energy surface and hence locate minima that is nearest to starting point

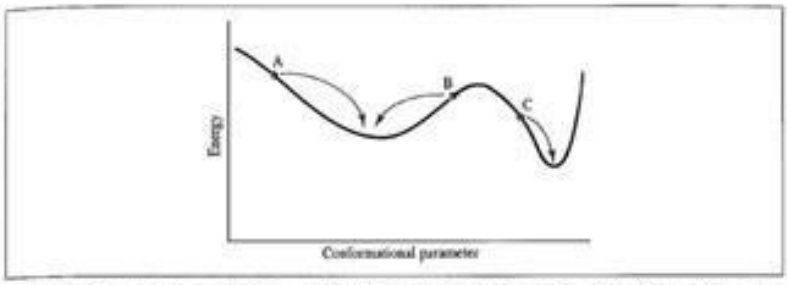


Fig. 3.3: A schematic one-dimensional energy surface. Minimisation methods move downhill to the nearest minimum. The statistical weight of the narrow, deep minimum may be less than a broad minimum which is higher in energy

The input to a minimisation program consists of a set of initial coordinates for the system. The initial coordinates may come from a variety of sources. They may be obtained from an experimental technique, such as X-ray crystallography or NMR. In other cases a theoretical method is employed, such as a conformational search algorithm. A combination of experimental and theoretical approaches may also be used. For example, to study the

- $x_{new} = x_{old}$ + correction .
- In the equation, x_{new} refers to the value of the geometry at the next step (for example, moving from step 1 to 2 in the figure),
- x_{old} refers to the geometry at the current step, and correction is some adjustment made to the geometry.
- In all these methods, a numerical test is applied to the new geometry (xnew) to decide if a minimum is reached.

Minimization Methods

Several methods exist for finding a minimum of an arbitrary continuous function. One way to classify a minimization method is based on what kind of derivatives are used to guide the minimization. In this classification, we can distinguish between:

- methods that use no derivatives (function values, such as the energy)
- methods that use only first derivatives (slope, or force)
- methods that use second derivatives (curvature, or force constants)

Derivative-free methods

In general, methods that use no derivatives spend the least amount of time at each point but require the most steps to reach the minimum. Methods such as the simplex minimization, simulated annealing, and optimization by genetic algorithms fall into this category. They are used in computational chemistry rarely because of their slow convergence. However, some docking programs implement simplex minimizers.

First Derivative Methods

Methods that use only the first derivatives are sometimes used in computational chemistry, especially for preliminary minimization of very large systems. The first derivative tells the downhill direction and also suggests how large steps should be taken when stepping down the hill (large steps on a steep slope, small steps on flatter areas that are hopefully near the minimum). Methods such as deepest descent and a variety of conjugate gradient minimization algorithms belong to this group. In the steepest descent method in one-dimension, the new position along the line is obtained as:

$$r_{new} = r_{old} - \frac{\left(\frac{\partial E}{\partial r}\right)_{r=r_{old}}}{1/stepsize}$$

.

In the case of multiple variables, the steepest descent step is calculated from the vector-matrix equation

$$\vec{r}_{new} = \vec{r}_{old} - stepsize \cdot \nabla E$$

The upside-down capital delta in the latter equation is called "nabla" or "del" and stands for the gradient. The gradient is a vector formed from individual partial derivatives of the function in all search directions. Specifically, for a search in two dimensions, the vector-matrix matrix equation becomes

$$\begin{pmatrix} x \\ y \end{pmatrix}_{new} = \begin{pmatrix} x \\ y \end{pmatrix}_{old} - stepsize \bullet \begin{pmatrix} \frac{\partial E}{\partial x} \\ \frac{\partial E}{\partial y} \end{pmatrix}_{old}$$

If the derivative of a function can be calculated analytically, the time spent at each step is not much higher than the time needed for evaluation of the energy at this step. However, if

the derivative must be calculated numerically, the program needs to carry out additional energy calculations near each point to obtain the derivative.

Second Derivative Methods

Methods that use both the fist derivative and the second derivative can reach the minimum in the least number of steps because the curvature information allows estimation of where the minimum is. The simplest method in this category is the Newton-Rhapson method. The Newton-Rhapson method used the gradient and the Hessian (H) at the current point R the new point R_new that is closer to the true minimum. In case of one-dimensional line search, the new position is calculated as R_new = R - F/H, or more explicitly:

$$r_{new} = r_{old} - \frac{\left(\frac{\partial E}{\partial r}\right)_{old}}{\left(\frac{\partial^2 E}{\partial r^2}\right)_{old}}$$

In the case of multiple variables, the coordinates of the new point are calculated from analogous vector-matrix expression:

$$\vec{r}_{new} = \vec{r}_{old} - \mathbf{H}^{-1}\nabla E$$

Notice that the Newton-Rhapson method requires the inversion of the second derivative matrix. Because the inversion of large matrixes can be very demanding on computer CPU and memory, a simple Newton-Rhapson algorithm becomes quickly slow or unfeasible for larger systems. In practice, optimization can be carried out calculating the second derivative matrix only once at the beginning of the optimization and updating it subsequentially using first derivatives. Alternatively, quasi-Newton methods such as the BFGS, which estimate approximate Hessian based on gradients on two consequtive search points can be used. Many common optimization routines in quantum chemistry that are used for locating the minimum energy geometries and optimize wave functions use quasi-Newton methods.

Minimisation algorithms

Simplex algorithm
* Not a gradient minimization method.
* Used mainly for very crude, high energy starting structures.

Steepest descent minimiser
* Follows the gradient of the energy function (b) at each step.
  This results in successive steps that are always mutually perpendicular, which can lead to backtracking.
* Works best when the gradient is large (far from a minimum).
* Tends to have poor convergence because the gradient becomes smaller as a minimum is approached.

Conjugate gradient and Powell minimiser
* Remembers the gradients calculated from previous steps to help reduce backtracking.
* Generally finds a minimum in fewer steps than Steepest Descent.
* May encounter problems when the initial conformation is far from a minimum.

Newton-Raphson and BFGS minimiser
* Predicts the location of a minimum, and heads in that direction.
* Calculates (Newton-Raphson) or approximates (BFGS) the second derivatives in **A**.
  Storage of the **A** term can require substantial amounts of computer memory.
* May find a minimum in fewer steps than the gradient-only methods.
* May encounter serious problems when the initial conformation is far from a minimum.

Minimisation algorithms

The **steepest descent** minimiser uses the numerically calculated first derivative of the energy function to approach the energy minimum. The energy is calculcated for the initial geometry and then again when one of the atoms has been moved in a small increment. This process will be repeated for all atoms which finally are moved to new positions downhill on the energy surface. The optimisation process is slow near the minimum. Usually used as a first run (e.g. start of crystallographic refinement).

The **conjugate gradient** method accumulates the information about the function from one iteration to the next. With this proceeding, the reverse of the progress made in an earlier iteration can be avoided. Computational effort and storage requirements are greater than for steepest descent, but conjugate gradient is the method of choice for larger systems.

The **Powell** method is very similar to the conjugate gradient approach. It is faster in finding convergence and suitable for a variety of problems. However, torsion angles may sometimes be modified dramatically.

The **Newton-Raphson** minimiser also uses the curvature of the energy function to identify the search direction. Its efficiency increases as convergenc eis approached. Main disadvantage is the computational effort and large storage requirements for calculating larger systems. Also, for structures with high starin, the minimisation process can become instable. This method is thus not recommended as the first method in a refinement procedure.

Computer simulation

Computer simulation is the process of mathematical modelling, performed on a computer, which is designed to predict the behaviour of or the outcome of a real-world or physical system. Since they allow to check the reliability of chosen mathematical models, computer simulations have become a useful tool for the mathematical modeling of many natural systems in physics (computationalphysics), astrophysics, climatology, chemistry, biology and manufactur ing, as well as human systems in economics, psychology, social science, health care and engineering. Simulation of a system is represented as the running of the system's model. It can be used to explore and gain new insights into new technology and to estimate the performance of systems too complex for analytical solutions.

A computer model is the algorithms and equations used to capture the behavior of the system being modeled. By contrast, computer simulation is the actual running of the program that contains these equations or algorithms. Simulation, therefore, is the process of running a model. Thus one would not "build a simulation"; instead, one would "build a model", and then either "run the model" or equivalently "run a simulation"

Benefits

- Gain greater understanding of a process
- Identify problem areas or bottlenecks in processes
- Evaluate effect of systems or process changes such as demand, resources, supply, and constraints
- Identify actions needed upstream or downstream relative to a given operation, organization, or activity to either improve or mitigate processes or events
- Evaluate impact of changes in policy prior to implementation

Types

- Discrete Models – Changes to the system occur at specific times
- Continuous Models – The state of the system changes continuously over time
- Mixed Models – Contains both discrete and continuous elements

Types of Data/Information Needed to Develop a Simulation Model:

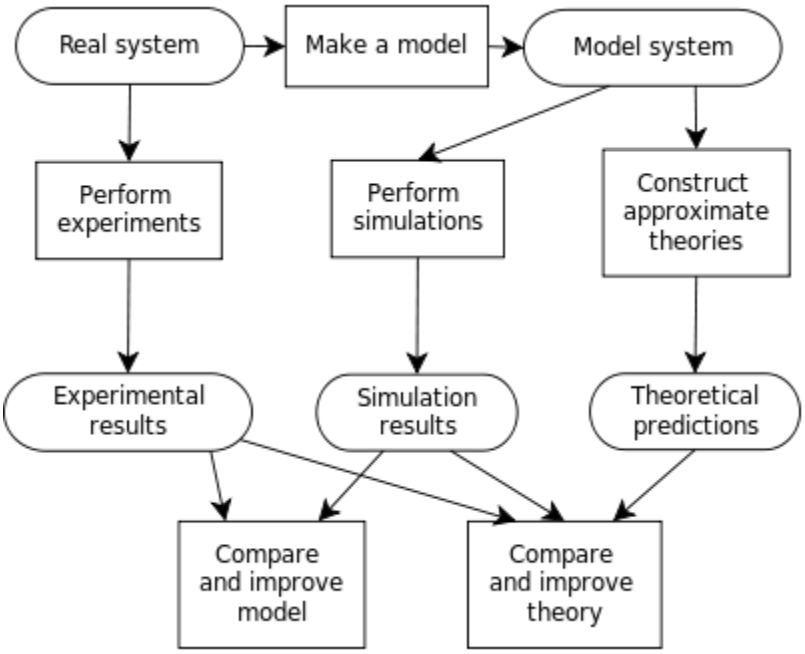
- The overall process flow and its associated resources
- What is being produced, served, or acted upon by the process (entities)
- Frequency at which the entities arrive in the process
- How long do individual steps in the process take
- Probability distributions that characterize real life uncertainties and variations in the process

- Computer simulation is the use of a computer to represent the dynamic responses of one system by the behavior of another system modeled after it.
- A simulation uses a mathematical description, or model, of a real system in the form of a computer program.
- This model is composed of equations that duplicate the functional relationships within the real system.

