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 BIOTECHNOLOGY

UNIT – I – Concepts in Molecular Modeling – SBI1310

I. Concepts in 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 scientists 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

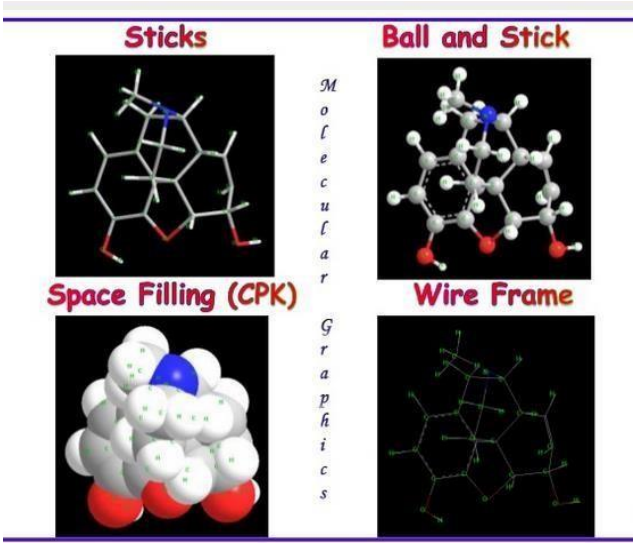
   These models enable the energy of any arrangement if the atoms and mol to be calculated and allow the modeller to determine how the energy of the system varies as the positions of the atoms and molecular changes

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.


Sticks
Ball and Stick
Molecular Graphics
Space Filling (CPK)
Wire Frame

Fig, 1.1

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

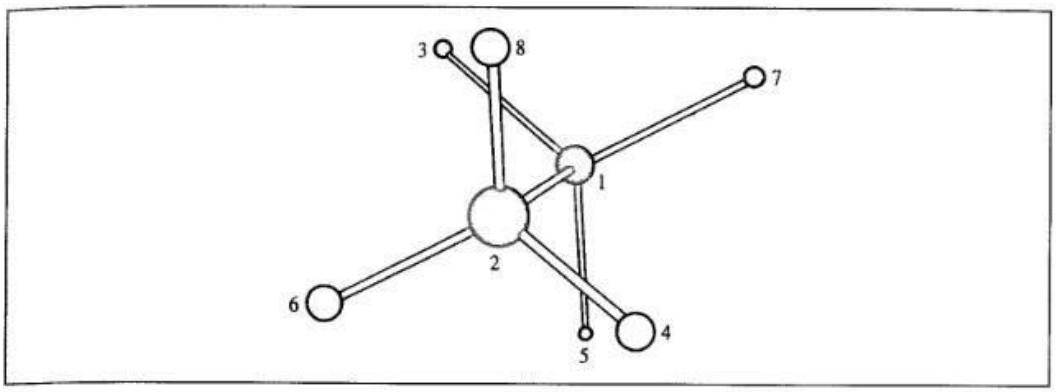


Fig. 1 1 The staggered conformation of ethane

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

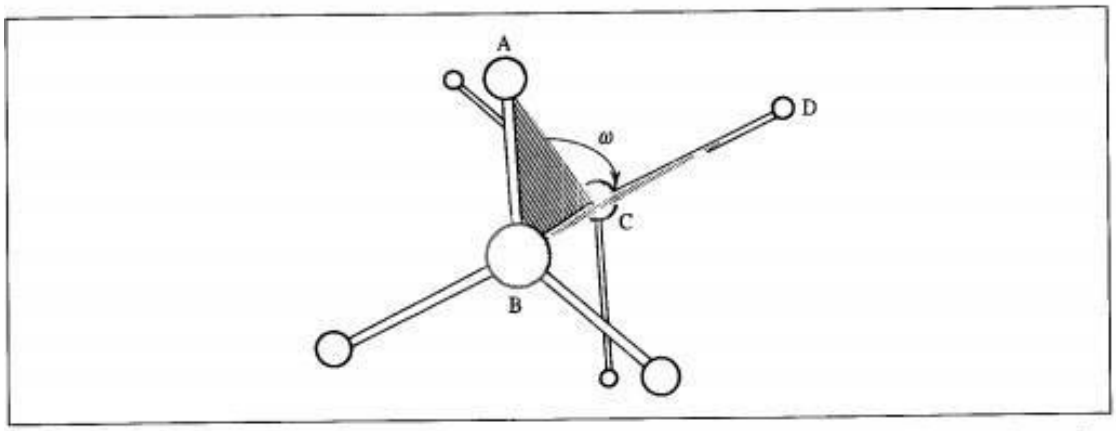


Fig. 1.2 A torsion angle A–B–C–D is defined as the angle between the planes A, B, C and B, C, D A torsion angle can vary through 360° although the range −180° to +180° is most commonly used We shall adopt the IUPAC definition of a torsion angle in which an eclipsed conformation corresponds to a torsion 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

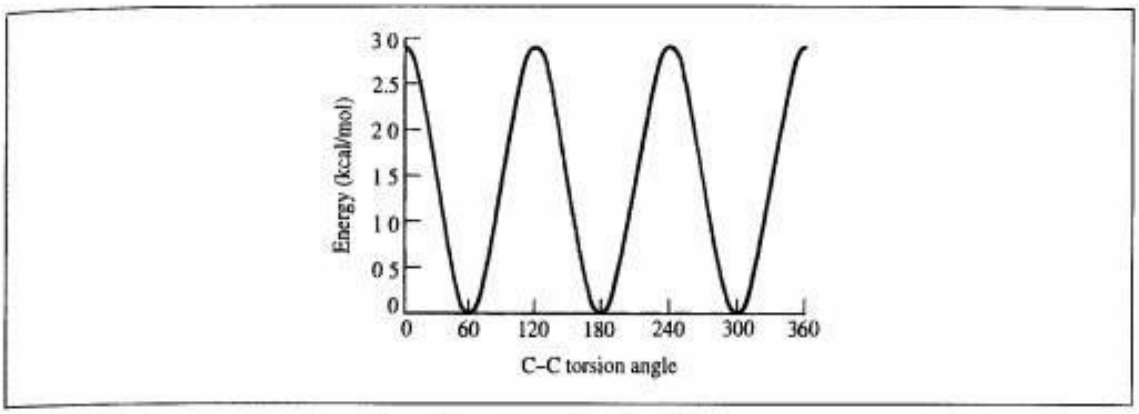


Fig. 1 3 Variation in energy with rotation of the carbon–carbon bond in ethane

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].

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,

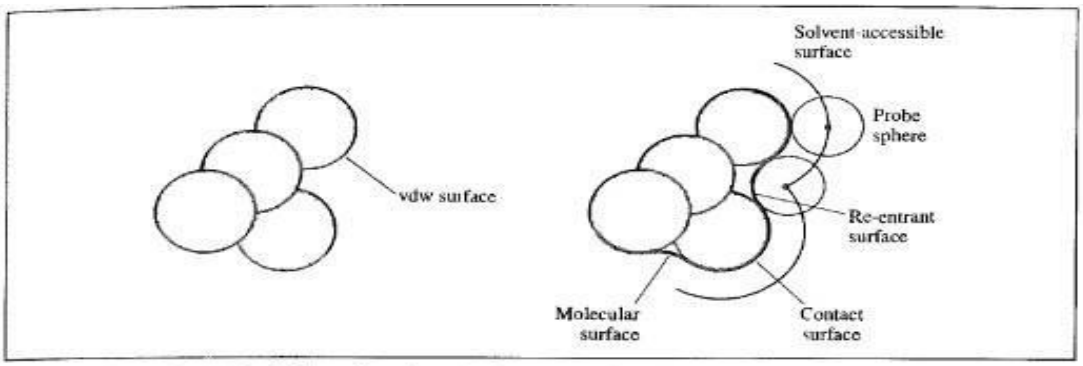


Fig 1.6· The van der Waals (vdw) surface of a molecule corresponds to the outward-facing surfaces of the van der Waals spheres of the atoms. The molecular surface is generated by rolling a spherical probe (usually of radius 1.4 Å to represent a water molecule) on the van der Waals surface. The molecular 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,

The Molecular Modeling Toolbox

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)

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

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: $\Psi(\mathbf{r},t)=\psi(\mathbf{r})T(t)$. 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.

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

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

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

2.1.1 Operators

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:

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

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:

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

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

$$-\frac{\hbar^2}{2m}\nabla^2 \tag{2.9}$$

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

2.2 One-electron Atoms

In an atom that contains a single electron, the potential energy depends upon the distance between the electron and the nucleus as given by the Coulomb equation. The Hamiltonian thus takes the following form:

$$\mathscr{H} = -\frac{\hbar^2}{2m}\nabla^2 - \frac{Ze^2}{4\pi\varepsilon_0 r} \tag{2.16}$$

In atomic units the Hamiltonian is:

$$\mathscr{H} = -\frac{1}{2}\nabla^2 - \frac{Z}{r} \tag{2.17}$$

For the hydrogen atom, the nuclear charge, Z, equals $+1$. r is the distance of the electron from the nucleus. The helium cation, He^+, is also a one-electron atom but has a nuclear charge of $+2$. As atoms have spherical symmetry it is more convenient to transform the Schrödinger equation to polar coordinates r, θ and ϕ, where r is the distance from the nucleus (located at the origin), θ is the angle to the z axis and ϕ is the angle from the x axis in the xy plane (Figure 2.1). The solutions can be written as the product of a radial function $R(r)$, which depends only on r, and an angular function $Y(\theta, \phi)$ called a *spherical harmonic*, which

2.3 Polyelectronic Atoms and Molecules

Solving the Schrödinger equation for atoms with more than one electron is complicated by a number of factors. The first complication is that the Schrödinger equation for such systems cannot be solved exactly, even for the helium atom. The helium atom has three particles (two electrons and one nucleus) and is an example of a *three-body problem*. No exact solutions can be found for systems that involve three (or more) interacting particles. Thus, any solutions we might find for polyelectronic atoms or molecules can only be approximations to the real, true solutions of the Schrödinger equation. One consequence of there being no exact solution is that the wavefunction may adopt more than one functional form; no form is necessarily more 'correct' than another. In fact, the most general form of the wavefunction will be an infinite series of functions.

A second complication with multi-electron species is that we must account for electron spin. Spin is characterised by the quantum number s, which for an electron can only take the value $\frac{1}{2}$. The spin angular momentum is quantised such that its projection on the z axis is either $+\hbar$ or $-\hbar$. These two states are characterised by the quantum number m_s, which can have values of $+\frac{1}{2}$ or $-\frac{1}{2}$, and are often referred to as 'up spin' and 'down spin' respectively. Electron spin is incorporated into the solutions to the Schrödinger equation by writing each one-electron wavefunction as the product of a spatial function that depends on the coordinates of the electron and a spin function that depends on its spin. Such solutions are called *spin orbitals*, which we will represent using the symbol χ. The spatial part (which will be referred to as an orbital and represented using ϕ for atomic orbitals

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 species, H_2^+ (and isotopically equivalent species such as HD^+), when the motion of the electrons is decoupled from the motion of the nuclei in accordance with the Born–Oppenheimer approximation. The masses of the nuclei are much greater than the masses of the electrons (the resting mass of the lightest nucleus, the proton, is 1836 times heavier than the resting mass of the electron). This means that the electrons can adjust almost instantaneously to any changes in the positions of the nuclei. The electronic wavefunction thus depends only on the positions of the nuclei and not on their momenta. Under the Born–Oppenheimer approximation the total wavefunction for the molecule can be written in the following form:

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

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

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.

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.

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 BIOTECHNOLOGY

UNIT –II – Molecular Mechanics – SBI1310

II. Molecular Mechanics

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 force field and mix them with parameters from another force field. The force fields 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 force field. 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 contains 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:

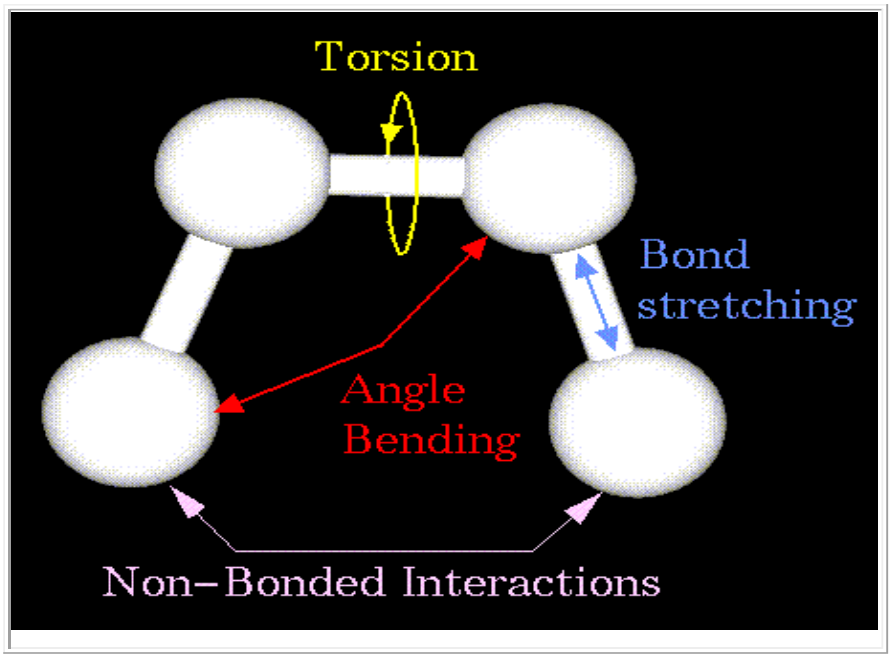


Fig 2.1

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.

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

$\mathscr{V}(\mathbf{r}^N)$ 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

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

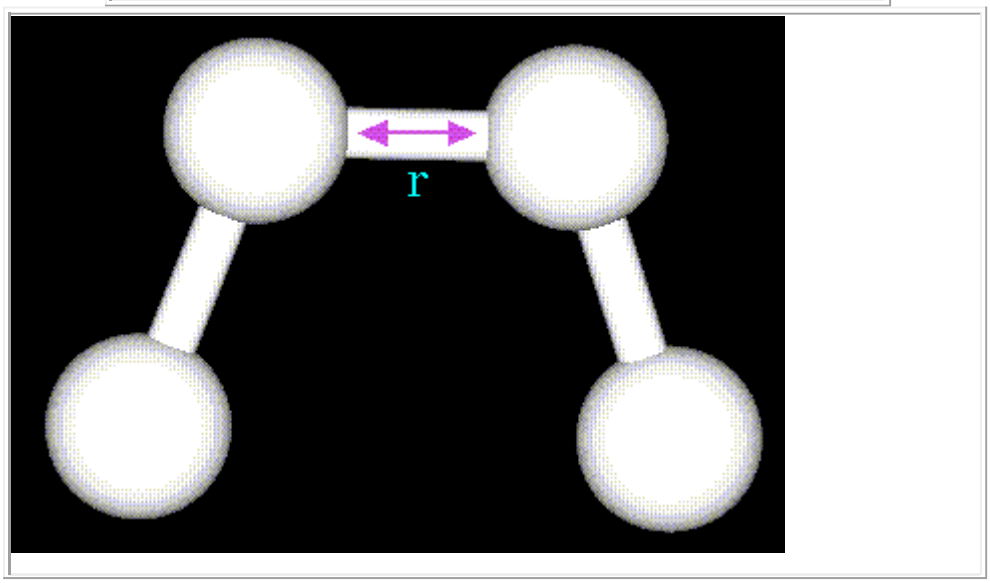


Fig 2.2

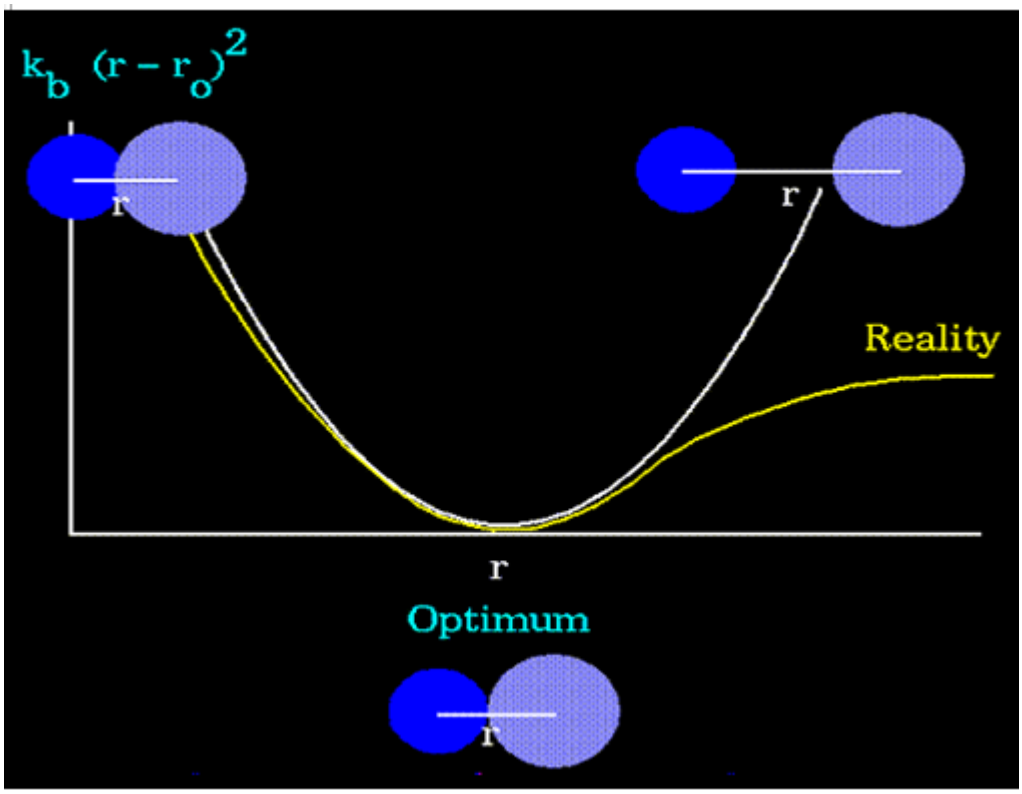


Fig 2.3

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

Bending Energy

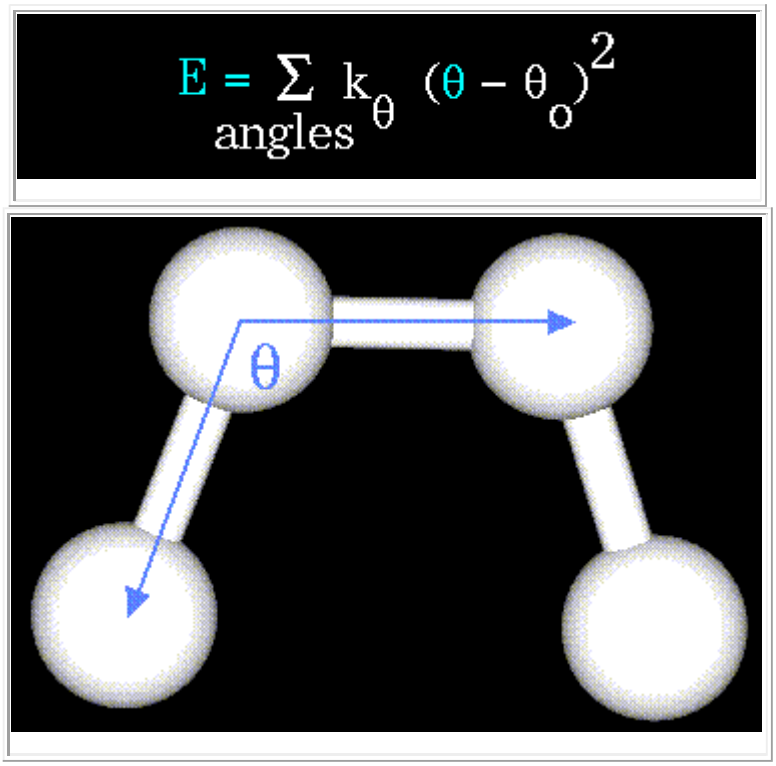


Fig 2.4

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:

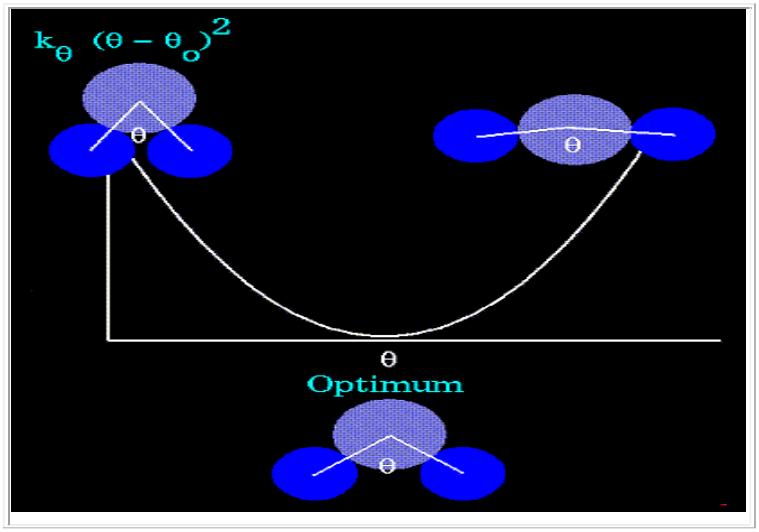


Fig 2.5

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":

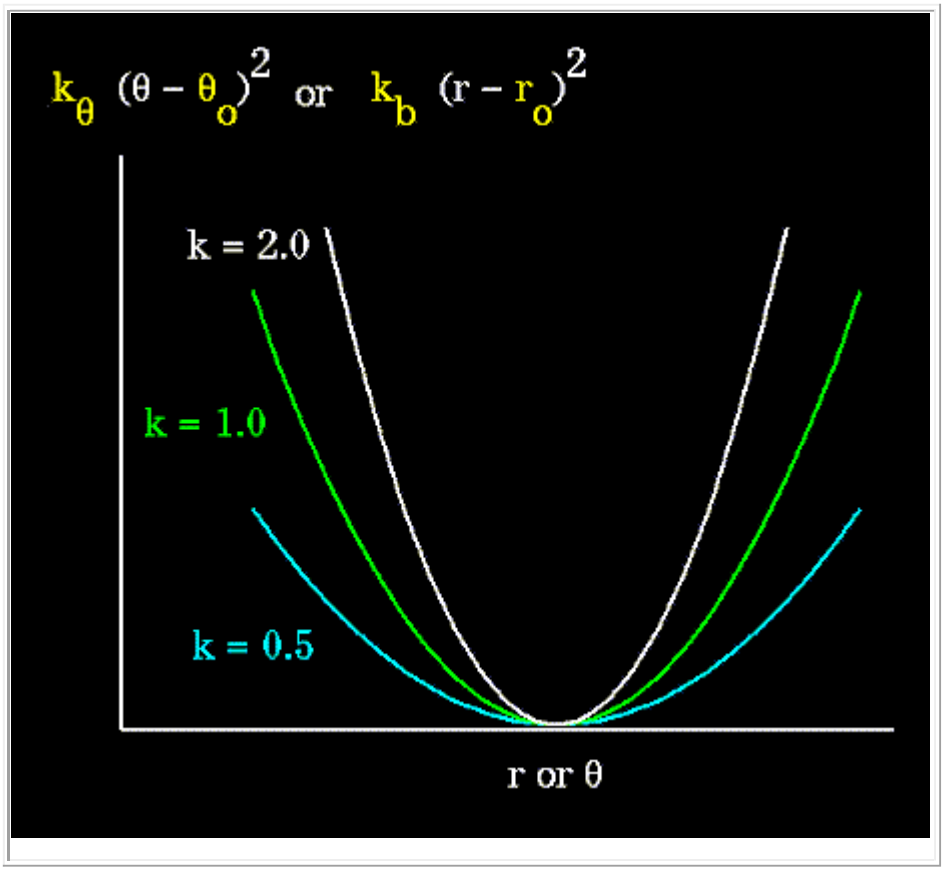


Fig 2.6

Torsion Energy

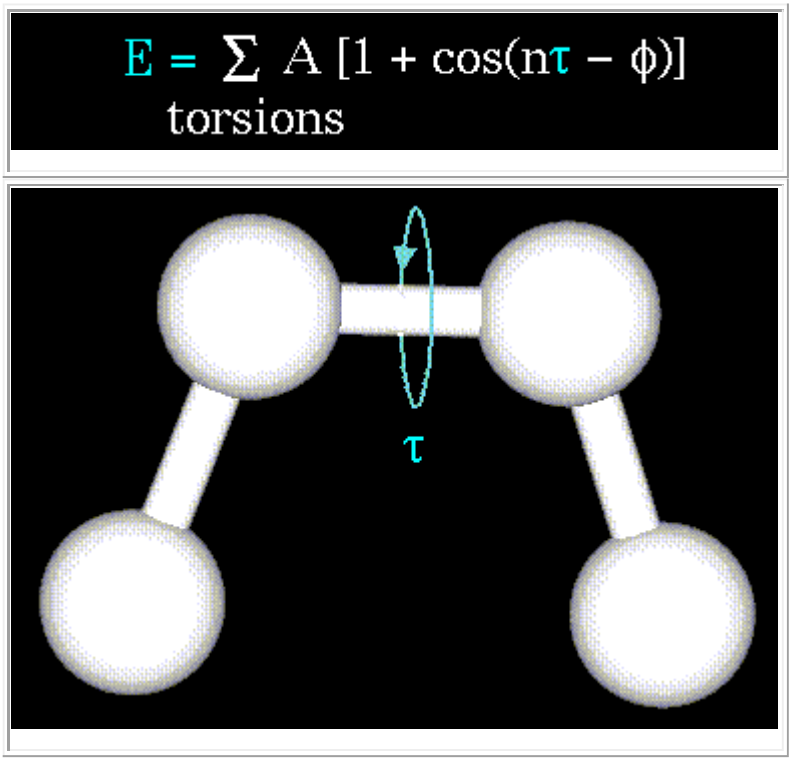


Fig 2.7

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

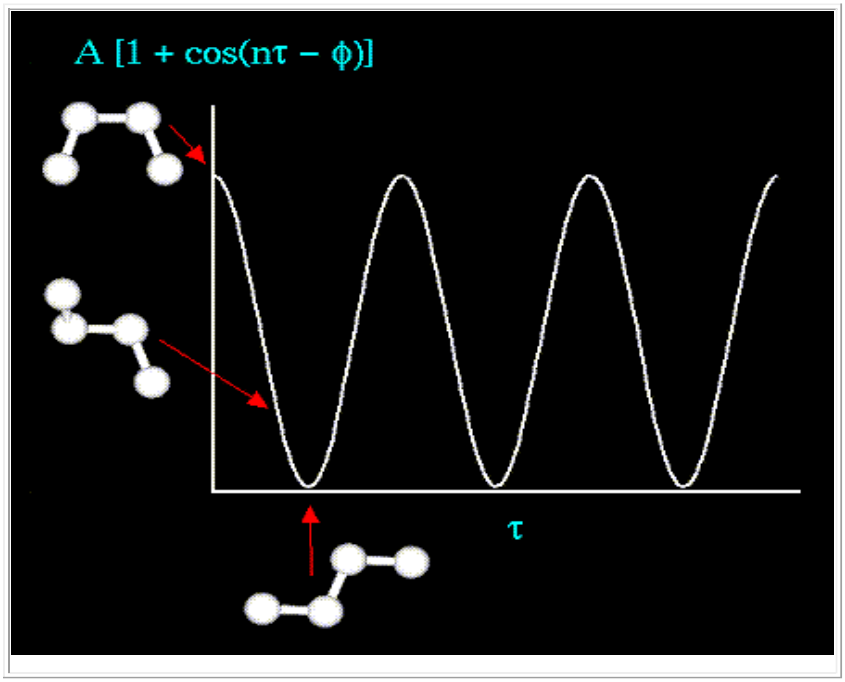


Fig 2.8

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 as 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:

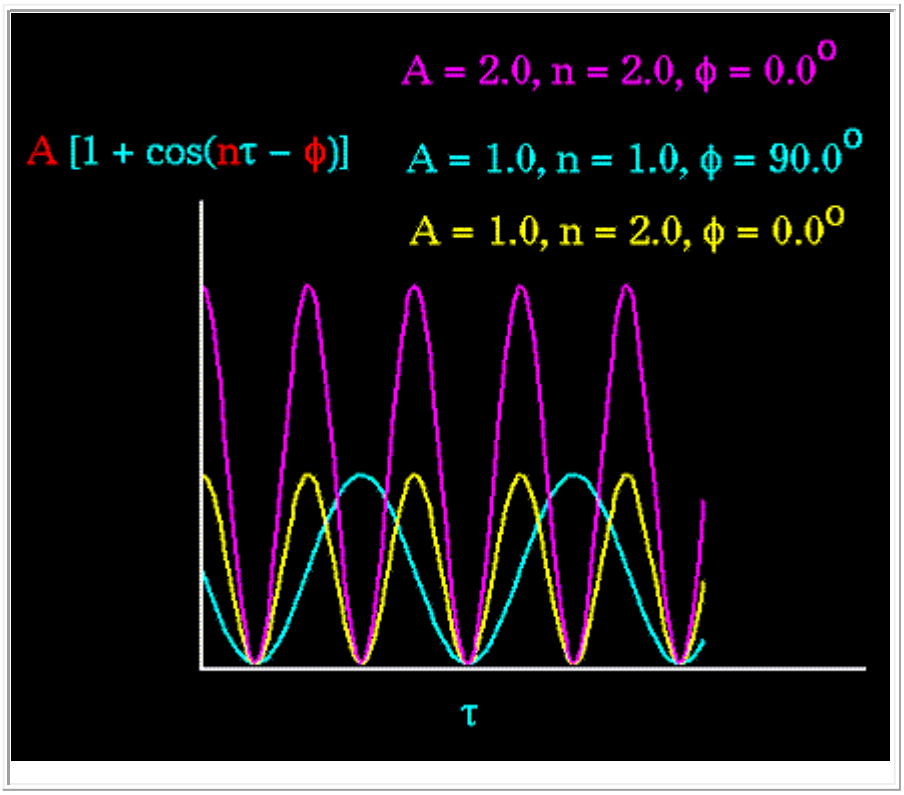


Fig 2.9

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.