- When the program is run, the resulting mathematical dynamics form an analog of the behavior of the real system, with the results presented in the form of data.
- A simulation can also take the form of a computer-graphics image that represents dynamic processes in an animated sequence.
- Computer simulations have become a useful part of mathematical modeling of many natural systems in physics, astrophysics, chemistry, biology, climatology, psychology, social science, etc

USES

- Computer simulations are used to study the dynamic behavior of objects or systems in response to conditions that cannot be easily or safely applied in real life.
- Simulations are especially useful in enabling observers to measure and predict how the functioning of an entire system may be affected by altering individual components within that system.
- Simulations have great military applications also. Many uses for a computer simulation can be found within various scientific fields of study such as meteorology, physical sciences, etc



Process of building a computer model, and the interplay between experiment, simulation, and theory.

Advantages of Computer Simulation

- To test a system without having to create the system for real (Building real-life systems can be expensive, and take a long time)
- To predict what might happen to a system in the future (An accurate model allows us to go forward in virtual time to see what the system will be doing in the future)
- To train people to use a system without putting them at risk (Learning to fly an airplane is very difficult and mistake will be made. In a real plane mistakes could be fatal!)
- To investigate a system in great detail (A model of a system can be zoomed in/out or rotated. Time can be stopped, rewound, etc.)

22

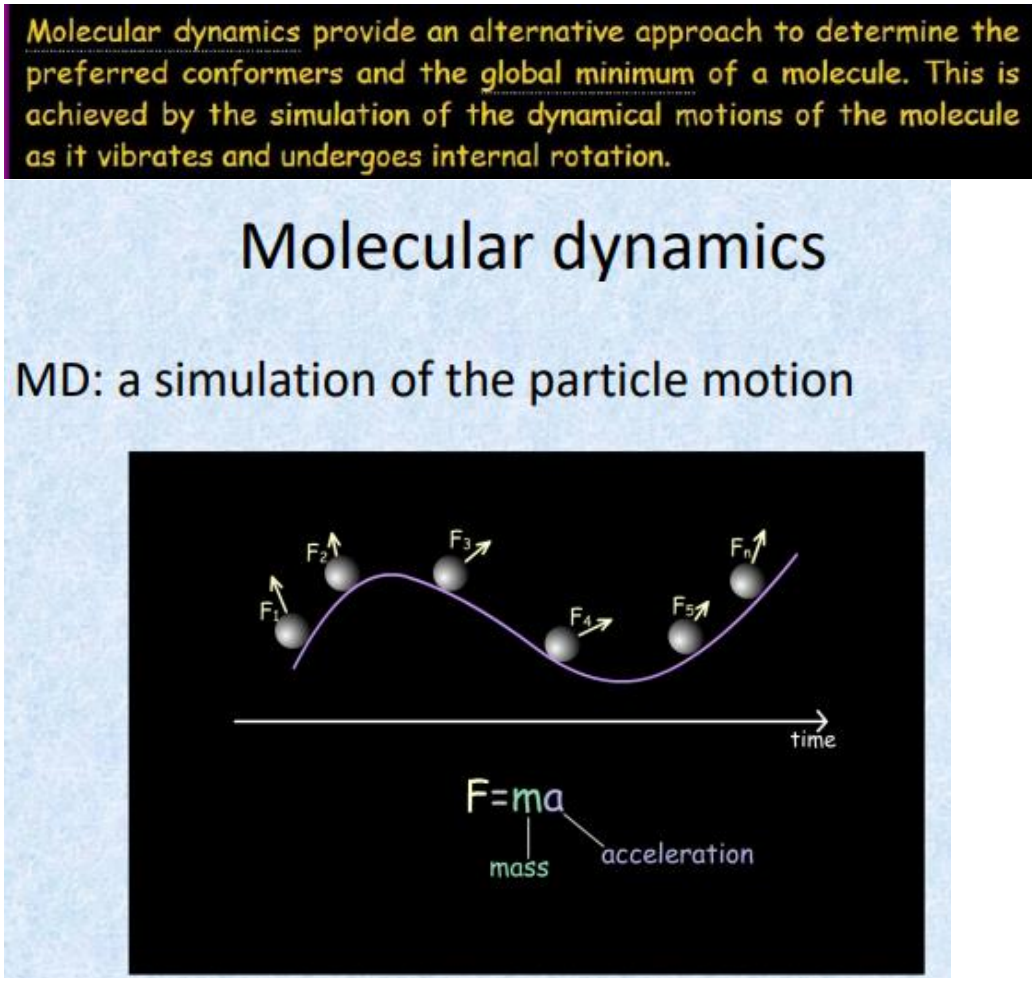
Disadvantages of Computer Simulation

- It depends upon the skills of the model maker
- Time consuming and high cost
- Sometimes, it does not give the best solution for system

Basic Simulation Techniques

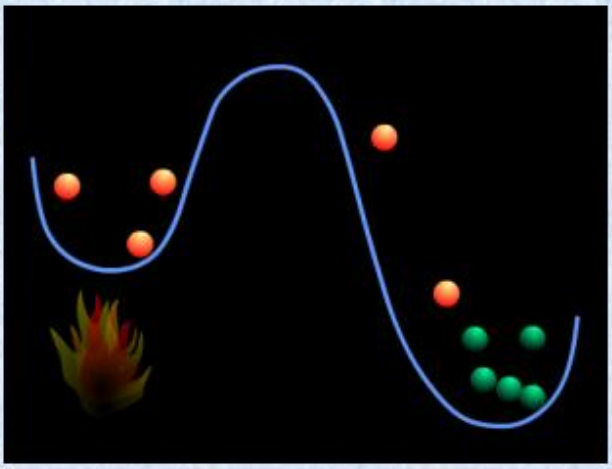
To explore the energy landscape described by the molecular mechanics force field, *i.e.* to sample molecular conformations, a simulation is required. This is also the route to relate the microscopic movements and positions of the atoms to the macroscopic or thermodynamic quantities that can be measured experimentally. There are two major simulation methods to sample biomolecular systems: molecular dynamics (MD) and Monte Carlo (MC)

Molecular dynamics (MD) is a computer simulation method for analyzing the physical movements of atoms and molecules. The atoms and molecules are allowed to interact for a fixed period of time, giving a view of the dynamic "evolution" of the system.



Molecular dynamics

The motion (determined by the temperature) allows conformational changes



Molecular dynamics

- Calculates the time dependent behaviour of a molecular system
- Provides detailed information on the fluctuations and conformational changes of macromolecules
- Routinely used to investigate the structure, dynamics and thermodynamics of biological molecules
- Used in the determination of structures from x-ray and NMR experiments

In a molecular dynamics (MD) simulation it is possible to explore the macroscopic properties of a system

The connection between microscopic simulation and macroscopic properties is made through statistical mechanics

Allows to study both thermodynamic properties and time dependent (kinetic) phenomenon

A MD simulation is practically carried out through the application of the Newton law:

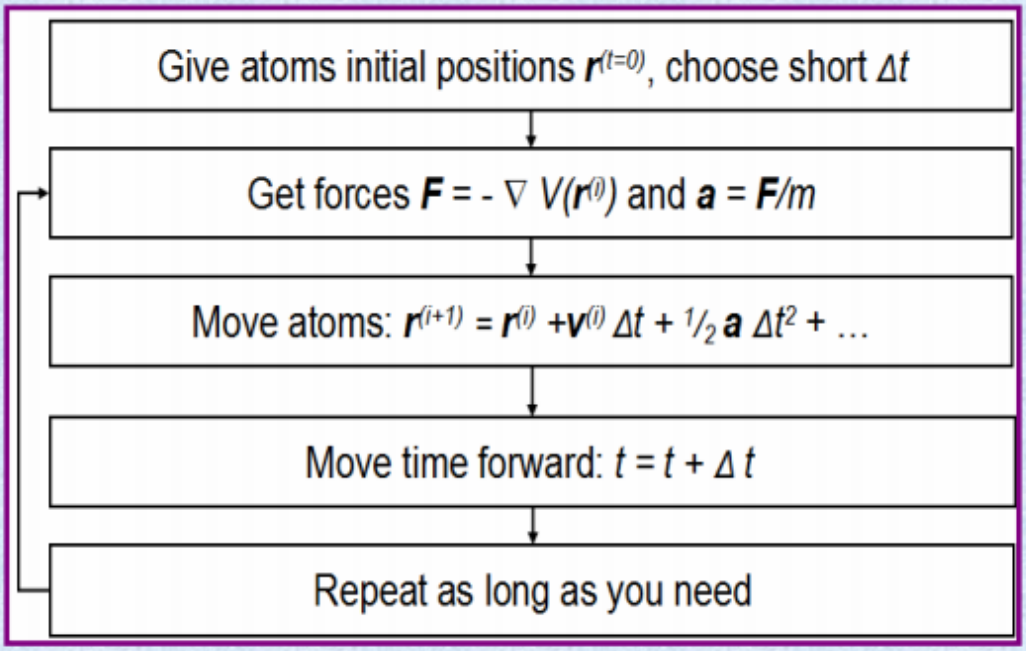
$$f = m \times a$$

The motion of each particle of the system is calculated from a

a is calculated from f

f is calculated from the potential V

A very simplified MD run



Molecular dynamics

- The potential ***V*** can be calculated at different accuracy level (from MM to QM)
- In biology the potential ***V*** is generally obtained by a MM force field
- This is a classical treatment allowing the calculation of conformational changes but usually it is not able to reproduce chemical reactions

Δt cannot be longer than the fastest atomic motion, therefore:

$$\Delta t = 10^{-15}$$

consequently a simulation of a microsecond needs one billion steps

Molecular dynamics

Temperature is directly correlated with kynetic energy:

$$K = \frac{3}{2} N k_B T$$

Generally a "free" evolution of the system is not allowed. Constraints on temperature and/or pressure are imposed in order to reproduce a particular ensemble.