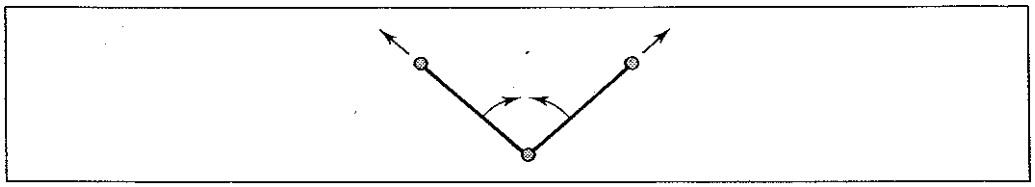


Fig. 4.12: Coupling between the stretching of the bonds as an angle closes.

Fig 2.10

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.

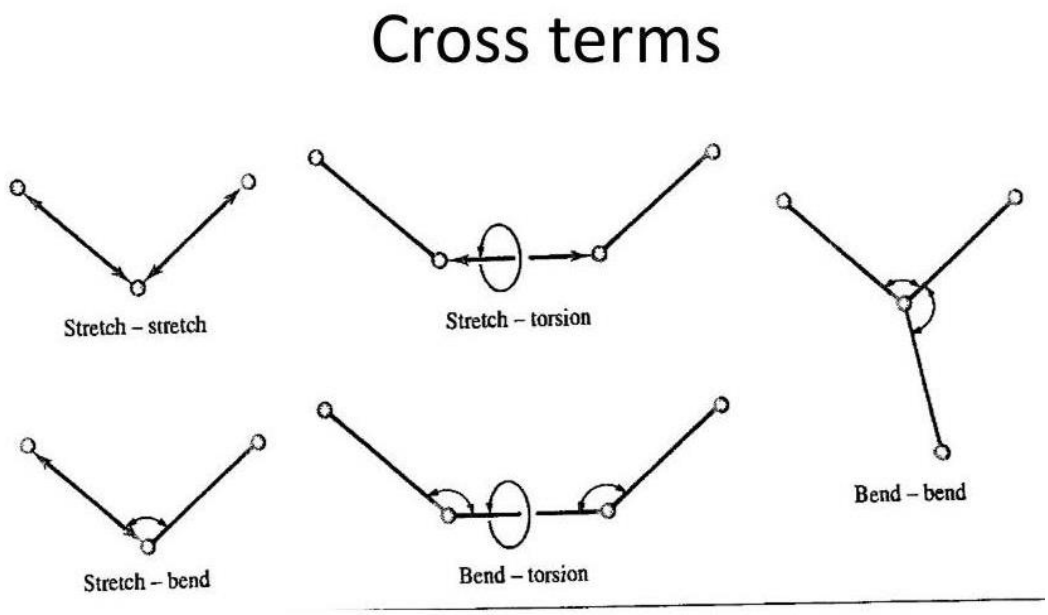


Fig 2.11

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:

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

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):

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

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 inverse power 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

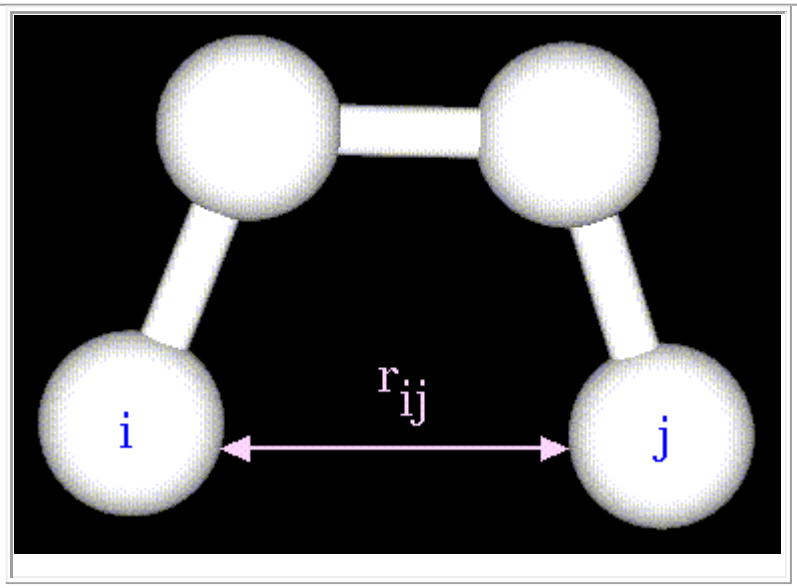


Fig 2.12

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.).

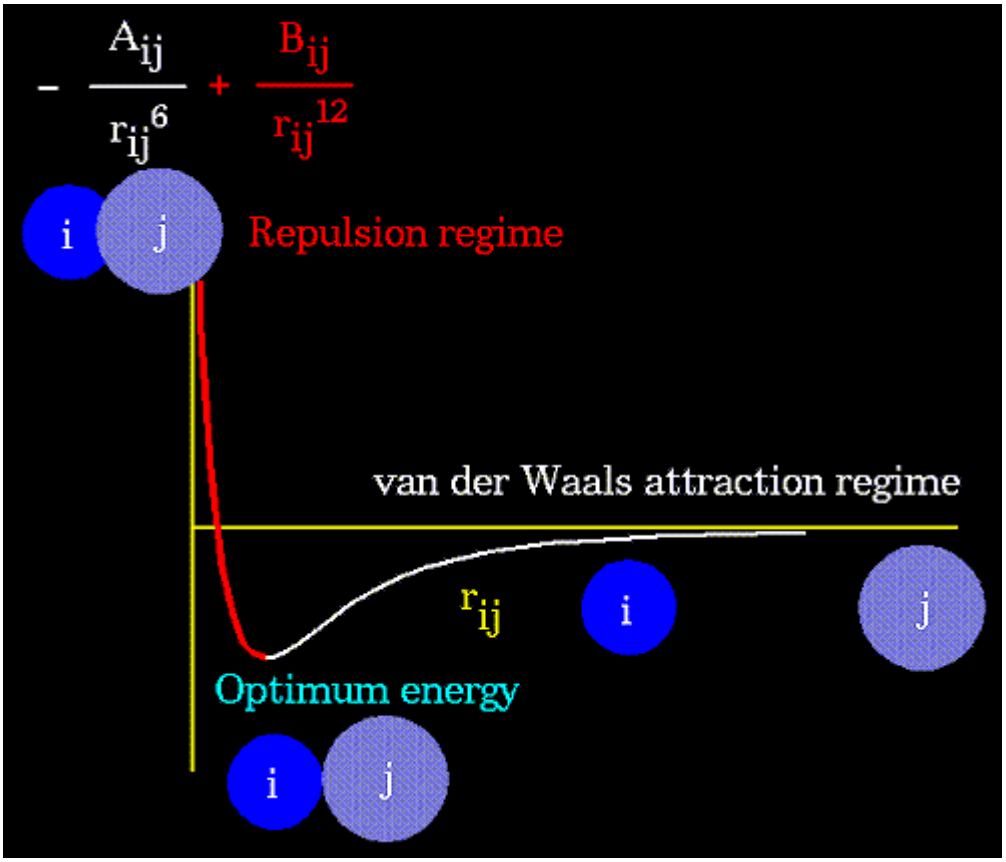


Fig 2.13

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:

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

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

ENERGY MINIMISATION

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.

As an example, when optimizing the geometry of a water molecule, one aims to obtain the hydrogen-oxygen bond lengths and the hydrogen-oxygen-hydrogen bond angle which minimize the forces that would otherwise be pulling atoms together or pushing them apart.
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

- 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

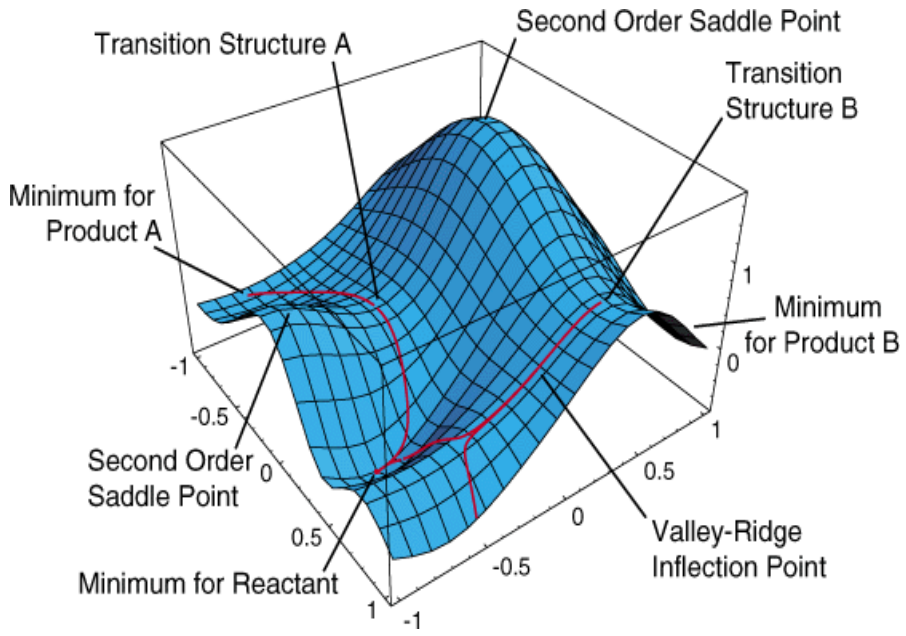


Fig 2.14

•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.

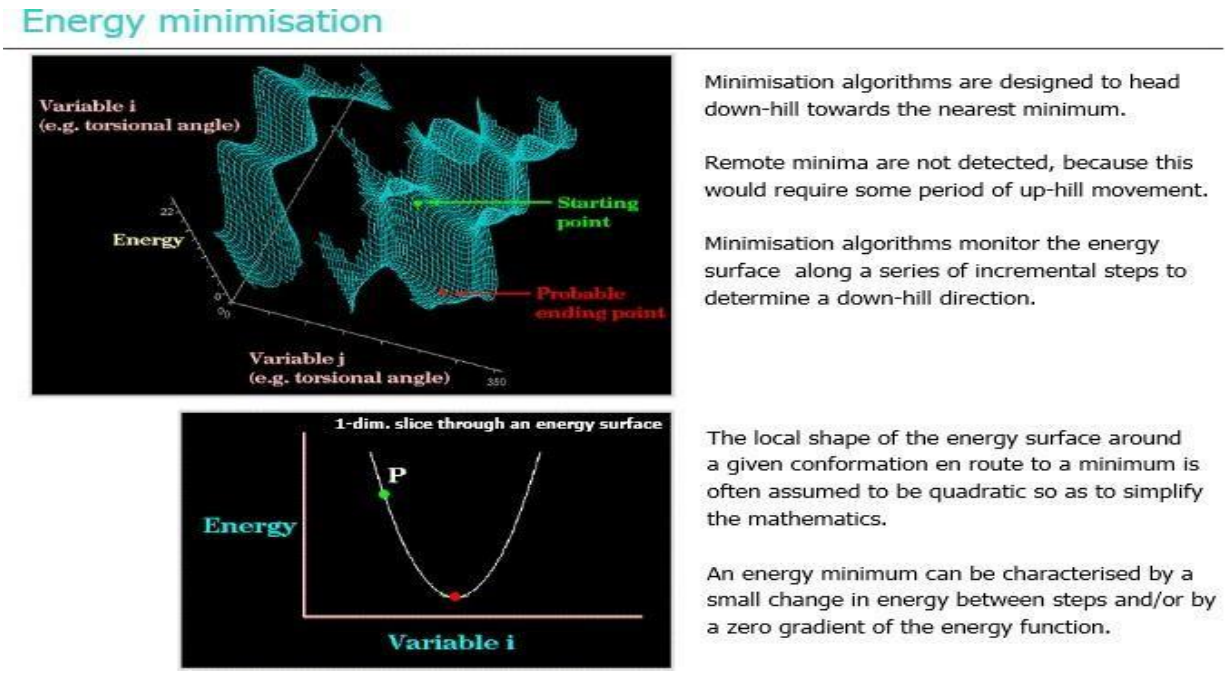


Fig 2.15

Energy minimization algorithms

- Two types
  - Uses the derivatives of energy with respect to coordinates
  - Those that donot use derivatives

Derivatives are useful – provide info on the shape of the energy surface, enhance the efficiency of minima location

Best algorithm – provide quick answer using the least amount of memory

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

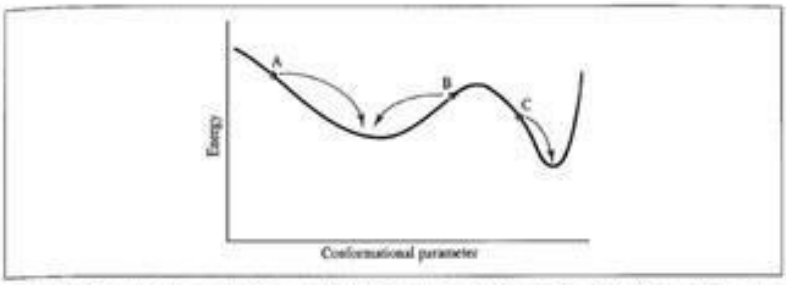


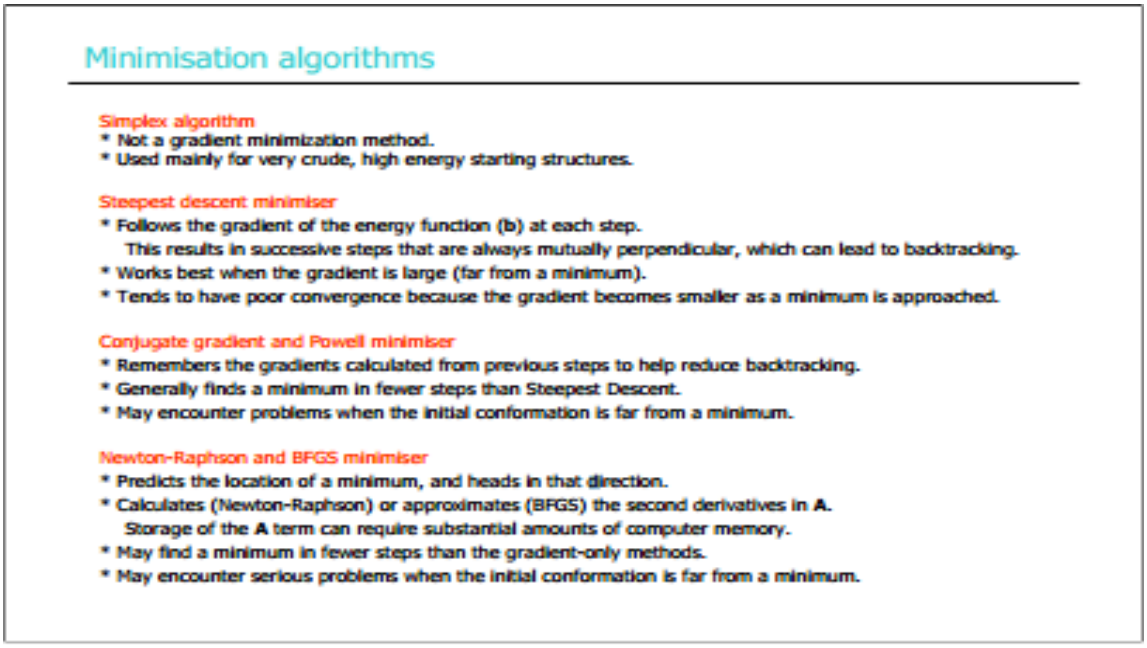
Fig. 3.3: A schematic one-dimensional energy surface. Minimisation methods move downhill to the nearest minimum. The statistical weight of the narrow, deep minimum may be less than a broad minimum which is higher in energy.

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

- First-order minimization: Steepest descent, Conjugate gradient minimization
- Second derivative methods: Newton-Raphson method
- Quasi-Newton methods: L-BFGS

Minimization Methods

- Non Derivative methods
  - Require energy evaluation only and may require many energy evaluations
  - Storage required ~ N^2
  - Simplex Method (Nelder and Mead)
  - Powell's Method (assumes quadratic function)
- Derivative Methods
  - Require evaluation of energy and first derivatives
  - Steepest Descent and Conjugate Gradient
  - Quasi-Newton Methods – DFP, BFGS
  - Full Newton-Raphson requires second derivatives
  - Storage requirements vary from 5N to ~N^2



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

A computer simulation is a simulation, run on a single computer, or a network of computers, to reproduce behavior of a system. The simulation uses an abstract model (a computer model, or a computational model) to simulate the system. Computer simulations have become a useful part of mathematical modeling of many natural systems in physics (computational physics), astrophysics, climatology, chemistry and biology, human systems in economics, psychology, social science, 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.

atoms or molecules. A simulation generates representative configurations of these small replications in such a way that accurate values of structural and thermodynamic properties can be obtained with a feasible amount of computation. Simulation techniques also enable the time-dependent behaviour of atomic and molecular systems to be determined, providing a detailed picture of the way in which a system changes from one conformation or configuration to another. Simulation techniques are also widely used in some experimental procedures, such as the determination of protein structures from X-ray crystallography.

These are smaller replications of larger macromolecules with manageable number of

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).

Two simulation techniques – Molecular dynamics and Monte Carlo

Molecular dynamics

- Calculates the real dynamics of the sys from which time averages s of a property can be calculated
- Deterministic method – state of a system at any future time can be predicted from current state
    - Time dep of the properties of the system
    - Any configuration can be predicted
    - Has KE
    - Constant no of particles (N), V, E

Monte Carlo

- Each configuration is dependent only upon the predecessor
- Generated configurations randiomly and uses a special set of criteria to decide whether or not to accept the config
- Time independent
- Config depends on predecessor
- PE

  Constant N, V, T

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

Conformational Space

- clash-free space - atoms are not in self-collision

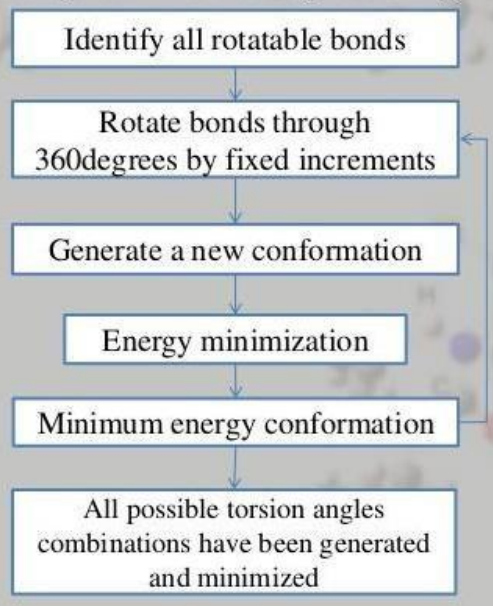
Conformational search methods

- Systematic search algorithms
- Model-building methods
- Random approaches – Generates conformers by random perturbation of Cartesian coordinates or the torsion angles of rotatable bonds
- Distance geometry – Determines the lower and upper distances for all pairs of atoms in the molecule and the distance matrix is generated
- Molecular dynamics

Systematic Method

- Simplest
- Deterministic
- *Grid search*

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

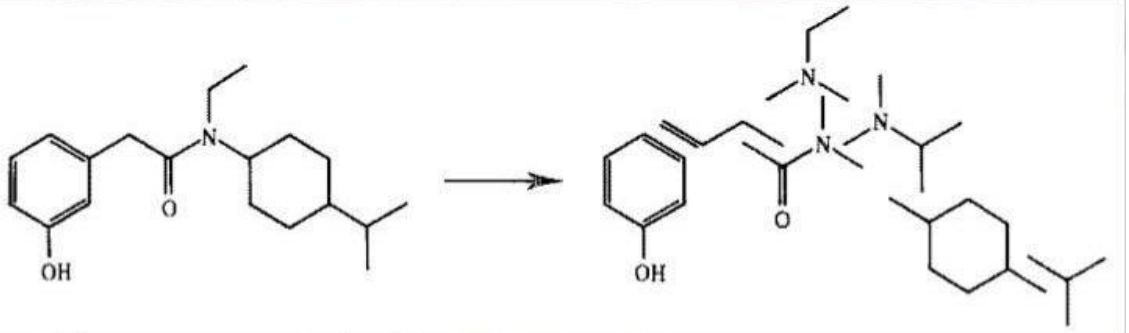


Advantages

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

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.

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 BIOTECHNOLOGY

UNIT – III –Drugs – SBI1310

III. Drugs

A Prodrug can be defined as "pharmacologically inert chemical derivative that can be converted *in vivo*, enzymatically and/or a chemical transformation, to the active drug that exerts the intended therapeutic effect(s)". Ideally, the prodrug should be converted to the parent drug as soon as it reach its goal, and then followed by the subsequent rapid metabolism and/or elimination of the released active group.

Prodrug design can be highly effective for solving many pharmaceutical and pharmacokinetic barriers in clinical drug application such as stability, solubility, permeability and targeting problems in drug discovery and development.

Prodrug advantages:

Prodrugs are used as a way to:

- Increase lipid or water solubility
- Improve that taste of a drug to make it more patient compatible
- Alleviate pain when the drug is administered parenterally by injection
- Reduce toxicity
- Increase chemical stability
- Increase biological stability
- Change the length of the time of duration of action
- Deliver the drug to a specific site in the body

Prodrug classification:

There are potentially many methods of classifying prodrugs and these are based on the following aspects:

1. Therapeutic categories; for example, anticancer prodrugs, antiviral prodrugs, antibacterial prodrugs, non-steroidal anti-inflammatory prodrugs, cardiovascular prodrugs, etc.

- Double prodrugs, pro-prodrugs or cascade-latentiated prodrugs, where a prodrug is further derivatized in a fashion such that only enzymatic conversion to prodrug is possible before the latter can cleave to release the active drug.

Macromolecular prodrugs, where macromolecules like polysaccharides, dextrans, cyclodextrins, proteins, peptides, and polymers are used as carriers. The development of macromolecular prodrugs of NSAIDs is advantageous because of the fact that these formulations show sustained release of drug, colon-targeted drug delivery, reduction in the administration frequency and better patient compliance.

Site-specific prodrugs where a carrier acts as a transporter of the active drug to a specific targeted site.

Mutual prodrug, where the carrier used is another biologically active drug instead of inert molecule. A mutual prodrug consists of two pharmacologically active agents coupled together so that each acts as a promoiety for the other agent and vice versa. The carrier selected may have the same biological action as that of the parent drug and thus might give synergistic action, or the carrier may have some additional biological action that is lacking in the parent drug, thus ensuring some additional benefit. The carrier may also be a drug that might help to target the parent drug to a specific site or organ or cells or may improve site specificity of a drug. The carrier drug may be useful to overcome some side effects of the parent drug as well.

DRUG TARGETS

A **biological target** is anything within a living organism to which some other entity, like an endogenous ligand or a drug is directed and/or binds. Examples of common classes of biological targets are proteins and nucleic acids. The definition is context-dependent and can refer to the biological target of a pharmacologically active drug compound, the receptor target of a hormone (like insulin), or some other target of an external stimulus. The implication is that a target is "hit" by a signal and its behavior or function is then changed. Biological targets are most commonly proteins such as enzymes, ion channels, and receptors.

Mechanism

The external stimulus (*i.e.*, chemical substance) physically binds to the biological target. The interaction between the substance and the target may be:

- noncovalent – A relatively weak interaction between the stimulus and the target where no chemical bond is formed between the two interacting partners and hence the interaction is completely reversible.

- reversible covalent - A chemical reaction occurs between the stimulus and target in which the stimulus becomes chemically bonded to the target, but the reverse reaction also readily occurs in which the bond can be broken.

- irreversible covalent - The stimulus is permanently bound to the target through irreversible chemical bond formation.

Depending on the nature of the stimulus, the following can occur:

- There is no direct change in the biological target, except that the binding of the substance prevents other endogenous substances such as activating hormone to bind to the target. Depending on the nature of the target, this effect is referred as receptor antagonism, enzyme inhibition, or ion channel blockade.
- A conformational change in the target is induced by the stimulus which results in a change in target function. This change in function can mimic the effect of the endogenous substance in which case the effect is referred to as receptor agonism (or channel or enzyme activation) or be the opposite of the endogenous substance which in the case of receptors is referred to as inverse agonism.

Drug targets

The term biological target is frequently used in pharmaceutical research to describe the native protein in the body whose activity is modified by a drug resulting in a specific effect, which may be a desirable therapeutic effect or an unwanted adverse effect. In this context, the biological target is often referred to as a **drug target**. The most common drug targets of currently marketed drugs include:

- proteins
  - G protein-coupled receptors (target of 50% of drugs)
  - enzymes (especially protein kinases, proteases, esterases, and phosphatases)
  - ion channels
  - ligand-gated ion channels
  - voltage-gated ion channels
  - nuclear hormone receptors
  - structural proteins such as tubulin

membrane transport proteins

- nucleic acid

Drug target identification

Identifying the biological origin of a disease, and the potential targets for intervention, is the first step in the discovery of a medicine. This has been a great challenge for both academia and industry. Number of different approaches and technologies are reviewed.

Databases

Databases containing biological targets information

- Therapeutic Targets Database (TTD)
- DrugBank
- Binding DB

RECEPTOR- DRUG TARGET

Salbutamol- drug

Salbutamol is a highly selective β2-adrenergic receptor stimulating drug that has a bronchodilator effect. It is used to relieve bronchospasm in bronchial asthama, chronic bronchitis, emphysema and other airway resistance diseases.

General pharmacology:

► The chemical name of salbutamol is 1-(4-hydroxy-3-hydroxymethylphenyl)-2-(t-butylamino)-ethanol sulphate, molecular formula is $(C_{13}H_{21}NO_3)_2 \cdot H_2SO_4$ and molecular weight is 576.7.

► After oral administration, approximately 50% of salbutamol is absorbed from the intestinal tract with a slower onset of action, reaching a peak at about 2 hours after intake. After inhalation, salbutamol reaches the lungs directly and acts within 3-5 minutes with a peak at 15-20 minutes. Overall duration of action of salbutamol is 4-6 hours. It is metabolized in the intestinal tract and in the liver and is excreted via the urine. Learn more.

Mechanism of action:

► Salbutamol stimulates β2 adrenergic receptors which are predominant receptors in bronchial smooth muscle of the lung. Stimulation of β2 receptors leads to the activation of enzyme adenyl cyclase that form cyclic AMP (adenosine-mono-phosphate) from ATP

(adenosine-tri-phosphate). This high level of cyclic AMP relaxes bronchial smooth muscle and decreases airway resistance by lowering intracellular ionic calcium concentrations. Salbutamol relaxes the smooth muscles of airways, from trachea to terminal bronchioles.