Molecular dynamics

Environment simulation

The solvent can be simulated in an implicit and in an explicit manner.

Implicit solvent (in most cases the *continuum* approximation is used): fast calculation but poor results

Explicit solvent (*periodic boundary conditions* are generally used): accurate results but time consuming

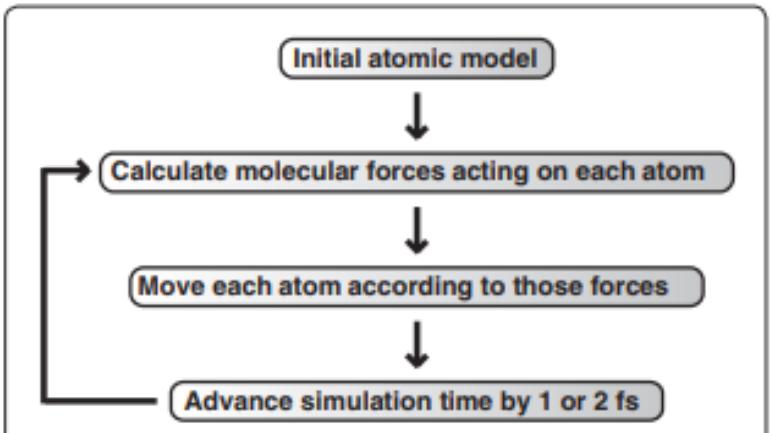


Figure 2. A schematic showing how a molecular dynamics simulation is performed. First, a computer model of the receptor-ligand system is prepared. An equation like that shown in Figure 3 is used to estimate the forces acting on each of the system atoms. The positions of the atoms are moved according to Newton's laws of motion. The simulation time is advanced, and the process is repeated many times. This figure was adapted from a version originally created by Kai Nordlund.

First, a computer model of the molecular system is prepared from nuclear magnetic resonance (NMR), crystallographic, or homology-modeling data. The forces acting on each of the system atoms are then estimated from an equation like that shown in Figure 3 [14]. In brief, forces arising from interactions between bonded and non-bonded atoms contribute. Chemical bonds and

atomic angles are modeled using simple virtual springs, and dihedral angles (that is, rotations about a bond) are modeled using a sinusoidal function that approximates the energy differences between eclipsed and staggered conformations. Non-bonded forces arise due to van der Waals interactions, modeled using the Lennard-Jones 6- 12 potential, and charged (electrostatic) interactions, modeled using Coulomb's law.

Once the forces acting on each of the system atoms have been calculated, the positions of these atoms are moved according to Newton's laws of motion. The simulation time is then advanced, often by only 1 or 2 quadrillionths of a second, and the process is repeated, typically millions of times. Because so many calculations are required, molecular dynamics simulations are typically performed on computer clusters or supercomputers using dozens if not hundreds of processors in parallel. Many of the most popular simulation software packages, which often bear the same names as their default force fields (for example AMBER, CHARMM, and NAMD), are compatible with the Message Passing Interface (MPI), a system of computerto-computer messaging that greatly facilitates the execution of complex tasks by one software application on multiple processors operating simultaneously.

SATHYABAMA
INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)
Accredited "A" Grade by NAAC | 12B Status by UGC | Approved by AICTE
www.sathyabama.ac.in

SCHOOL OF BIO AND CHEMICAL ENGINEERING
DEPARTMENT OF BIOINFORMATICS

UNIT – 4- SBIA5203 – Biomolecular modeling

Analyses in MD Simulation

The common output from MD simulations includes positions, velocities, potential energies. Some other useful information can also be analyzed from the trajectory file.

1. Conformational analysis: analyzing conformational changes of proteins (stability, folding or unfolding), nucleic acids or polymers in different solutions or temperatures. 16
2. Hydrogen bonds, coordination bonds analysis: analyzing number and occupancy of hydrogen bonds or coordination bonds of selected groups.
3. Chemical shift analysis: predicting the chemical shift of each atom in a molecule in nuclear magnetic resonance (NMR) spectroscopy.
4. pKa value analysis: predicting protonation states of residues on proteins or polymers in aqueous solution at various pH values.
5. Protein-ligand docking: searching the binding site for ligands on the surface of a protein based on geometric complementary and scoring functions for drug design or protein purification.
6. Interaction energy analysis: calculating VdW and electrostatic interaction energies between two selected groups.
7. Water dynamics analysis: calculating residence time, the self-diffusion coefficient, or molecular orientations in selected regions.
8. Free energy analysis: calculating relative free energies between different states such as solvation free energy and binding free energy.
9. Mechanistic analysis: constructing the free energy surface with defined reaction coordinates to investigate biological processes such as the ion channel, as well as the reaction mechanisms of the enzyme or catalysis

Free Energy Calculation

Based on the statistical mechanics, the relative free energy differences between two states can be calculated in MD simulation. The relative free energy is directly related to many chemical quantities such as the solubility or the binding strength. In this dissertation, thermodynamics integration (TI) and Molecular Mechanics Poisson-Boltzmann Surface Area (MM/PBSA) were conducted to calculate relative free energies.

Biological molecules exhibit a wide range of time scales over which specific processes occur; for example

- Local Motions (0.01 to 5 Å, 10^{-15} to 10^{-1} s)
  - Atomic fluctuations
  - Sidechain Motions
  - Loop Motions
- Rigid Body Motions (1 to 10Å, 10^{-9} to 1s)
  - Helix Motions
  - Domain Motions (hinge bending)
  - Subunit motions
- Large-Scale Motions (> 5Å, 10^{-7} to 10^{4} s)
  - Helix coil transitions
  - Dissociation/Association
  - Folding and Unfolding

Molecular dynamics simulations permit the study of complex, dynamic processes that occur in biological systems. These include, for example,

- Protein stability
- Conformational changes
- Protein folding
- Molecular recognition: proteins, DNA, membranes, complexes
- Ion transport in biological systems

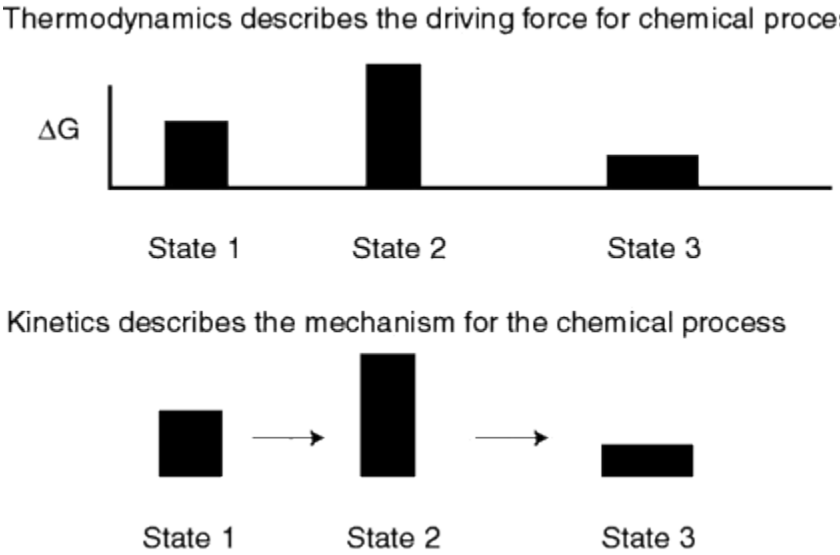
and provide the mean to carry out the following studies,

- Drug Design
- Structure determination: X-ray and NMR

The first protein simulations appeared in 1977 with the simulation of the bovine pancreatic trypsin inhibitor (BPTI) (McCammon, *et al*, 1977). Today in the literature, one routinely finds molecular dynamics simulations of solvated proteins, protein-DNA complexes as well as lipid systems addressing a variety of issues including the thermodynamics of ligand binding and the folding of small proteins.

In a molecular dynamics simulation, one often wishes to explore the macroscopic properties of a system through microscopic simulations, for example, to calculate changes in the binding free energy of a particular drug candidate, or to examine the energetics and mechanisms of conformational change. The connection between microscopic simulations and macroscopic properties is made via statistical mechanics which provides the rigorous mathematical expressions that relate macroscopic properties to the distribution and motion of the atoms and molecules of the N-body system; molecular dynamics simulations provide the means to solve the equation of motion of the particles and evaluate these mathematical formulas. With molecular

dynamics simulations, one can study both thermodynamic properties and/or time dependent (kinetic) phenomenon.



Statistical mechanics is the branch of physical sciences that studies macroscopic systems from a molecular point of view. The goal is to understand and to predict macroscopic phenomena from the properties of individual molecules making up the system. The system could range from a collection of solvent molecules to a solvated protein-DNA complex. In order to connect the macroscopic system to the microscopic system, time independent statistical averages are often introduced.

Definitions

The thermodynamic state of a system is usually defined by a small set of parameters, for example, the temperature, T, the pressure, P, and the number of particles, N. Other thermodynamic properties may be derived from the equations of state and other fundamental thermodynamic equations.

The mechanical or microscopic state of a system is defined by the atomic positions, q, and momenta, p; these can also be considered as coordinates in a multidimensional space called phase space. For a system of N particles, this space has 6N dimensions. A single point in phase space, denoted by G, describes the state of the system. An ensemble is a collection of points in phase space satisfying the conditions of a particular thermodynamic state. A molecular dynamics simulations generates a sequence of points in phase space as a function of time; these points belong to the same ensemble, and they correspond to the different conformations of the system and their respective momenta. Several different ensembles are described below.