► High level of cyclic AMP are also inhibits the release of broncho constrictor mediators such as histamine, leukotreine from the mast cells in the airway.

Usage:

1. Bronchospasm with reversible obstructive airway diseases

Salbutamol is indicated for the preventation or treatment of bronchospasm with reversible obstructive airway diseases such as

► Bronchial asthama

► Chronic obstructive pulmonary disease (COPD) which includes chronic bronchitis and emphysema

2. Exercise-induced bronchospasm

► Salbutamol is used for the prevention of exercise-induced bronchospasm.

3. Any other situations known to induce bronchospasm.

LIPID AS A DRUG TARGET

Drugs acting on cell membrane lipids - Anaesthetics and some Antibiotics

Action of amphotericin B (antifungal agent) - builds tunnels through membrane and drains cell

Amphotericin B is an antifungal drug often used intravenously for serious systemic fungal infections and is the only effective treatment for some fungal infections. Common side effects include a reaction of fever, shaking chills, headaches and low blood pressure soon after it is infused, as well as kidney and electrolyte problems. Allergic symptoms including anaphylaxis may occur. It was originally extracted from *Streptomyces nodosus,* a filamentous bacterium, in 1955, at the Squibb Institute for Medical Research. Its name originates from the chemical's amphoteric properties. It is on the World Health Organization's List of Essential Medicines, the most important medications needed in a basic health system.

Mechanism of action

As with other polyene antifungals, amphotericin B binds with ergosterol, a component of fungal cell membranes, forming a transmembrane channel that leads to monovalent ion (K^+,

Na^+, H^+ and Cl^-) leakage, which is the primary effect leading to fungal cell death. Researchers have found evidence that pore formation is not the only mechanism responsible for cell death.

By an unknown mechanism, amphotericin B also causes oxidative stress within the fungal cell and the addition of free radical scavengers or induction of antioxidant enzymes in pathogens can lead to amphotericin resistance in species such as *Scedosporium prolificans* without having to effect cell wall ergosterol.

Two amphotericins, amphotericin A and amphotericin B, are known, but only B is used clinically, because it is significantly more active *in vivo*. Amphotericin A is almost identical to amphotericin B (having a double C=C bond between the 27th and 28th carbons), but has little antifungal activity.

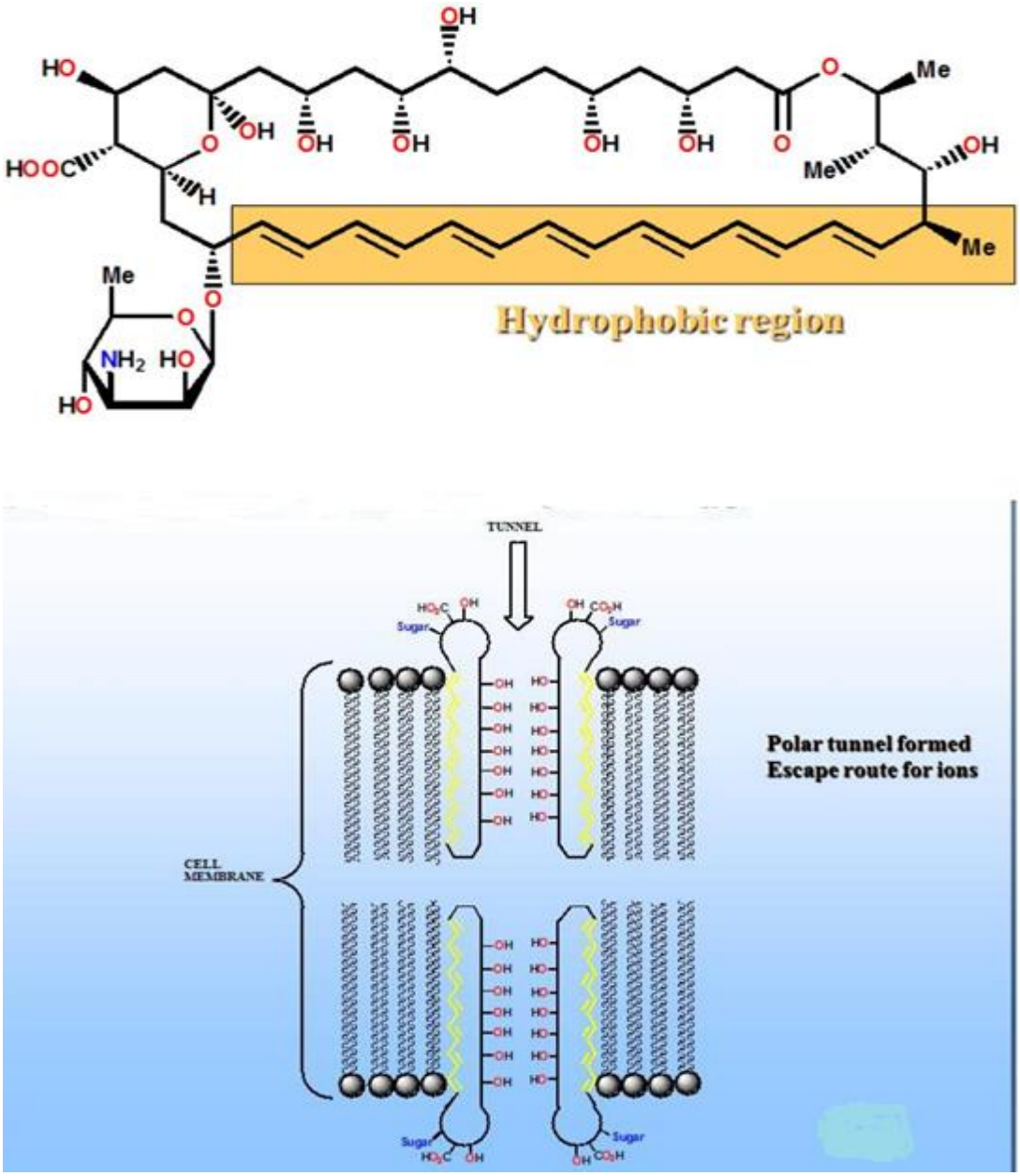


Fig 3.1 -Structure of amphotericin B bound to cell membrane

Fig 2 – structure of **amphotericin B bound** cell membrane

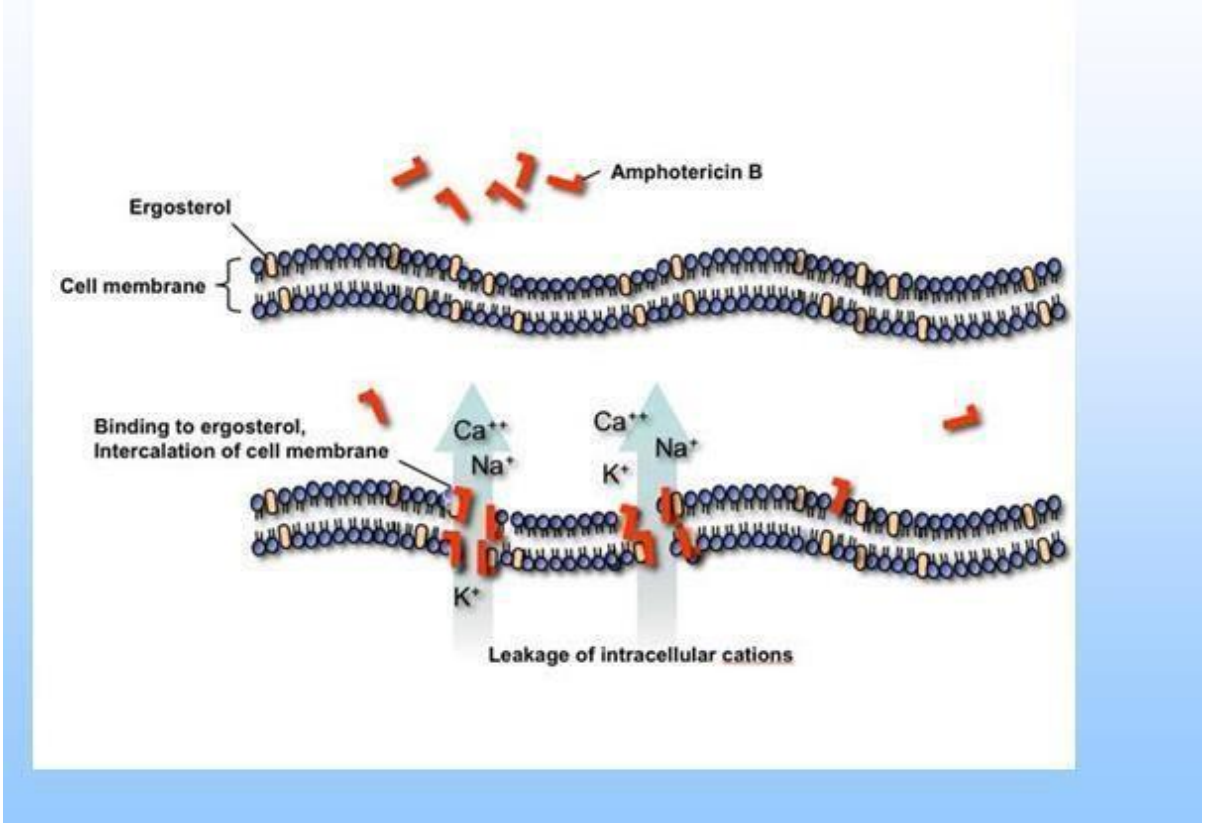


Fig 3.2 Drug action of amphotercin

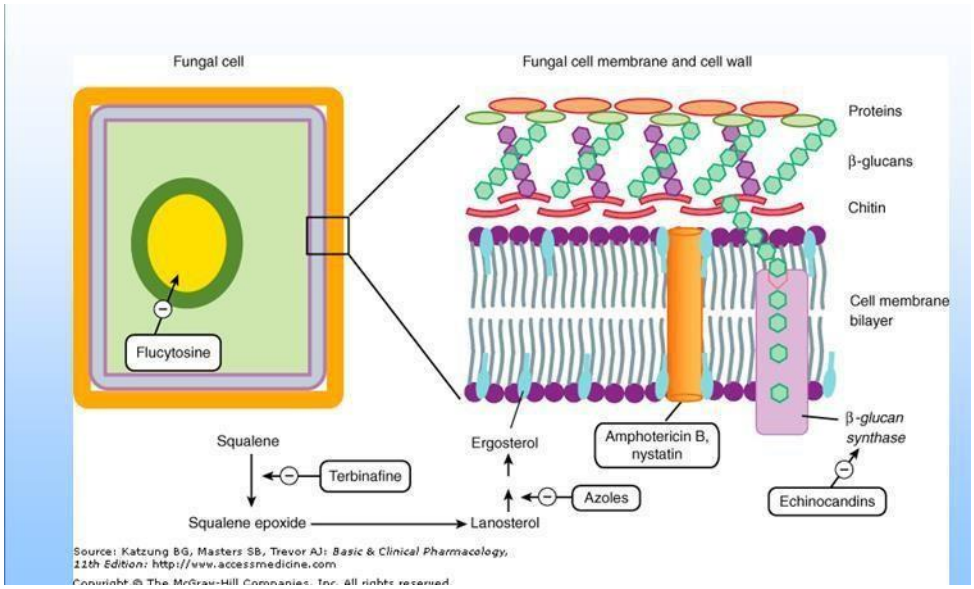


Fig 3.3 Fungal drug target

ION CHANNEL AS DRUG TARGET

Local Anaesthetics

Koller introduced the ester cocaine into clinical practice for eye surgery in 1884 because the conditions provided by general anaesthesia were poor. It is interesting that the use of local anaesthesia for eye surgery has once more become very popular, although much safer drugs than cocaine are now employed. In 1948 Gordh was the first to use the amide drug, lignocaine; the amide local anaesthetics are used now in preference to the esters in the U.K. as they have fewer undesirable effects. Local anaesthetics are either aminoesters (e.g. procaine) or aminoamides (e.g. lignocaine) which reversibly inhibit nerve conduction.

Mechanisms of action

Local anaesthetics inhibit nerve conduction by interfering with the physiological changes in ionic permeability during an action potential. Nerve cells are selective in their permeability to ions and consequently have an electrical potential across their membrane; at rest this is of the order of minus 50 to minus 80 mV, with the inside being negative. Cell membranes are composed mainly of lipids and do not permit ions to pass through them, but they are crossed by specialised protein ion channels, which allow potassium, sodium and other ions to pass through.

At rest, the potassium channels in nerve cell membranes are open and the sodium gates are closed; when a nerve cell is excited, the membrane suddenly becomes transiently permeable to sodium as that ionic channel opens. The membrane potential is reversed so that it has a positive charge inside, and a propagated action potential is passed along the fibre. Local anaesthetics block sodium channels, prevent the evolution of the action potential and so prevent or decrease sensation arising in the affected area. It is thought that most local anaesthetics work by blocking the sodium channel from the inside of the cell into which they must first diffuse before they can act. In infected tissues, acidic conditions prevent this diffusion and thus local anaesthetics then tend to be less effective.

Pain Management

Local anaesthetics are used on their own and combination with opioids for epidural and spinal blocks. Local anaesthetics are also used for local blocks and are used extensively for day case surgery, limb surgery and hand surgery. Local anaesthetics can also be used

systemically for pain management. Sodium channel blockers can be used to reduce pain due to nerve damage and intravenous lignocaine and oral mexiletine (an oral analogue) can both reduce neuropathic pain in nonmalignant and cancer pain.

Side effects/complications

Adverse effects to local anaesthetics can be due to the use of excessive doses, abnormal reactions to normal doses, or to toxicity or depression of vital centres after inadvertent injection into the bloodstream or the cerebrospinal fluid. Toxic reactions to local anaesthetics can be reduced by slow administration, and intravenous access should always be secured before a block is performed in case of untoward events occurring. Resuscitation equipment and drugs should be immediately available. The effects of local anaesthetics are as follows:

- Central nervous system - is particularly sensitive to the effects of local anaesthetics and with increasing blood concentrations predictable consequences present. Early signs of toxicity are shivering, confusion, and twitching and tremors followed by generalised seizures. Eventually, with large doses, generalised central nervous system depression ensues with cessation of seizures, respiratory arrest and hypoxia. Treatment comprises the administration of anticonvulsants (thiopentone or diazepam) and oxygenation, with tracheal intubation and respiratory support if necessary.
- Cardiovascular system - is more resistant to local anaesthetics, but vasodilatation, myocardial depression and disorders of rhythm occur and can lead to cardiac arrest and circulatory collapse. Cardiovascular toxicity may be precipitated and worsened by hypoxia, hypercarbia and acidosis consequent to inadequate treatment of the convulsions and respiratory arrest described above. In particular, hypoxia and acidosis potentiate the cardiodepressant effects and arrhythmias associated with bupivacaine toxicity. Cardiac massage, ventricular defibrillation, intravenous fluids and inotropic support are indicated and resuscitation may be prolonged, especially with bupivacaine.

- Allergic reactions - to local anaesthetics are rare and most involve the aminoesters. There is also cross-sensitivity between the para-aminobenzoic acid derivatives and methylparaben, a preservative commonly used in local anaesthetic preparations.

  Allergy to amide local anaesthetics is rare, and almost all have been related to methylparaben.

- Methaemoglobin - The administration of large doses of prilocaine (10 mg/kg) may lead to the accumulation of an oxidising agent, which converts haemoglobin to methaemoglobin. Patients may appear cyanosed at a methaemoglobin concentration of 3 - 5 g/100 ml of blood, but in healthy individuals this should not present a problem. In patients who have other cardiorespiratory abnormalities, immediate treatment for methaemoglobinaemia may be required and the reducing agent methylene blue, 1 - 5 mg/kg, should be given intravenously. Lignocaine also produces methaemoglobin, but a clinical problem rarely presents. Prilocaine may be a problem if EMLA is used in large quantities on premature babies.

Specific drugs

Lignocaine

This is the most commonly used agent in the U.K.; it is available in solutions of 0.5 - 2%. The effect of lignocaine is prolonged considerably by the addition of the vasoconstrictor adrenaline.

Bupivacaine

Bupivacaine is more potent than lignocaine; 0.5% bupivacaine is as effective as 2% lignocaine. It is available in 0.25 - 0.75% concentrations. It is more cardiotoxic than other local anaesthetics and is not recommended for intravenous regional analgesia. The duration of action is from 4 to 16 hours, and bupivacaine produces more sensory than motor block. Levo-bupivacaine is also available and while the concentrations and usage are the same as bupivacaine, evidence suggests that it may be less cardiotoxic.

Prilocaine

The potency of prilocaine is similar to lignocaine but as it is metabolised in the lung as well as the liver it is cleared from the body more quickly than the other amides (this makes it particularly useful for intravenous regional analgesia). Methaemoglobinaemia is associated with the use of high doses and it is unsuitable for use in obstetrics because of this risk to the unborn child.

Ropivacaine

Ropivacaine is a long acting local anaesthetics like bupivacaine but is associated with less cardiovascular toxicity. It is one of the newer local anaesthetics.

ENZYME AS DRUG TARGET

Arbutin

- Arbutin is one of the most widely prescribed skin-lightening and de-pigmenting agent worldwide.
- Arbutin, the b-D-glucopyranoside derivative of hydroquinone, is a naturally occurring plant derived compound found in dried leaves of a number of different plant species including, bearberry, blueberry, cranberry and pear trees.
- Arbutin, inhibits tyrosinase activity competitively but at non-cytotoxic concentrations in a dose dependent manner in cultured melanocytes.
- It also inhibits melanosome maturation and is less cytotoxic to melanocytes than hydroquinone.
- Although, higher concentrations may be more efficacious, greater risk for paradoxical hyperpigmentation exists.
- Controlled trials on treating hyperpigmentation are lacking.

How alpha arbutin works

Arbutin inhibits the formation of melanin pigment by inhibiting Tyrosinase activity.

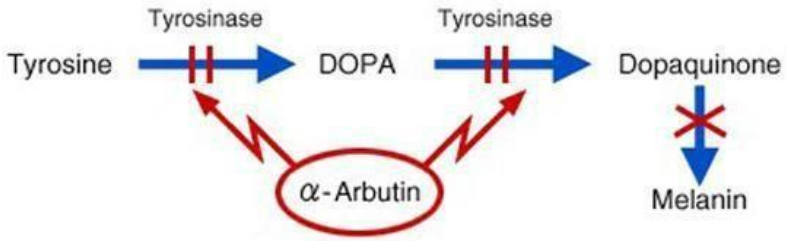


Fig 3.4

PENICILLINS- EXAMPLE FOR ENZYME AS DRUG TARGET

Bacteria constantly remodel their peptidoglycan cell walls, simultaneously building and breaking down portions of the cell wall as they grow and divide. *β*-Lactam antibiotics inhibit the formation of peptidoglycan cross-links in the bacterial cell wall; this is achieved through binding of the four-membered *β*-lactam ring of penicillin to the enzyme DD-transpeptidase. As a consequence, DD-transpeptidase cannot catalyze formation of these cross-links, and an imbalance between cell wall production and degradation develops, causing the cell to rapidly die. The enzymes that hydrolyze the peptidoglycan cross-links continue to function, even while those that form such cross-links do not. This weakens the cell wall of the bacterium, and osmotic pressure becomes increasingly uncompensated—eventually causing cell death (cytolysis). In addition, the build-up of peptidoglycan precursors triggers the activation of bacterial cell wall hydrolases and autolysins, which further digest the cell wall's peptidoglycans. The small size of the penicillins increases their potency, by allowing them to penetrate the entire depth of the cell wall. This is in contrast to the glycopeptide antibiotics vancomycin and teicoplanin, which are both much larger than the penicillins.

Gram-positive bacteria are called protoplasts when they lose their cell walls. Gram- negative bacteria do not lose their cell walls completely and are called spheroplasts after treatment with penicillin. Penicillin shows a synergistic effect with aminoglycosides, since the inhibition of peptidoglycan synthesis allows aminoglycosides to penetrate the bacterial cell wall more easily, allowing their disruption of bacterial protein synthesis within the cell. This results in a lowered MBC for susceptible organisms.

Penicillins, like other *β*-lactam antibiotics, block not only the division of bacteria, including cyanobacteria, but also the division of cyanelles, the photosynthetic organelles of the glaucophytes, and the division of chloroplasts of bryophytes. In contrast, they have no effect on the plastids of the highly developed vascular plants. This supports the endosymbiotic theory of the evolution of plastid division in land plants.

The chemical structure of penicillin is triggered with a very precise, pH-dependent

directed mechanism, effected by a unique spatial assembly of molecular components, which can activate by protonation. It can travel through bodily fluids, targeting and inactivating enzymes responsible for cell-wall synthesis in gram-positive bacteria, meanwhile avoiding the surrounding non-targets.

Penicillin can protect itself from spontaneous hydrolysis in the body in its anionic form, while storing its potential as a strong acylating agent, activated only upon approach to the target transpeptidase enzyme and protonated in the active centre.

This targeted protonation neutralizes the carboxylic acid moiety, which is weakening of the β- lactam ring N–C(=O) bond, resulting in a self-activation. Specific structural requirements are equated to constructing the perfect mouse trap for catching targeted prey.

DNA AS DRUG TARGET

Ex- Doxorubicin

Mechanism of action

Diagram of two doxorubicin molecules intercalating DNA, from PDB:1D12. Doxorubicin interacts with DNA by intercalation and inhibition of macromolecular biosynthesis. This inhibits the progression of the enzyme topoisomerase II, which relaxes supercoils in DNA for transcription. Doxorubicin stabilizes the topoisomerase II complex after it has broken the DNA chain for replication, preventing the DNA double helix from being resealed and thereby stopping the process of replication. It may also increase quinone type free radical production, hence contributing to its cytotoxicity. The planar aromatic chromophore portion of the molecule intercalates between two base pairs of the DNA, while the six-membered daunosamine sugar sits in the minor groove and interacts with flanking base pairs immediately adjacent to the intercalation site, as evidenced by several crystal structures. By intercalation, doxorubicin can also induce histone eviction from transcriptionally active chromatin. As a result, DNA damage response, epigenome and transcriptome are deregulated in doxorubicin-exposed cells.

PROTEIN AS DRUG TARGET

Omeprazole, sold under the brand names **Prilosec** and **Losec** among others, is a medication used to treat gastroesophageal reflux disease, peptic ulcer disease, and Zollinger–Ellison syndrome. It is also used to prevent upper gastrointestinal bleeding in people who are at high risk. It is taken by mouth.

Mechanism of action

Omeprazole is a selective and irreversible proton pump inhibitor. It suppresses stomach acid secretion by specific inhibition of the H^+/K^+-ATPase system found at the secretory surface of gastric parietal cells. Because this enzyme system is regarded as the acid (proton, or H^+) pump within the gastric mucosa, omeprazole inhibits the final step of acid production. Omeprazole also inhibits both basal and stimulated acid secretion irrespective of the stimulus.

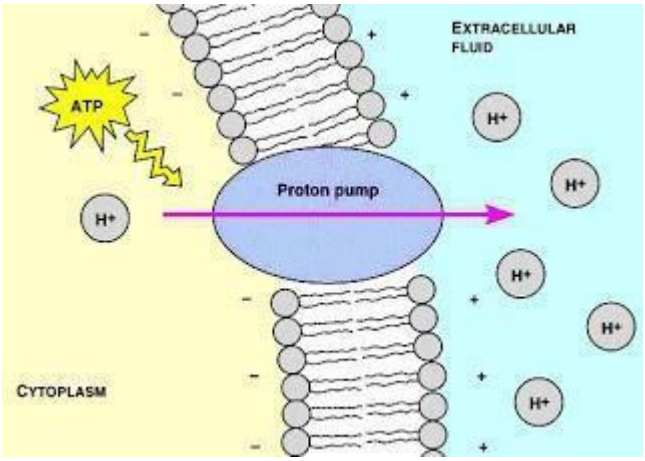


Fig 3.5 - Proton pump

A **proton pump** is an integral membrane protein that is capable of moving **protons** across a biological membrane. Mechanisms are based on conformational changes of the protein structure or on the Q cycle.

PHARMACO KINETICS

OVERVIEW

Pharmacokinetics is the study of drug disposition in the body and focuses on the changes in **drug plasma concentration** . For any given drug and dose, the plasma concentration of the drug will rise and fall according to the rates of three processes: **absorption, distribution**, and **elimination.**

Absorption of a drug refers to the movement of drug into the bloodstream, with the rate dependent on the physical characteristics of the drug and its formulation. Distribution of a drug refers to the process of a drug leaving the bloodstream and going into the organs and tissues. Elimination of a drug from the blood relies on two processes: **biotransformation (metabolism)** of a drug to one or more metabolites, primarily in the liver; and the **excretion**
of the parent drug or its metabolites, primarily by the kidneys. The relationship between these processes is shown in Figure.

DRUG ABSORPTION

Drug absorption refers to the **passage of drug molecules** from the site of administration into the circulation. The process of drug absorption applies to all routes of administration, except for the topical route, where drugs are applied directly on the target tissue, and intravenous administration, where the drug is already in the circulation.

Drug absorption requires that drugs cross one or more layers of cells and cell membranes. Drugs injected into the subcutaneous tissue and muscle bypass the epithelial barrier and are more easily absorbed through spaces between capillary endothelial cells. In the gut, lungs, and skin, drugs must first be absorbed through a layer of epithelial cells that have tight junctions. For this reason, drugs face a greater **barrier** to absorption after oral administration than after parenteral administration.

Processes of Absorption

Most drugs are absorbed by **passive diffusion** across a biologic barrier and into the circulation. The rate of absorption is proportional to the drug concentration

gradient across the barrier and the surface area available for absorption at that site, known as **Fick's Law**. Drugs can be absorbed passively through cells either by lipid diffusion or by aqueous diffusion. **Lipid diffusion** is a process in which the drug dissolves in the lipid components of the cell membranes. This process is facilitated by a high degree of lipid solubility of the drug. **Aqueous diffusion** occurs by passage through aqueous pores in cell membranes. Because aqueous diffusion is restricted to drugs with low molecular weights, many drugs are too large to be absorbed by this process.

A few drugs are absorbed by **active transport** or by **facilitated diffusion.** Active transport requires a **carrier molecule** and a form of **energy**, provided by hydrolysis of the terminal high-energy phosphate bond of ATP. Active transport can transfer drugs against a concentration gradient. For example, the antineoplastic drug, **5-fluorouracil**, undergoes active transport. Facilitated diffusion also requires a carrier molecule, but no energy is needed. Thus drugs or substances cannot be transferred against a concentration gradient but diffuse faster than without a carrier molecule present. Some cephalosporin antibiotics, such as **cephalexin**, undergo facilitated diffusion by an oligopeptide transporter protein located in intestinal epithelial cells.

Effect of pH on Absorption of Weak Acids and Bases

Many drugs are weak acids or bases that exist in both ionized and non-ionized forms in the body. Only the **non-ionized form** of these drugs is sufficiently soluble in membrane lipids to cross cell membranes. The ratio of the two forms at a particular site influences the **rate of absorption** and is also a factor in distribution and elimination.

The protonated form of a weak acid is non-ionized, whereas the protonated form of a weak base is ionized. The ratio of the protonated form to the nonprotonated form of these drugs can be calculated using the **Henderson- Hasselbalch equation**.

The pKa is the negative log of the ionization constant, particular for each acidic or basic drug. At a pH equal to the pKa , **equal** amounts of the protonated and nonprotonated forms are present. If the pH is less than the pKa, the protonated

form predominates. If the pH is greater than the pKa, the nonprotonated form predominates.

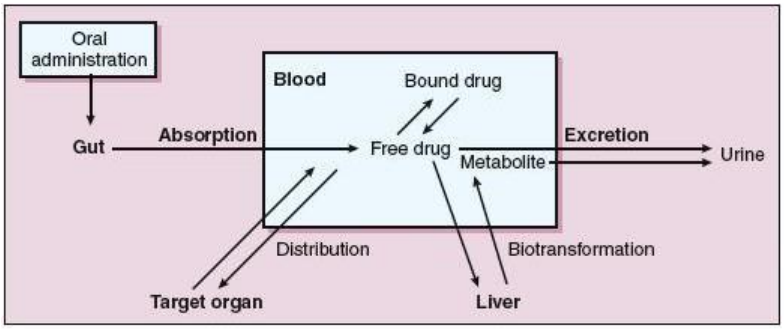


Figure 2–1. The absorption, distribution, biotransformation (metabolism), and excretion of a typical drug after its oral administration.