An ensemble is a collection of all possible systems which have different microscopic states but have an identical macroscopic or thermodynamic state.

There exist different ensembles with different characteristics.

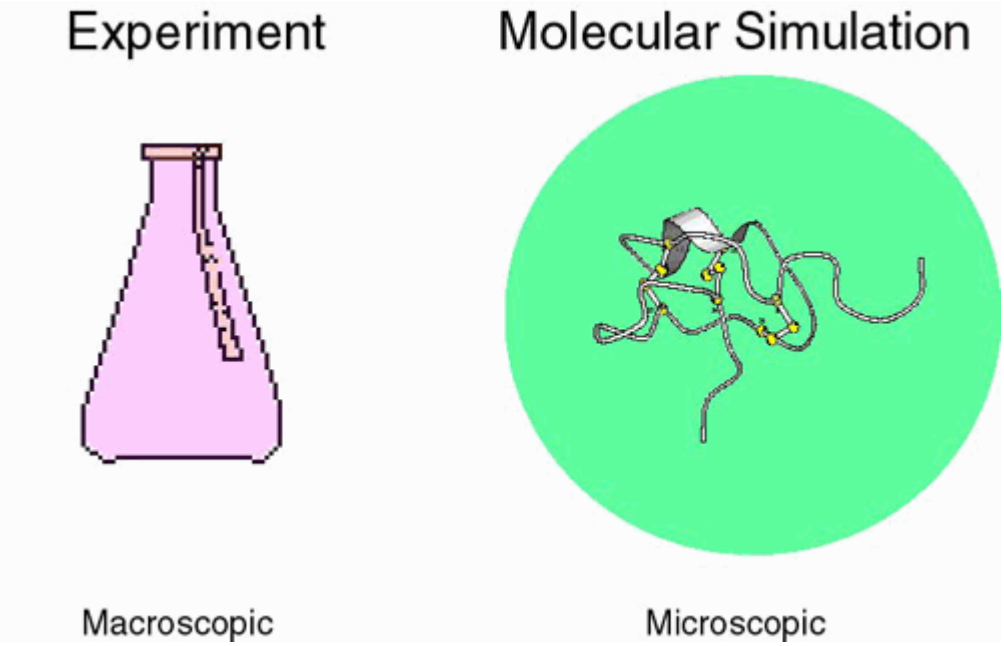
Microcanonical ensemble (NVE) : The thermodynamic state characterized by a fixed number of atoms, N, a fixed volume, V, and a fixed energy, E. This corresponds to an isolated system.

Canonical Ensemble (NVT): This is a collection of all systems whose thermodynamic state is characterized by a fixed number of atoms, N, a fixed volume, V, and a fixed temperature, T.

Isobaric-Isothermal Ensemble (NPT): This ensemble is characterized by a fixed number of atoms, N, a fixed pressure, P, and a fixed temperature, T.
Grand canonical Ensemble (mVT): The thermodynamic state for this ensemble is characterized by a fixed chemical potential, m, a fixed volume, V, and a fixed temperature, T.

Calculating Averages from a Molecular Dynamics Simulation

An experiment is usually made on a macroscopic sample that contains an extremely large number of atoms or molecules sampling an enormous number of conformations. In statistical mechanics, averages corresponding to experimental observables are defined in terms of ensemble averages; one justification for this is that there has been good agreement with experiment. An ensemble average is average taken over a large number of replicas of the system considered simultaneously.



6.1.1 Time Averages, Ensemble Averages and Some Historical Background

Suppose we wish to determine experimentally the value of a property of a system such as the pressure or the heat capacity. In general, such properties will depend upon the positions and momenta of the N particles that comprise the system. The instantaneous value of the property A can thus be written as $A(\mathbf{p}^N(t)\ \mathbf{r}^N(t))$, where $\mathbf{p}^N(t)$ and $\mathbf{r}^N(t)$ represent the N momenta and positions respectively at time t (i.e. $A(\mathbf{p}^N(t),\mathbf{r}^N(t)) \equiv A(p_{1x}, p_{1y}, p_{1z}, p_{2x}, \ldots, x_1, y_1, z_1, x_2, \ldots, t)$ where p_{1x} is the momentum of particle 1 in the x direction and x_1 is its x coordinate). Over time, the instantaneous value of the property A fluctuates as a result of interactions between the particles. The value that we measure experimentally is an average of A over the time of the measurement and is therefore known as a *time average*. As the time over which the measurement is made increases to infinity, so the value of the following integral approaches the 'true' average value of the property:

$$A_{ave} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A(\mathbf{p}^N(t), \mathbf{r}^N(t))\, dt \qquad (6.1)$$

To calculate average values of the properties of the system, it would therefore appear to be necessary to simulate the dynamic behaviour of the system (i.e. to determine values of $A(\mathbf{p}^N(t), \mathbf{r}^N(t))$, based upon a model of the intra- and intermolecular interactions present).

In statistical mechanics, average values are defined as ensemble averages.

The ensemble average is given by

$$\langle A \rangle_{ensemble} = \iint dp^N dr^N A\left(p^N, r^N\right)\rho\left(p^N, r^N\right)$$

where

$$A\left(p^N, r^N\right)$$

is the observable of interest and it is expressed as a function of the momenta, *p*, and the positions, *r*, of the system. The integration is over all possible variables of *r* and *p*.

The probability density of the ensemble is given by

$$\rho\left(p^N, r^N\right) = \frac{1}{Q}\exp\left[-H\left(p^N, r^N\right) / k_B T\right]$$

where *H* is the Hamiltonian, *T* is the temperature, *kB* is Boltzmann's constant and *Q* is the partition function

$$Q = \iint dp^N dr^N \exp\left[-H\left(p^N, r^N\right) / k_B T\right]$$

This integral is generally *extremely* difficult to calculate because one must calculate all possible states of the system. In a molecular dynamics simulation, the points in the ensemble are calculated sequentially in time, so to calculate an ensemble average, the molecular dynamics simulations must pass through all possible states corresponding to the particular thermodynamic constraints.

Another way, as done in an MD simulation, is to determine a time average of A, which is expressed as

$$\langle A \rangle_{time} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A\left(p^N(t), r^N(t)\right) dt \approx \frac{1}{M} \sum_{t=1}^{M} A\left(p^N, r^N\right)$$

where t is the simulation time, M is the number of time steps in the simulation and *A(pN,rN)* is the instantaneous value of *A*.

To calculate average values of the properties of the system, it would therefore appear to be necessary to simulate the dynamic behaviour of the system (i.e. to determine values of $A(\mathbf{p}^N(t), \mathbf{r}^N(t))$, based upon a model of the intra- and intermolecular interactions present). In principle, this is relatively straightforward to do. For any arrangement of the atoms in the system, the force acting on each atom due to interactions with other atoms can be calculated by differentiating the energy function. From the force on each atom it is possible to determine its acceleration via Newton's second law. Integration of the equations of motion should then yield a trajectory that describes how the positions, velocities and accelerations of the particles vary with time, and from which the average values of properties can be determined using the numerical equivalent of Equation (6.1). The difficulty is that for 'macroscopic' numbers of atoms or molecules (of the order of 10^{23}) it is not even feasible to determine an initial configuration of the system, let alone integrate the equations of motion and calculate a trajectory. Recognising this problem, Boltzmann and Gibbs developed statistical mechanics, in which a single system evolving in time is replaced by a large number of replications of the system that are considered simultaneously. The time average is then replaced by an *ensemble average*:

$$\langle A \rangle = \iint d\mathbf{p}^N \, d\mathbf{r}^N \, A(\mathbf{p}^N, \mathbf{r}^N) \rho(\mathbf{p}^N, \mathbf{r}^N) \tag{6.2}$$

The angle brackets $\langle\,\rangle$ indicate an ensemble average, or *expectation value*; that is, the average value of the property A over all replications of the ensemble generated by the simulation. Equation (6.2) is written as a double integral for convenience but in fact there should be $6N$ integral signs on the integral for the $6N$ positions and momenta of all the particles. $\rho(\mathbf{p}^N\mathbf{r}^N)$ is the *probability density* of the ensemble; that is, the probability of finding a configuration with momenta $\mathbf{p}^N$ and positions $\mathbf{r}^N$. The ensemble average of the property A is then determined by integrating over all possible configurations of the system. In accordance with the *ergodic hypothesis*, which is one of the fundamental axioms of statistical mechanics, the ensem-

uous nature of the more realistic potentials requires the equations of motion to be integrated by breaking the calculation into a series of very short time steps (typically between 1 femtosecond and 10 femtoseconds; 10^{-15} s to 10^{-14} s). At each step, the forces on the atoms are computed and combined with the current positions and velocities to generate new positions and velocities a short time ahead. The force acting on each atom is assumed to be constant during the time interval. The atoms are then moved to the new positions, an updated set of forces is computed, and so on. In this way a molecular dynamics simulation generates a

trajectory that describes how the dynamic variables change with time. Molecular dynamics simulations are typically run for tens or hundreds of picoseconds (a 100 ps simulation using a 1 fs time step requires 100 000 steps). Thermodynamic averages are obtained from molecular dynamics as time averages using numerical integration of Equation (6.2):

$$\langle A \rangle = \frac{1}{M} \sum_{i=1}^{M} A(\mathbf{p}^N, \mathbf{r}^N) \qquad (6.5)$$

M is the number of time steps. Molecular dynamics is also extensively used to investigate the conformational properties of flexible molecules as will be discussed in Chapters 7 and 9.

The dilemma appears to be that one can calculate time averages by molecular dynamics simulation, but the experimental observables are assumed to be ensemble averages. Resolving this leads us to one of the most fundamental axioms of statistical mechanics, the **ergodic hypothesis,** which states that the time average equals the ensemble average.

The **Ergodic hypothesis** states

$$\langle A \rangle_{ensemble} = \langle A \rangle_{time}$$

Ensemble average = Time average

The basic idea is that if one allows the system to evolve in time indefinitely, that system will eventually pass through all possible states. One goal, therefore, of a molecular dynamics simulation is to generate enough representative conformations such that this equality is satisfied. If this is the case, experimentally relevant information concerning structural, dynamic and thermodynamic properties may then be calculated using a feasible amount of computer resources. Because the simulations are of fixed duration, one must be certain to sample a sufficient amount of phase space.

6.1.3 The Basic Elements of the Monte Carlo Method

In a molecular dynamics simulation the successive configurations of the system are connected in time. This is not the case in a Monte Carlo simulation, where each configuration depends only upon its predecessor and not upon any other of the configurations previously visited. The Monte Carlo method generates configurations randomly and uses a special set of criteria to decide whether or not to accept each new configuration. These criteria ensure that the probability of obtaining a given configuration is equal to its Boltzmann factor, $\exp\{-\mathscr{V}(\mathbf{r}^N)/k_B T\}$, where $\mathscr{V}(\mathbf{r}^N)$ is calculated using the potential energy function. States with a low energy are thus generated with a higher probability than configurations with a higher energy. For each configuration that is accepted the values of the desired properties are calculated and at the end of the calculation the average of these properties is obtained by simply averaging over the number of values calculated, M:

$$\langle A \rangle = \frac{1}{M} \sum_{i=1}^{M} A(\mathbf{r}^N) \tag{6.6}$$

Most Monte Carlo simulations of molecular systems are more properly referred to as Metropolis Monte Carlo calculations after Metropolis and his colleagues, who reported the first such calculation. The distinction can be important because there are other ways in which an ensemble of configurations can be generated. As we shall see in Chapter 7, the Metropolis scheme is only one of a number of possibilities, though it is by far the most popular.

In a Monte Carlo simulation each new configuration of the system may be generated by randomly moving a single atom or molecule. In some cases new configurations may also be obtained by moving several atoms or molecules or by rotating about one or more bonds. The energy of the new configuration is then calculated using the potential energy

In a Monte Carlo simulation each new configuration of the system may be generated by randomly moving a single atom or molecule. In some cases new configurations may also be obtained by moving several atoms or molecules or by rotating about one or more bonds. The energy of the new configuration is then calculated using the potential energy function. If the energy of the new configuration is lower than the energy of its predecessor then the new configuration is accepted. If the energy of the new configuration is higher than the energy of its predecessor then the *Boltzmann factor* of the energy difference is calculated: $\exp[-(\mathscr{V}_{\text{new}}(\mathbf{r}^N) - \mathscr{V}_{\text{old}}(\mathbf{r}^N))/k_B T]$. A random number is then generated between 0 and 1 and compared with this Boltzmann factor. If the random number is higher than the Boltzmann factor then the move is rejected and the original configuration is retained for the next iteration; if the random number is lower then the move is accepted and the new

configuration becomes the next state. This procedure has the effect of permitting moves to states of higher energy. The smaller the uphill move (i.e. the smaller the value of $\mathscr{V}_{\text{new}}(\mathbf{r}^N) - \mathscr{V}_{\text{old}}(\mathbf{r}^N)$) the greater is the probability that the move will be accepted.

UNIT – 5- SBIA5203 – Biomolecular modeling

Protein structure prediction

Determining Protein Structures

- Protein structures can be determined experimentally (in most cases) by
  - x-ray crystallography
  - nuclear magnetic resonance (NMR)
  - cryo-electron microscopy (cryo-EM)
- But this is very expensive and time-consuming
- There is a large sequence-structure gap
  - ≈ 550K protein sequences in SwissProt database
  - ≈ 100K protein structures in PDB database
- Key question: can we predict structures by computational means instead?

Types of Protein Structure Predictions

- Prediction in 1D
  - secondary structure
  - solvent accessibility (which residues are exposed to water, which are buried)
  - transmembrane helices (which residues span membranes)
- Prediction in 2D
  - inter-residue/strand contacts
- Prediction in 3D
  - homology modeling
  - fold recognition (e.g. via threading)
  - *ab initio* prediction (e.g. via molecular dynamics)

Performance of Structure Prediction Methods

There are four major classes of algorithms for the prediction of proteins structure.

Homology based methods work from the assumption that proteins with shared ancestry will have mutually conserved sequence and structure. The objective is to identify homologous proteins with known structures and to use these similar structures to predict the structure for an unknown protein. This is done using template elements analogous to building blocks, such as Legos. Correspondence information is derived from primary structure (i.e. sequence) similarity. Accordingly, homology based algorithms are the most reliant of the four classes, generally requiring a database of known proteins with at least 30% sequence similarity and that provide at least 90% template coverage. Since homology algorithms work closely with similar, known natural structures, they also tend to have the best accuracy, producing predictions with an RMSD on the order of 1 - 3Å. They also tend to be the fastest and easiest to implement with running times on the order of seconds.

The second class of algorithms does not use sequence similarity to establish and instead attempts to establish correspondence between the unknown and the known structures by recognising commonalities in the folds. While this relaxes the algorithms' needs in terms of sequence identity (generally 20 - 30% is the minimum) and template coverage (reduced to 75%), this is paid for with losses in accuracy and computation time. Accuracy degrades is a general range of 2 - 5Å and computation time is on the order of minutes instead of seconds.

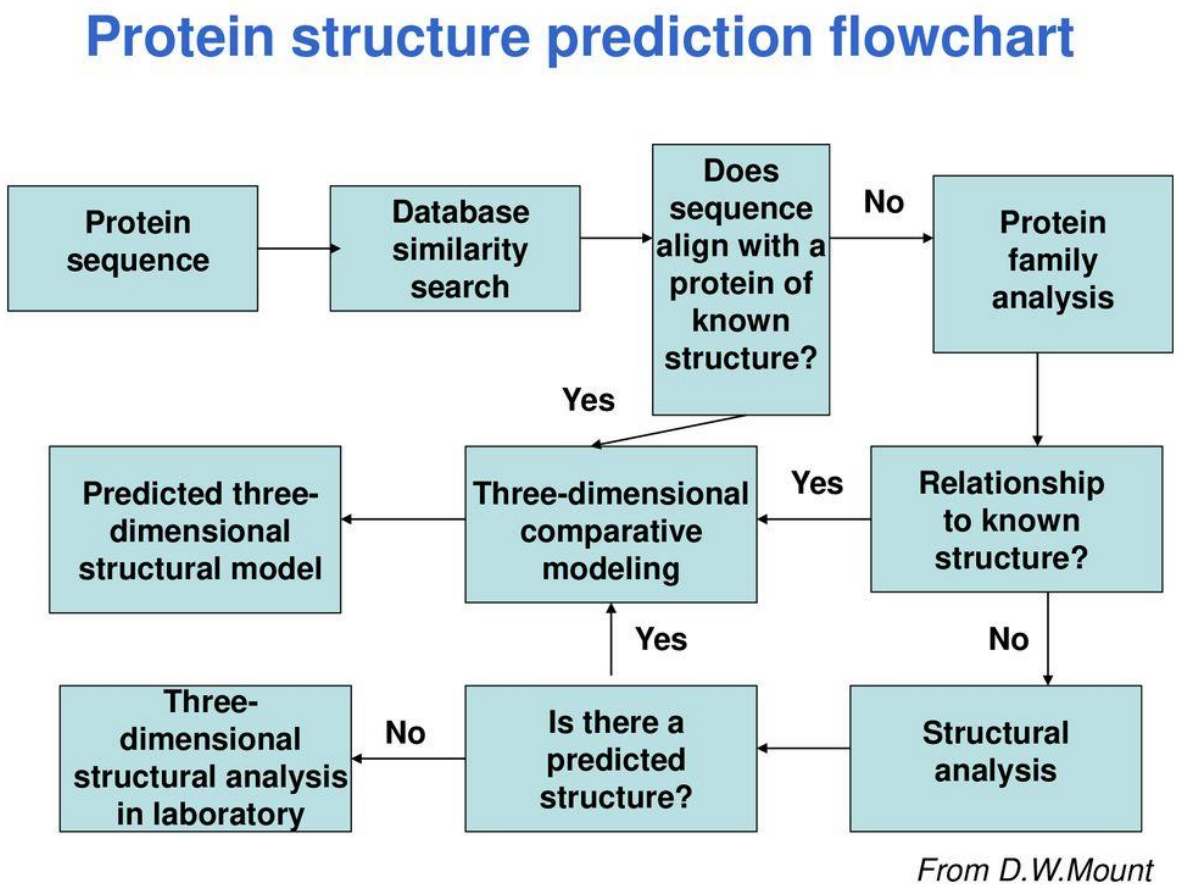
The third class of algorithms completely foregoes homology-based comparison and instead works in terms of fold recognition alone. Again relaxations in the sequence similarity and template coverage requirements, which fall to less than 20% and greater than 50% respectively, are offset by worse accuracy and increased computation time. Accuracy is generally between 3 - 10Å after computation requiring hours for these methods.

Ab initio methods comprise the last class of algorithms. While the previous three classes all use no templates, *ab initio* algorithms can predict structures with an accuracy in the range of 5 - 20Å after computation lasting on the order of days.

Method	Sequence Similarity	Template Coverage	Accuracy	Difficulty	Comput. Expense
Homology	>30%	>90%	1-3 Å	Trivial	Seconds
FR/Homology	20-30%	>75%	2-5 Å	Easy	Minutes
Fold Recognition	<20%	>50%	3-10 Å	Moderate	Hours
Ab initio	<10%	0	5-20 Å	Hard	Days

Figure 1: Summary of Requirements and Performance

some manner Informatic about how the conformations of known proteins, pure ab initio techniques explore structure space with the guidance of an energy function that they seek to minimise without regard to known structures. These techniques are useful to predict novel structures for which comparisons against known structures do not yield useful information. In addition, as ab initio techniques improve, they may yield insight into the ways proteins fold in nature. Since these techniques start with less information, their performance is not as good. On the other hand, they can be applied in certain contexts where other techniques' information requirements are not satisfied. With sequence similarity below 10% (or zero for pure ab initio) and



Homology modeling

- Homology modeling is also known as comparative modeling predicts protein structures based on sequence homology with known structures.
- It is based on the principle that "if two proteins share a high enough sequence similarity, they are likely to have very similar three-dimensional structures."
- It hence relies on the identification of one or more known protein structures likely to resemble the structure of the query sequence, and on the production of an alignment that maps residues in the query sequence to residues in the template sequence.
- Thus, if one of the protein sequences has a known structure, then the structure can be copied to the unknown protein with a high level of confidence.