In the stomach, with a pH of 1, weak acids and bases are highly protonated. At this site, the non-ionized form of weak acids (pK a = 3 – 5) and the ionized form of weak bases (pK a = 8 – 10) will predominate. Hence, weak acids are more readily absorbed from the stomach than are weak bases. In the intestines, with a pH of 7, weak bases are also mostly ionized, but much less so than in the stomach, and weak bases are absorbed more readily from the intestines than from the stomach. However, weak acids can also be absorbed more readily from the intestines than from the stomach, despite their greater ionization in the intestines, because the intestines have a greater surface area than the stomach for absorption of the non-ionized form of a drug, and this outweighs the influence of greater ionization in the intestines.

DRUG DISTRIBUTION

Drugs are distributed to organs and tissues via the circulation, diffusing into interstitial fluid and cells from the circulation. Most drugs are not uniformly distributed throughout total body water, and some drugs are restricted to the extracellular fluid or plasma compartment. Drugs with sufficient lipid solubility can simply diffuse through membranes into cells. Other drugs are concentrated in cells by the phenomenon of **ion trapping**, which is described further below.

Drugs can also be actively transported into cells. For example, some drugs are actively transported into hepatic cells, where they may undergo enzymatic biotransformation. In the intestines, drug transport by **P-glycoprotein (Pgp)** in the blood-to-lumen direction leads to a secretion of various drugs into the intestinal

tract, thereby serving as a detoxifying mechanism. The Pgp proteins also remove many drugs from tissues throughout the body, including anticancer agents. Inhibition of Pgp by amiodarone, erythromycin, propranolol, and other agents can increase tissue levels of these drugs and augment their pharmacologic effects.

Factors Affecting Distribution

The rate at which a drug is distributed to various organs after a drug dose is administered depends largely on the proportion of **cardiac output** received by the organs. Drugs are rapidly distributed to highly perfused tissues, namely the brain, heart, liver, and kidney, and this enables a rapid onset of action of drugs affecting these tissues. Drugs are distributed more slowly to less perfused tissues such as skeletal muscle and even more slowly to those with the lowest blood flow, such as skin, bone, and adipose tissue.

Plasma Protein Binding

Almost all drugs are reversibly bound to plasma proteins, primarily **albumin**, but also lipoproteins, glycoproteins, and β globulins. The extent of binding depends on the affinity of a particular drug for protein-binding sites and ranges from less than 10% to as high as 99% of the plasma concentration. As the free (unbound) drug diffuses into interstitial fluid and cells, drug molecules dissociate from plasma proteins to maintain the equilibrium between free drug and bound drug. In general, acid drugs bind to albumin and basic drugs to glycoproteins and β globulins. Plasma protein binding is **saturable** and a drug can be displaced from binding sites by other drugs that have a high affinity for such sites. However, most drugs are not used at high enough plasma concentrations to occupy the vast number of plasma protein binding sites. There are a few agents that may cause drug interactions by competing for plasma protein binding sites.

Molecular Size

Molecular size is a factor affecting the distribution of extremely large molecules, such as those of the anticoagulant **heparin**. Heparin is largely confined to the plasma compartment, although it does undergo some biotransformation in the liver.

Lipid Solubility

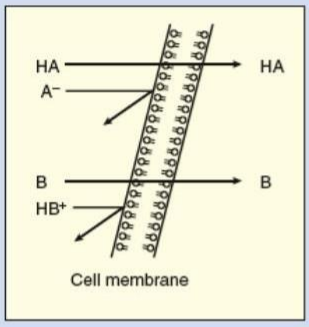
Lipid solubility is a major factor affecting the extent of drug distribution, particularly to the brain, where the **bloodbrain barrier** restricts the penetration of polar and ionized molecules. The barrier is formed by tight junctions between the capillary endothelial cells and also by the glial cells that surround the capillaries, which inhibit the penetration of polar molecules into brain neurons.

BOX 2–1 EFFECT OF pH ON THE ABSORPTION OF A WEAK ACID AND A WEAK BASE

Weak **acids** (HA) donate a proton (H^+) to form anions (A^-), whereas **weak bases** (B) accept a proton to form cations (HB^+).

$HA \rightleftharpoons H^+ + A^-$		For weak acids, the protonated form is nonionized.
$B + H^+ \rightleftharpoons HB^+$		For weak bases, the protonated form is ionized.

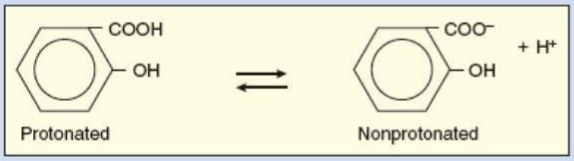
Only the **non-ionized form** of a drug can readily penetrate cell membranes.



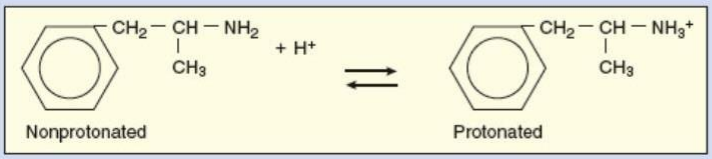
The pK_a of a weak acid or weak base is the **pH** at which there are equal amounts of the protonated form and the nonprotonated form. The **Henderson-Hasselbalch equation** can be used to determine the ratio of the two forms:

$$\log \frac{[\text{protonated form}]}{[\text{Nonrotonated form}]} = pK_a - pH$$

For **salicylic acid**, which is a weak acid with a pK_a of 3, log [HA]/[A^-] is 3 minus the pH. At a pH of 2, then, log [HA]/[A^-] = 3 − 2 = 1. Therefore, [HA]/[A^-] = 10/1.



For **amphetamine**, which is a weak base with a pK_a of 10, log [HB^+]/[B] is 10 minus the pH. At a pH of 8, then, log [HB^+]/[B] = 10 − 8 = 2. Therefore, [HB^+]/[B] = 100/1.



(Continued)

Fig 3.6

DRUG BIOTRANSFORMATION

Drug **biotransformation** and **excretion** are the two processes responsible for the decline of the plasma drug concentration over time. Both of these processes contribute to the **elimination** of active drug from the body, and as discussed later in the chapter, **clearance** is a measure of the rate of elimination. Biotransformation, or **drug metabolism**, is the enzyme- catalyzed conversion of drugs to their metabolites. Most drug biotransformation takes place in the liver, but drug-metabolizing enzymes are found in many other tissues, including the gut, kidneys, brain, lungs, and skin.

Role of Drug Biotransformation

The fundamental role of drug-metabolizing enzymes is to **inactivate and detoxify** drugs and other foreign compounds (xenobiotics) that can harm the body. Drug metabolites are usually more water soluble than is the parent molecule and, therefore, they are more readily excreted by the kidneys. No particular relationship exists between biotransformation and pharmacologic activity. Some drug metabolites are active, whereas others are inactive. Many drug molecules undergo attachment of polar groups, a process called **conjugation,** for more rapid excretion. As a general rule, most conjugated drug metabolites are inactive, but a few exceptions exist.

Formation of Active Metabolites

Many pharmacologically active drugs, such as the sedative hypnotic agent **diazepam** (**VALIUM**), are biotransformed to active metabolites. Some agents, known as **prodrugs**, are administered as inactive compounds and then biotransformed to active metabolites. This type of agent is usually developed because the prodrug is better absorbed than its active metabolite. For example, the antiglaucoma agent **dipivefrin** (**PROPINE**) is a prodrug that is converted to its active metabolite, epinephrine, by corneal enzymes after topical ocular administration. Orally administered prodrugs, such as the antihypertensive agent **enalapril** (**VASOTEC**), are converted to their active metabolite by hepatic enzymes during their first pass through the liver.

First-Pass Biotransformation

Drugs that are absorbed from the gut reach the liver via the hepatic portal vein before entering the systemic circulation.

Many drugs, such as the antihypertensive agent metabolites during their first pass through the gut wall and liver, and have low **bioavailability** after oral administration. This phenomenon is called the **first-pass effect**. Drugs administered by the sublingual or rectal route undergo less first-pass metabolism and have a higher degree of bioavailability than do drugs administered by the oral route.

Phases of Drug Biotransformation

Drug biotransformation can be divided into two phases, each carried out by unique sets of metabolic enzymes. In many cases, phase I enzymatic reactions create or unmask a chemical group required for a phase II reaction. In some cases, however, drugs bypass phase I biotransformation and go directly to phase II. Although some phase I drug metabolites are pharmacologically active, most phase II drug metabolites are inactive.

Phase I Biotransformation

Phase I biotransformation includes oxidative, hydrolytic, and reductive reactions.

OXIDATIVE REACTIONS

Oxidative reactions are the most common type of phase I biotransformation. They are catalyzed by enzymes isolated in the microsomal fraction of liver homogenates (the fraction derived from the endoplasmic reticulum) and by cytoplasmic enzymes.

The **microsomal cytochrome P450 (CYP) monooxygenase system** is a family of enzymes that catalyzes the biotransformation of drugs with a wide range of chemical structures. The microsomal **monooxygenase reaction** requires the following: CYP (a hemoprotein); a flavoprotein that is reduced by nicotinamide adenine dinucleotide phosphate (NADPH), called NADPH CYP reductase; and membrane lipids in which the system is embedded.

In the drug- oxidizing reaction, one atom of oxygen is used to form a hydroxylated metabolite of a drug, whereas the other atom of oxygen forms water when combined with electrons contributed by NADPH. The hydroxylated metabolite may be the end product of the reaction or serve as an intermediate that leads to the formation of another metabolite.

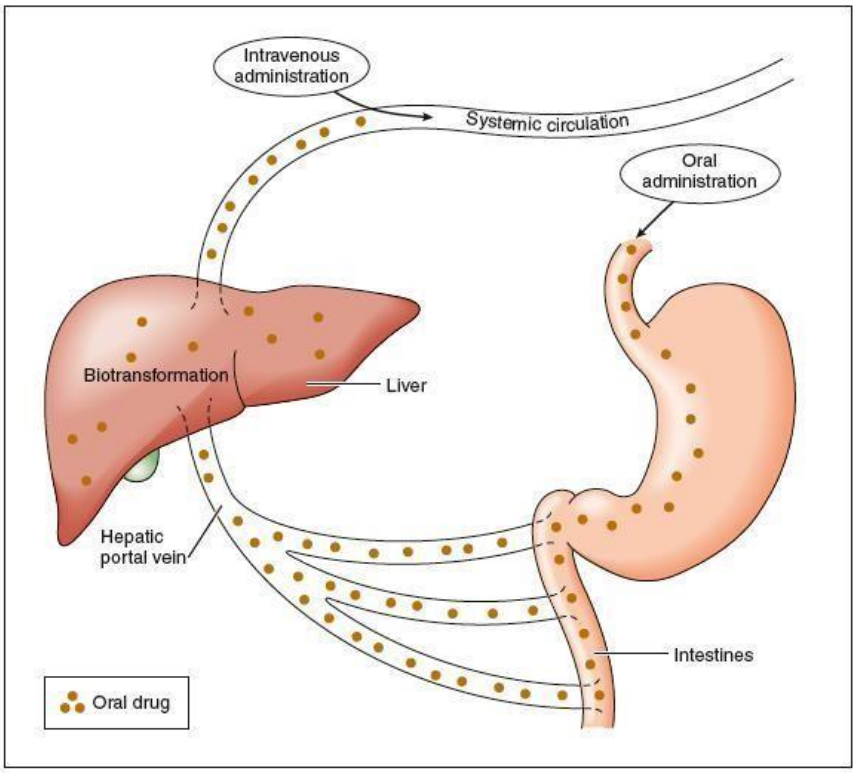


Figure 2–2. First-pass drug biotransformation. Drugs that are absorbed from the gut can be biotransformed by enzymes in the gut wall and liver before reaching the systemic circulation. This process lowers their degree of bioavailability.

The most common chemical reactions catalyzed by CYP enzymes are aliphatic hydroxylation, aromatic hydroxylation, N-dealkylation, and O-dealkylation. Many **CYP isozymes** have been identified and cloned, and their role in metabolizing specific drugs elucidated. Each isozyme catalyzes a different but overlapping spectrum of oxidative reactions. Most drug biotransformation is catalyzed by three CYP families named CYP1, CYP2, and CYP3. The different CYP families are likely related by gene duplication and each family is divided into subfamilies, also clearly related by homologous protein sequences. The **CYP3A** subfamily catalyzes more than half of all microsomal drug oxidations. Many drugs alter drug metabolism by inhibiting or inducing CYP enzymes, and **drug interactions** can occur when these drugs are administered concurrently with other drugs that are metabolized by CYP. Two examples of **inducers of CYP** are the barbiturate, **Phenobarbital**, and the antitubercular drug, **rifampin**. The inducers stimulate the transcription of genes encoding CYP enzymes, resulting in increased messenger RNA (mRNA) and protein synthesis. Drugs that induce CYP enzymes activate the binding of **nuclear receptors** to enhancer domains of

CYP genes, increasing the rate of gene transcription. A few drugs are oxidized by **cytoplasmic enzymes.**

For example, **ethanol** is oxidized to aldehyde by alcohol dehydrogenase, and **caffeine** and the bronchodilator **theophylline** are metabolized by xanthine oxidase. Other cytoplasmic oxidases include **monoamine oxidase**, a site of action for some psychotropic medications.

HYDROLYTIC REACTIONS

Esters and amides are hydrolyzed by a variety of enzymes. These include cholinesterase and other plasma esterases that inactivate choline esters, local anesthetics, and drugs such as **esmolol** (**BREVIBLOC**), an agent for the treatment of tachycardia that blocks cardiac 1-adrenergic receptors. There are few CYP enzymes that carry out hydrolytic reactions.

REDUCTIVE REACTIONS

Reductive reactions are less common than are oxidative and hydrolytic reactions. **Chloramphenicol** , an antimicrobial agent, and a few other drugs are partly metabolized by a hepatic nitro reductase, and this process involves CYP enzymes. **Nitroglycerin**, a vasodilator, undergoes reductive hydrolysis catalyzed by glutathione- organic nitrate reductase.