It predicts the three-dimensional structure of a given protein sequence (target) based on alignment to one or more known protein structures (templates). If the similarity between the target sequence and the template sequence is detected, structural similarity can be assumed. In general, 30% sequence identity is required to generate a useful model.

Steps

The overall homology modeling procedure consists of six steps.

1. The first step is template selection, which involves the **identification of homologous sequences** in the protein structure database to be used as templates for modeling.

- It most commonly relies on serial pairwise sequence alignments aided by database search techniques such as FASTA and BLAST but may employ other approaches such as PSI-BLAST, Protein threading etc in addition to these.

2. The second step is the **alignment of the target and template sequences.**

- Once the structure with the highest sequence similarity is identified as a template, the full-length sequences of the template and target proteins need to be realigned using refined alignment algorithms to obtain optimal alignment.
- The best possible multiple alignment algorithms, such as Praline and T-coffee, should be used for this purpose followed by manual refinement of the alignment such as to improve alignment quality.

3. The third step is to **build a framework structure for the target protein** consisting of main chain atoms.

- Once optimal alignment is achieved, coordinates of the corresponding residues of the template proteins can be simply copied onto the target protein.
- If the two aligned residues are identical, coordinates of the side chain atoms are copied along with the main chain atoms. If the two residues differ, only the backbone atoms can be copied.

4. The fourth step of model building includes the **addition and optimization of side chain** atoms and loops.

- In the sequence alignment for modeling, there are often regions caused by insertions and deletions producing gaps in the sequence alignment. The gaps cannot be directly modeled, creating “holes” in the model.
- Closing the gaps requires loop modeling, which is a very difficult problem.
- Currently, there are two main techniques used to approach the problem: the database searching method and the ab initio method.
- Once main chain atoms are built, the positions of side chains that are not modeled must be determined.
- Modeling side chain geometry is very important in evaluating protein-ligand interactions at active sites and protein-protein interactions at the contact interface.
- Most modeling packages incorporate the side chain refinement function. A specialized side-chain modeling program that has reasonably good performance is SCWRL (sidechain placement with a rotamer library), a UNIX program that works by placing side chains on a backbone template according to preferences in the backbone-dependent rotamer library.

5. The fifth step is to **refine and optimize** the entire model according to energy criteria.

- The entire raw homology model is made free of structural irregularities such as unfavorable bond angles, bond lengths, or close atomic contacts.
- If structural irregularities are seen, it can be corrected by applying the energy minimization procedure on the entire model.

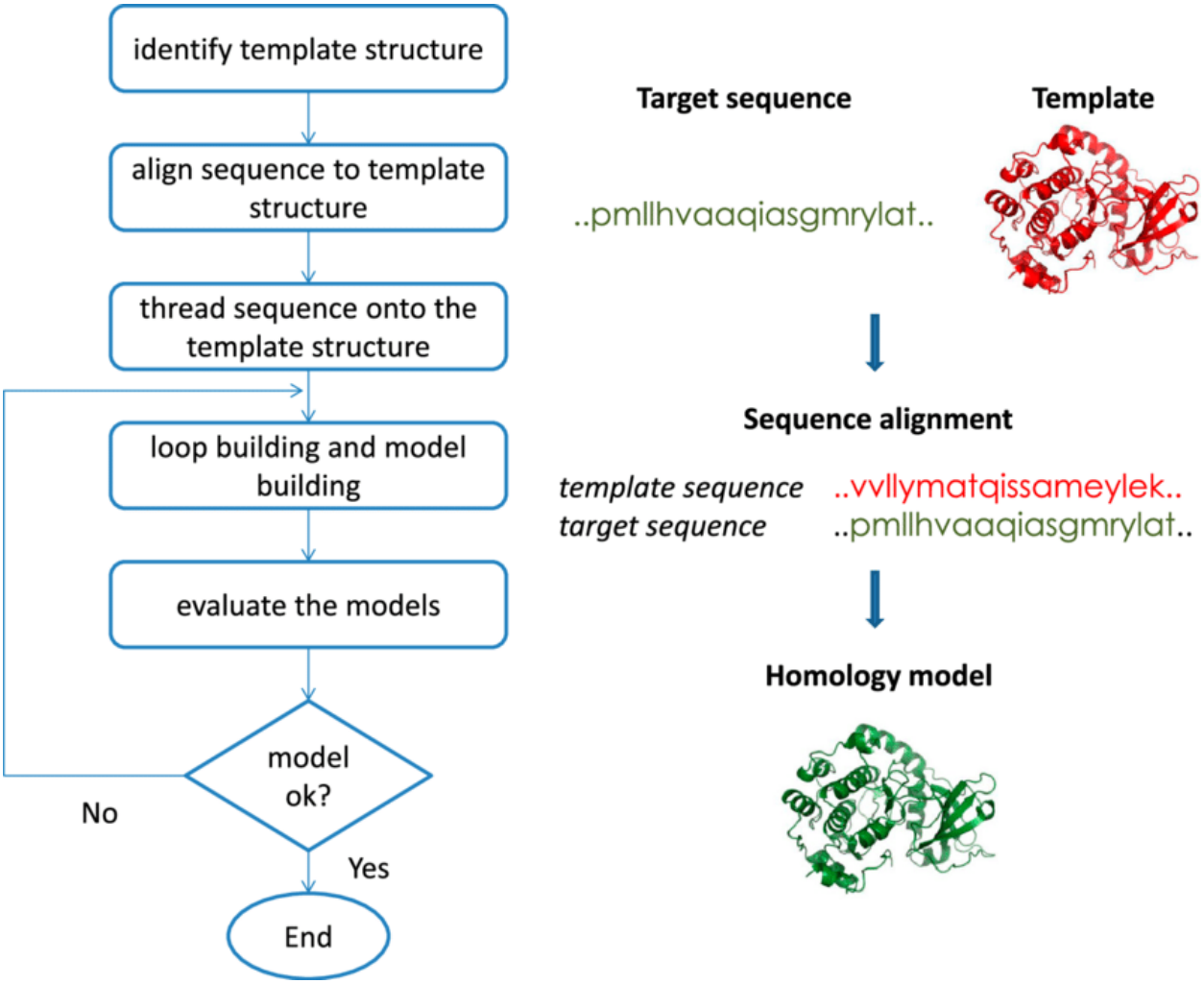
- Another often used structure refinement procedure is a molecular dynamics simulation. GROMOS (www.igc.ethz.ch/gromos/) is a UNIX program for molecular dynamics simulation.

6. The final step involves **evaluating the overall quality** of the model obtained.

- The final homology model has to be evaluated to make sure that the structural features of the model are consistent with the physicochemical rules.
- If necessary, alignment and model building are repeated until a satisfactory result is obtained.

Uses

- Protein modeling Provide a solid basis for:
- Rational design of proteins with increased stability or novel functions
- Analysis of protein function, interactions, antigenic behavior
- Structure-based drug design
- Because it is difficult and time-consuming to obtain experimental structures from methods such as X-ray crystallography and protein NMR for every protein of interest, homology modeling can provide useful structural models for generating hypotheses about a protein's function and directing further experimental work.



Threading

Protein threading, also known as **fold recognition**, is a method of protein modeling which is used to model those proteins which have the same fold as proteins of known structures, but do not have homologous proteins with known structure. It differs from the homology modeling method of structure prediction as it (protein threading) is used for proteins which do not have their homologous protein structures deposited in the Protein Data Bank (PDB), whereas homology modeling is used for those proteins which do. Threading works by using statistical knowledge of the relationship between the structures deposited in the PDB and the sequence of the protein which one wishes to model.

The prediction is made by "threading" (i.e. placing, aligning) each amino acid in the target sequence to a position in the template structure, and evaluating how well the target fits the template. After the best-fit template is selected, the structural model of the sequence is built based on the alignment with the chosen template. Protein threading is based on two basic observations: that the number of different folds in nature is fairly small (approximately 1300); and that 90% of the new structures submitted to the PDB in the past three years have similar structural folds to ones already in the PDB.

Method

A general paradigm of protein threading consists of the following four steps:

The construction of a structure template database: Select protein structures from the protein structure databases as structural templates. This generally involves selecting protein structures from databases such as PDB, FSSP, SCOP, or CATH, after removing protein structures with high sequence similarities.

The design of the scoring function: Design a good scoring function to measure the fitness between target sequences and templates based on the knowledge of the known relationships between the structures and the sequences. A good scoring function should contain mutation potential, environment fitness potential, pairwise potential, secondary structure compatibilities, and gap penalties. The quality of the energy function is closely related to the prediction accuracy, especially the alignment accuracy.

Threading alignment: Align the target sequence with each of the structure templates by optimizing the designed scoring function. This step is one of the major tasks of all threading-based structure prediction programs that take into account the pairwise contact potential; otherwise, a dynamic programming algorithm can fulfill it.

Threading prediction: Select the threading alignment that is statistically most probable as the threading prediction. Then construct a structure model for the target by placing the backbone atoms of the target sequence at their aligned backbone positions of the selected structural template.

Fold recognition methods

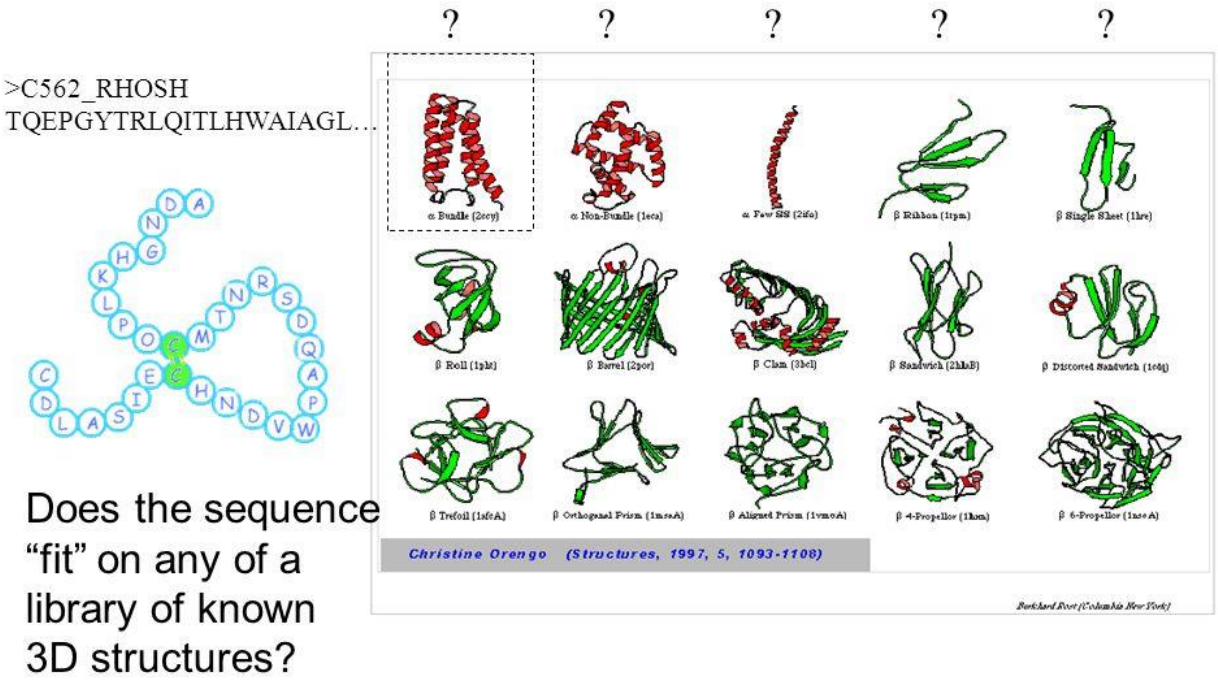
Fold recognition methods can be broadly divided into two types: *1*, those that derive a 1-D profile for each structure in the fold library and align the target sequence to these profiles; and *2*, those that consider the full 3-D structure of the protein template. A simple example of a profile

representation would be to take each amino acid in the structure and simply label it according to whether it is buried in the core of the protein or exposed on the surface. More elaborate profiles might take into account the local secondary structure (e.g. whether the amino acid is part of an alpha helix) or even evolutionary information (how conserved the amino acid is). In the 3-D representation, the structure is modeled as a set of inter-atomic distances, i.e. the distances are calculated between some or all of the atom pairs in the structure. This is a much richer and far more flexible description of the structure, but is much harder to use in calculating an alignment. The profile-based fold recognition approach was first described by Bowie, Lüthy and David Eisenberg in 1991.[1] The term *threading* was first coined by David Jones, William R. Taylor and Janet Thornton in 1992,[2] and originally referred specifically to the use of a full 3-D structure atomic representation of the protein template in fold recognition. Today, the terms threading and fold recognition are frequently (though somewhat incorrectly) used interchangeably.

Fold recognition methods are widely used and effective because it is believed that there are a strictly limited number of different protein folds in nature, mostly as a result of evolution but also due to constraints imposed by the basic physics and chemistry of polypeptide chains. There is, therefore, a good chance (currently 70-80%) that a protein which has a similar fold to the target protein has already been studied by X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy and can be found in the PDB. Currently there are nearly 1300 different protein folds known, but new folds are still being discovered every year due in significant part to the ongoing structural genomics projects.

Many different algorithms have been proposed for finding the correct threading of a sequence onto a structure, though many make use of dynamic programming in some form. For full 3-D threading, the problem of identifying the best alignment is very difficult (it is an NP-hard problem for some models of threading).[*citation needed*] Researchers have made use of many combinatorial optimization methods such as Conditional random fields, simulated annealing, branch and bound and linear programming, searching to arrive at heuristic solutions. It is interesting to compare threading methods to methods which attempt to align two protein structures (protein structural alignment), and indeed many of the same algorithms have been applied to both problems.

Fold Recognition ('Threading')



Abinitio prediction

The *ab-initio* method is often preferred for structure prediction when there is no or very low amount of similarity for the protein (let's say query protein sequence). It is the most difficult and general approach where the query protein is folded with a random conformation. The *ab-initio* method is based on the thermodynamic hypothesis proposed by Anfinsen [4], according to which the native structure corresponds to the global free energy minimum under a given set of conditions.

There are several ab-initio structure prediction approaches available such as ROSETTA, TOUCHSTONE-II, and the most widely preferred I-Tasser . These approaches are based on the Monte-Carlo algorithm. It has been found that I-Tasser outperforms the ROSETTA and TOUCHSTONE-II approaches with a far lower CPU cost .

In ab initio methods, an initial effort to elucidate secondary structures (alpha helix, beta sheet, beta turn, etc.) from primary structure is made by utilization of physicochemical parameters and neural net algorithms. From that point, algorithms predict tertiary folding. One drawback to this strategy is that it is not yet capable of incorporating the locations and orientation of amino acid side chains.

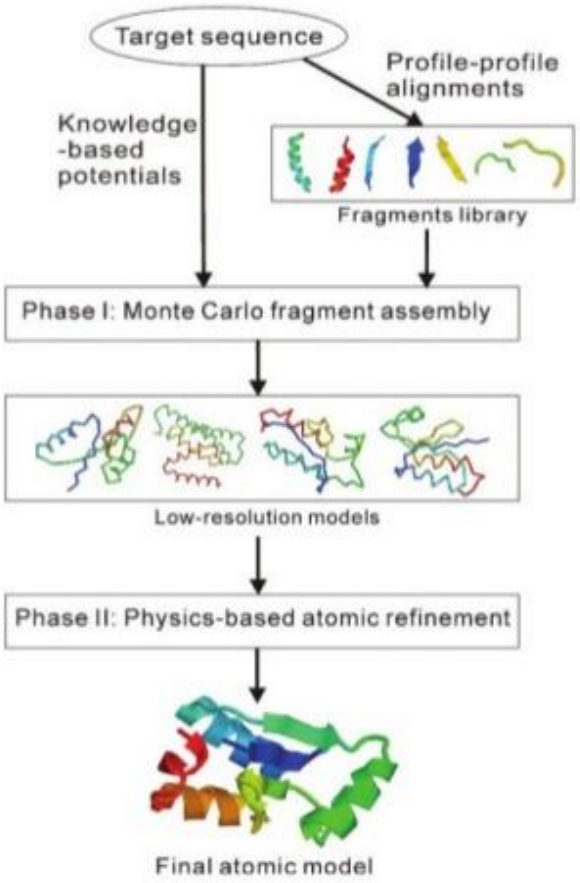


Fig.: Flowchart of the ROSETTA protocol

- A successful *ab initio* modelling depends on three factors:
  - an **accurate energy function** with which the native structure of a protein corresponds to the most thermodynamically stable state, compared to all possible decoy structures;
  - an efficient search method which can quickly identify the low-energy states through **conformational search**;
  - **selection of native-like models** from a pool of decoy structures.

Energy Functions ● Energy classified into two groups: ➔ Physics-based energy functions ➔ Knowledge-based energy functions

Physics-Based Energy Functions "In a strictly-defined physics-based ab initio method, interactions between atoms should be based on quantum mechanics and the coulomb potential

with only a few fundamental parameters such as the electron charge and the Planck constant; all atoms should be described by their atom types where only the number of electrons is relevant.

A compromised force field with a large number of selected atom types is used. In each atom type, the chemical and physical properties of the atoms are enough alike with the parameters calculated from crystal packing or quantum mechanical theory

Knowledge-Based Energy Function ● Refers to the empirical energy terms derived from the statistics of the solved structures in deposited PDB. ● Can be divided into two types: ➢ generic and sequence-independent terms such as the hydrogen bonding and the local backbone stiffness of a polypeptide chain ➢ amino-acid or protein-sequence dependent terms, e.g. pair wise residue contact potential, distance dependent atomic contact potential , and secondary structure propensities

Conformational Search Methods ● Successful ab initio modelling of protein structures depends on the availability of a powerful conformation search method which can efficiently find the global minimum energy structure for a given energy function with complicated energy landscape. ● Types: ➔ Monte Carlo Simulations ➔ Molecular Dynamics ➔ Genetic Algorithm ➔ Mathematical Optimization

Sequence-Structure Compatibility Function ● Best models are selected not purely based on energy functions. ● They are selected based on the compatibility of target sequences to model structures. ● The earliest and still successful example is that by Luthy et al. (1992), who used threading scores to evaluate structures. ● Colovos and Yeates (1993) later used a quadratic error function to describe the non-covalently bonded interactions among CC, CN, CO, NN, NO and OO, where near-native structures have fewer errors than other decoys

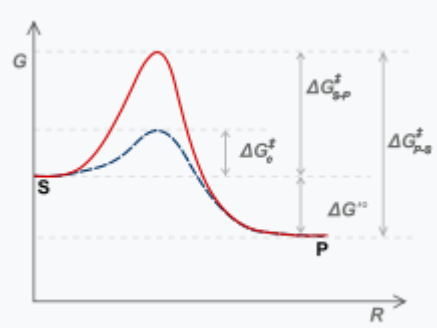
Clustering of Decoy Structures ● Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). ● The cluster-centre conformation of the largest cluster is considered closer to native structures than the majority of decoys. ● In the work by Shortle et al. (1998), for all 12 cases tested, the cluster-centre conformation of the largest cluster was closer to native structures than the majority of decoys. Cluster-centre structures were ranked as the top 1–5% closest to their native structures

Binding site

In biochemistry and molecular biology, a **binding site** is a region on a macromolecule such as a protein that binds to another molecule with specificity.[1] The binding partner of the macromolecule is often referred to as a ligand.[2] Ligands may include other proteins (resulting in a protein-protein interaction),[3] enzyme substrates,[4] second messengers, hormones, or allosteric modulators.[5] The binding event is often, but not always, accompanied by a conformational change that alters the protein's function.[6] Binding to protein binding sites is most often reversible (transient and non-covalent), but can also be covalent reversible[7] or irreversible.

Binding of a ligand to a binding site on protein often triggers a change in conformation in the protein and results in altered cellular function. Hence binding site on protein are critical parts of signal transduction pathways.[9] Types of ligands include neurotransmitters, toxins, neuropeptides, and steroid hormones.[10] Binding sites incur functional changes in a number of contexts, including enzyme catalysis, molecular pathway signaling, homeostatic regulation, and physiological function. Electric charge, steric shape and geometry of the site selectively allow for highly specific ligands to bind, activating a particular cascade of cellular interactions the protein is responsible for

Catalysis



Activation energy is decreased in the presence of an enzyme to catalyze the reaction.

Enzymes incur catalysis by binding more strongly to transition states than substrates and products. At the catalytic binding site, several different interactions may act upon the substrate. These range from electric catalysis, acid and base catalysis, covalent catalysis, and metal ion catalysis.[10] These interactions decrease the activation energy of a chemical reaction by providing favorable interactions to stabilize the high energy molecule. Enzyme binding allows for closer proximity and exclusion of substances irrelevant to the reaction. Side reactions are also discouraged by this specific binding.[13][10]

Types of enzymes that can perform these actions include oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases.[14]

For instance, the transferase hexokinase catalyzes the phosphorylation of glucose to make glucose-6-phosphate. Active site residues of hexokinase allow for stabilization of the glucose

molecule in the active site and spur the onset of an alternative pathway of favorable interactions, decreasing the activation energy

Inhibition

Protein inhibition by inhibitor binding may induce obstruction in pathway regulation, homeostatic regulation and physiological function.

Competitive inhibitors compete with substrate to bind to free enzymes at active sites and thus impede the production of the enzyme-substrate complex upon binding. For example, carbon monoxide poisoning is caused by the competitive binding of carbon monoxide as opposed to oxygen in hemoglobin.

Uncompetitive inhibitors, alternatively, bind concurrently with substrate at active sites. Upon binding to an enzyme substrate (ES) complex, an enzyme substrate inhibitor (ESI) complex is formed. Similar to competitive inhibitors, the rate at product formation is decreased also.[4]

Lastly, mixed inhibitors are able to bind to both the free enzyme and the enzyme-substrate complex. However, in contrast to competitive and uncompetitive inhibitors, mixed inhibitors bind to the allosteric site. Allosteric binding induces conformational changes that may increase the protein's affinity for substrate. This phenomenon is called positive modulation. Conversely, allosteric binding that decreases the protein's affinity for substrate is negative modulation.

Types

Active site

At the active site, a substrate binds to an enzyme to induce a chemical reaction.[17][18] Substrates, transition states, and products can bind to the active site, as well as any competitive inhibitors.[17] For example, in the context of protein function, the binding of calcium to troponin in muscle cells can induce a conformational change in troponin. This allows for tropomyosin to expose the actin-myosin binding site to which the myosin head binds to form a cross-bridge and induce a muscle contraction.[19]

In the context of the blood, an example of competitive binding is carbon monoxide which competes with oxygen for the active site on heme. Carbon monoxide's high affinity may outcompete oxygen in the presence of low oxygen concentration. In these circumstances, the binding of carbon monoxide induces a conformation change that discourages heme from binding to oxygen, resulting in carbon monoxide poisoning

Allosteric site

At the regulatory site, the binding of a ligand may elicit amplified or inhibited protein function.[4][20] The binding of a ligand to an allosteric site of a multimeric enzyme often induces positive cooperativity, that is the binding of one substrate induces a favorable conformation change and increases the enzyme's likelihood to bind to a second substrate.[21] Regulatory site ligands can involve homotropic and heterotropic ligands, in which single or multiple types of molecule affects enzyme activity respectively.[22]

Enzymes that are highly regulated are often essential in metabolic pathways. For example, phosphofructokinase (PFK), which phosphorylates fructose in glycolysis, is largely regulated by ATP. Its regulation in glycolysis is imperative because it is the committing and rate

limiting step of the pathway. PFK also controls the amount of glucose designated to form ATP through the catabolic pathway. Therefore, at sufficient levels of ATP, PFK is allosterically inhibited by ATP. This regulation efficiently conserves glucose reserves, which may be needed for other pathways. Citrate, an intermediate of the citric acid cycle, also works as an allosteric regulator of PFK

Single- and multi-chain binding sites

Binding sites can be characterized also by their structural features. Single-chain sites (of "monodesmic" ligands, μόνος: single, δεσμός: binding) are formed by a single protein chain, while multi-chain sites (of "polydesmic" ligands, πολοί: many) [24] are frequent in protein complexes, and are formed by ligands that bind more than one protein chain, typically in or near protein interfaces. Recent research shows that binding site structure has profound consequences for the biology of protein complexes (evolution of function, allostery).[25][26]

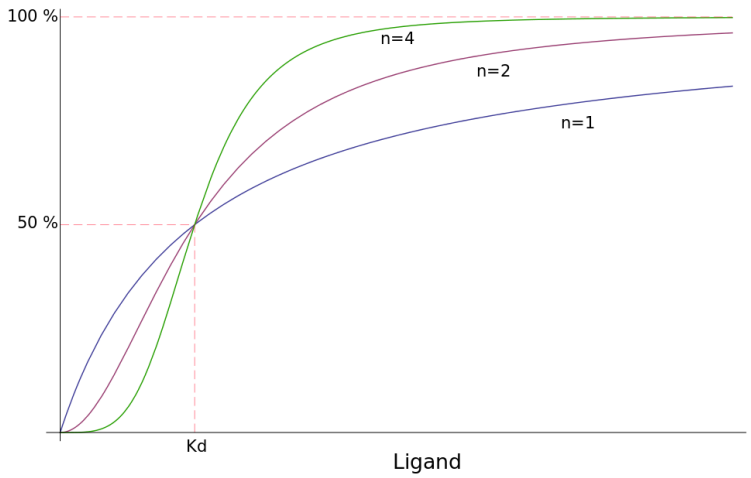
Cryptic binding sites

Cryptic binding sites are the binding sites that are transiently formed in an apo form or that are induced by ligand binding. Considering the cryptic binding sites increases the size of the potentially "druggable" human proteome from ~40% to ~78% of disease-associated proteins.[27] The binding sites have been investigated by: support vector machine applied to "CryptoSite" data set,[27] Extension of "CryptoSite" data set,[28] long timescale molecular dynamics simulation with Markov state model and with biophysical experiments,[29] and cryptic-site index that is based on relative accessible surface area

Binding curves

Binding curves describe the binding behavior of ligand to a protein. Curves can be characterized by their shape, sigmoidal or hyperbolic, which reflect whether or not the protein exhibits cooperative or noncooperative binding behavior respectively.[31] Typically, the x-axis describes the concentration of ligand and the y-axis describes the fractional saturation of ligands bound to all available binding sites.[4] The Michaelis Menten equation is usually used when determining the shape of the curve. The Michaelis Menten equation is derived based on steady-state conditions and accounts for the enzyme reactions taking place in a solution. However, when the reaction takes place while the enzyme is bound to a substrate, the kinetics play out differently.[32]

Modeling with binding curves are useful when evaluating the binding affinities of oxygen to hemoglobin and myoglobin in the blood. Hemoglobin, which has four heme groups, exhibits cooperative binding. This means that the binding of oxygen to a heme group on hemoglobin induces a favorable conformation change that allows for increased binding favorability of oxygen for the next heme groups. In these circumstances, the binding curve of hemoglobin will be sigmoidal due to its increased binding favorability for oxygen. Since myoglobin has only one heme group, it exhibits noncooperative binding which is hyperbolic on a binding curve



Sigmoidal versus hyperbolic binding patterns demonstrate cooperative and noncooperative character of enzymes.

Prediction

A number of computational tools have been developed for the prediction of the location of binding sites on proteins.[20][41][42] These can be broadly classified into sequence based or structure based.[42] Sequence based methods rely on the assumption that the sequences of functionally conserved portions of proteins such as binding site are conserved. Structure based methods require the 3D structure of the protein. These methods in turn can be subdivided into template and pocket based methods.[42] Template based methods search for 3D similarities between the target protein and proteins with known binding sites. The pocket based methods search for concave surfaces or buried pockets in the target protein that possess features such as hydrophobicity and hydrogen bonding capacity that would allow them to bind ligands with high affinity.[42] Even though the term pocket is used here, similar methods can be used to predict binding sites used in protein-protein interactions that are usually more planar, not in pockets.

Pockets Identification	
CASTp	Automatic Identification of pockets and cavities in proteins structure, and quantitation of their volumes using Delaunay triangulation. Available also as PyMOL plugin
Pocket-Finder	Automatic identification of pockets and cavities in proteins structure, and quantitation of their volumes.
PocketPicker	Grid-based technique for the analysis of protein pockets. PocketPicker available as a plugin for PyMOL
Binding Site Prediction	
ConSurf	Identification of functional regions in proteins by surface-mapping of phylogenetic information
CRESCENDO	Identification protein interaction sites. It uses sequence conservation patterns in homologous proteins to distinguish

	between residues that are conserved due to structural restraints from those due to functional restraints.
Ligand Binding Sites	
3DLigandSite	The server utilizes protein-structure prediction to provide structural models of the binding site. Ligands bound to structures are superimposed onto the model and use to predict the binding site.
FINDSITE	A threading-based method for ligand-binding site prediction and functional annotation based on binding-site similarity across superimposed groups of threading templates.
LIGSITE^{csc}	Prediction of binding site by pocket identification using the Connolly surface and degree of conservation
metaPocket	A meta server for ligand-binding site prediction. metaPocket use LIGSITE^{csc}, PASS, Q-SiteFinder and SURFNET