Phase II Biotransformation

In phase II biotransformation, drug molecules undergo **conjugation reactions** with an endogenous substance such as **acetate**, **glucuronate**, **sulfate**, or **glycine.** Conjugation enzymes, which are present in the liver and other tissues, join various drug molecules with one of these endogenous substances to form water-soluble metabolites that are more easily excreted. Except for microsomal **glucuronosyltransferases** , these enzymes are located in the cytoplasm. Most conjugated drug metabolites are pharmacologically inactive.

GLUCURONIDE FORMATION

Glucuronide formation, the most common conjugation reaction, utilizes glucuronosyltransferases to conjugate a glucuronate molecule with the parent drug molecule.

ACETYLATION Acetylation is accomplished by **N- acetyltransferase** enzymes that utilize acetyl coenzyme A (**acetyl CoA**) as a source of the acetate group.

SULFATION

Sulfotransferases catalyze the conjugation of several drugs, including the vasodilator **minoxidil** and the potassium-sparing diuretic **triamterene**, whose sulfate metabolites are pharmacologically active.

Pharmacogenomics

Since the completion of the human genome, it is now fully realized that there is a great degree of individual variation, called **polymorphism**, in the genes coding for drug- metabolizing enzymes. Modern genetic studies were triggered by rare fatalities in children being treating for leukemia using the thiopurine agent, 6-mercaptopurine (6-MP). It was discovered that the children died as a result of drug toxicity because they expressed a faulty variant of thiopurine methyltransferase, the enzyme that metabolizes 6-MP.

Variations in Acetyltransferase Activity

Individuals exhibit slow or fast acetylation of some drugs because of genetically determined differences in Nacetyltransferase. Slow acetylators (SAs) were first identified by neuropathic effects of **isoniazid**, a drug to treat tuberculosis. These patients had higher plasma levels of isoniazid compared to other patients classified as rapid acetylators (RAs). The SA phenotype is autosomal recessive, although there are more than 20 allelic variants of the gene for N- acetyltransferase identified. In individuals with one wild-type enzyme and one faulty variant, an intermediate phenotype is observed. The **distribution** of these phenotypes varies from population to population. About 15% of Asians, 50% of Caucasians and Africans, and more than 80% of Mideast populations have the SA phenotype. Other drugs that may cause toxicity in the SA patient are **sulfonamide antibiotics**, the antiarrhythmic agent **procainamide**, and the antihypertensive agent **hydralazine.**

Variations in CYP2D6 and CYP2C19 Activity

Variations in oxidation of some drugs have been attributed to genetic differences in certain CYP enzymes. Genetic polymorphisms of CYP2D6 and CYP2C19 enzymes are well characterized and human populations of “extensive metabolizers” and “poor metabolizers” have been identified.

These differences are caused by more than 70 identified variants in the CYP2D6 gene and more than 25 variants of the CYP2C19 genes, resulting from point mutations, deletions, or additions; gene rearrangements, or deletion or duplication of the entire gene. This gives rise to an increase, reduction, or complete loss of

enzyme activity and to different levels of enzyme expression that result in **altered rates** of enzymatic reactions.

Most individuals are extensive metabolizers of CYP2D6 substrates, but 10% of Caucasians and a smaller fraction of Asians and Africans are poor metabolizers of substrates for CYP2D6. Psychiatric patients who are poor metabolizers of CYP2D6 drugs have been found to have a higher rate of adverse drug reactions than do those who are extensive metabolizers because of higher psychotropic drug plasma levels. In addition, poor metabolizers of CYP2D6 drugs have a reduced ability to metabolize **codeine** to morphine sufficiently to obtain adequate pain relief when codeine is administered for analgesia. Poor metabolizers of CYP2C19 substrates have higher plasma levels of proton pump inhibitors, such as **omeprazole** (**PRILOSEC**), whereas some extensive metabolizers of CYP2C19 drugs require larger doses of omeprazole to treat peptic ulcer.

Other Variations in Drug Metabolism Enzymes

About 1 of 3000 individuals exhibits a familial **atypical cholinesterase** that will not metabolize succinylcholine, a neuromuscular blocking agent, at a normal rate. Affected individuals are subject to prolonged apnea after receiving the usual dose of the drug. For this reason, patients should be screened for atypical cholinesterase before receiving succinylcholine. There are many more polymorphisms in both phase I and phase II metabolic enzymes. With more than 30 families of drug-metabolizing enzymes, all with genetic variants, a major development in pharmacotherapy will be the individual tailoring of drug and dose to each patient's genomic identity.

DRUG EXCRETION

Excretion is the removal of drug from body fluids and occurs primarily in the **urine**. Other routes of excretion from the body include in bile, sweat, saliva, tears, feces, breast milk, and exhaled air.

Renal Drug Excretion

Most drugs are excreted in the urine, either as the parent compound or as a drug metabolite. Drugs are handled by the kidneys in the same manner as are endogenous substances, undergoing processes of glomerular filtration, active

tubular secretion, and passive tubular reabsorption. The amount of drug excreted is the sum of the amounts filtered and secreted minus the amount reabsorbed.

Glomerular Filtration

Glomerular filtration is the first step in renal drug excretion. In this process, the free drug enters the renal tubule as a dissolved solute in the plasma filtrate. If a drug has a large fraction bound to plasma proteins, as is the case with the anticoagulant **warfarin**, it will have a low rate of glomerular filtration.

Active Tubular Secretion

Some drugs, particularly weak acids and bases, undergo active tubular secretion by transport systems located primarily in proximal tubular cells. This process is competitively inhibited by other drugs of the same chemical class. For example, the secretion of penicillins and other weak acids is inhibited by **probenecid**, an agent used to treat gout. Active tubular secretion is not affected by plasma protein binding. This is due to the equilibrium of free drug and bound drug such that when free drug is actively transported across the renal tubule, this fraction of free drug is replaced by a fraction that dissociates from plasma proteins.

Passive Tubular Reabsorption

The extent to which a drug undergoes passive reabsorption across renal tubular cells and into the circulation depends on the **lipid solubility** of the drug. Drug biotransformation facilitates drug elimination by forming polar drug metabolites that are not as readily reabsorbed as the less-polar parent molecules. Most nonelectrolytes, including **ethanol**, are passively reabsorbed across tubular cells. Ionized weak acids and bases are not reabsorbed across renal tubular cells, and they are more rapidly excreted in the urine than are non-ionized drugs that undergo passive reabsorption. The proportion of ionized and non-ionized drugs is affected by **renal tubular pH**, which can be manipulated to increase the excretion of a drug after a drug overdose.

Biliary Excretion and Enterohepatic Cycling

Many drugs are excreted in the bile as the parent compound or a drug metabolite. Biliary excretion favors compounds with molecular weights that are higher than

300 and with both polar and lipophilic groups; smaller molecules are excreted only in negligible amounts. Conjugation, particularly with **glucuronate**, increases biliary excretion. Numerous conjugated drug metabolites, including both the glucuronate and sulfate metabolites of steroids, are excreted in the bile.

After the bile empties into the intestines, a fraction of the drug may be reabsorbed into the circulation and eventually return to the liver. This phenomenon is called **enterohepatic cycling**. Excreted conjugated drugs can be hydrolyzed back to the parent drug by intestinal bacteria, and this facilitates the drug's reabsorption. Thus, biliary excretion eliminates substances from the body only to the extent that enterohepatic cycling is incomplete, that is, when some of the excreted drug is not reabsorbed from the intestine

Other Routes of Excretion

Sweat and saliva represent minor routes of excretion for some drugs. In pharmacokinetic studies, saliva measurements are sometimes used because the saliva concentration of a drug often reflects the intracellular concentration of the drug in target tissues.

QUANTITATIVE PHARMACOKINETICS

To derive and use expressions for pharmacokinetic parameters, the first step is to establish a mathematical model that accurately relates the plasma drug concentration to the rates of drug absorption, distribution, and elimination. The **one-compartment model** is the simplest model of drug disposition, but the **two-compartment model** provides a more accurate representation of the pharmacokinetic behavior of many drugs. With the one-compartment model, drug undergoes absorption into the blood according to the rate constant, ka, and elimination from the blood with a rate constant, k e . In the two-compartment model, drugs are absorbed into the central compartment (blood), distributed from the central compartment to the peripheral compartment (the tissues), and eliminated from the central compartment. Regardless of the model used, rate constants can be determined for each process and used to derive expressions for other pharmacokinetic parameters, such as the **elimination half-life (t1/2)** of a drug.

Drug Plasma Concentration Curves

In a standardized **drug plasma concentration curve** over time after oral administration of a typical drug, the Y-axis is a linear scale of drug plasma concentration, often in μ g/mL or mg/L, and the X-axis is a time scale, usually in hours.

Parameters of the plasma drug concentration curve are the **maximum concentration** (**C max)** , the time needed to reach the maximum (**T max)** , the **minimum effective concentration** (**MEC)** , and the **duration of action** . A measure of the total amount of drug during the time course is given by the **area under the curve (AUC)**. These measures are useful for comparing the **bioavailability** of different pharmaceutical formulations or of drugs given by different routes of administration.

Bioavailability (F)

Bioavailability is defined as the **fraction (F)** of the administered dose of a drug that reaches the systemic circulation in an active form. The oral bioavailability of a particular drug is determined by dividing the AUC of an orally administered dose of the drug (AUC oral) by the AUC of an intravenously administered dose of the same drug (AUC IV). By definition, an intravenously administered drug has 100% bioavailability. The bioavailability of drugs administered intramuscularly or via other routes can be determined in the same manner as the bioavailability of drugs administered orally.

The bioavailability of orally administered drugs is of particular concern because it can be reduced by many pharmaceutical and biologic factors. Pharmaceutical factors include the rate and extent of tablet disintegration and drug dissolution.

Biologic factors include the effects of food, which can sequester or inactivate a drug; the effects of gastric acid, which can inactivate a drug; and the effects of gut and liver enzymes, which can metabolize a drug during its absorption and first pass through the liver. The CYP3A4 isozyme found in intestinal enterocytes and hepatic cells is a particularly important catalyst of first-pass drug metabolism. CYP3A4 works in conjunction with Pgp as the 3A4 isozyme located in enterocytes inactivates drugs transported into the intestinal lumen by Pgp.

Volume of Distribution

The volume of distribution (**Vd**) is defined as the volume of fluid in which a drug would need to be dissolved to have the **same concentration** as it does in plasma.

The V d does not represent the volume in a particular body fluid compartment, instead, it is an apparent volume that represents the relationship between the dose of a drug and the resulting plasma concentration of the drug.

Calculation of V d

After intravenous drug administration, the plasma drug concentration falls rapidly at first, as the drug is distributed from the central compartment to the peripheral compartment. The V d is calculated by dividing the dose of a drug given intravenously by the plasma drug concentration immediately after the distribution phase (α). As shown in Figure 2 – 9C, this drug concentration can be determined by extrapolating the plasma drug concentration back to time zero from the linear part of the elimination phase (β). Note that the Y-axis in this case is plotted on a **log scale** so that the exponential elimination phase is converted to a straight line. The plasma drug concentration at time zero (**C 0**) represents the plasma concentration of a drug that would be obtained if it were instantaneously dissolved in its V d. The equation for calculating V d is rearranged to determine the dose of a drug that is required to establish a specified plasma drug concentration.

Interpretation of V d

Although the V d does not correspond to an actual body fluid compartment, it does provide a measure of the extent of distribution of a drug. A low V d that approximates plasma volume or extracellular fluid volume usually indicates that the drug's distribution is restricted to a particular compartment (the plasma or extracellular fluid). The anticoagulant **warfarin** has a V d of about 8 L, which reflects a high degree of plasma protein binding. When the V d of a drug is equivalent to total body water (about 40 L, as occurs with ethanol), this usually indicates that the drug has reached the intracellular fluid as well. Some drugs have a V d that is much larger than total body water. A large V d may indicate that the drug is concentrated intracellularly, with a resulting low concentration in the

plasma. Many weak bases, such as the antidepressant **fluoxetine** (**PROZAC**), have a large V d (40 – 55 L) because of the phenomenon of intracellular **ion trapping.** Weak bases are less ionized within plasma than they are within cells because intracellular fluid usually has a lower pH than extracellular fluid.

After a weak base diffuses into a cell, a larger fraction is ionized in the more acidic intracellular fluid. This restricts its diffusion out of a cell and results in a large Vd. A large V d may also result from sequestration into fat tissue, such as occurs with the antimalarial agent **chloroquine.**

Drug Clearance

Clearance (**Cl**) is the most fundamental expression of drug elimination. It is defined as the volume of body fluid (blood) from which a drug is removed per unit of time. Whereas the clearance of a particular drug is **constant**, it is important to note that the amount of drug contained in the clearance volume will **vary** with the plasma drug concentration.

RENAL CLEARANCE

Renal clearance can be calculated as the renal excretion rate divided by the plasma drug concentration. Drugs that are eliminated primarily by glomerular filtration, with little tubular secretion or reabsorption, will have a renal clearance that is approximately equal to the creatinine clearance, which is normally about 100 mL/min in an adult. A renal drug clearance that is higher than the creatinine clearance indicates that the drug is a substance that undergoes tubular secretion. A renal drug clearance that is lower than the creatinine clearance suggests that the drug is highly bound to plasma proteins or that it undergoes passive reabsorption from the renal tubules.

HEPATIC CLEARANCE

Hepatic clearance is more difficult to determine than renal clearance. This is because hepatic drug elimination includes the biotransformation and biliary excretion of parent compounds. For this reason, hepatic clearance is usually determined by multiplying hepatic blood flow by the arteriovenous drug concentration difference.

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 BIOTECHNOLOGY

UNIT – IV – Computer Aided Drug Design – SBI1310

IV. Computer Aided Drug Design

The drug discovery process

In the fields of medicine, biotechnology and pharmacology despite advances in technology and ever-greater understanding of biological systems, the drug discovery process can too often be a lengthy, expensive, difficult, and inefficient process. Once a compound has shown its value in testing it will begin the process of drug development prior to clinical trials. Peira reduces the research and development costs associated with the drug discovery process.

Drug discovery and design requires the identification of candidates, synthesis, characterisation, screening, and assays for therapeutic efficacy. Peira's expertise lets researchers move away from time-consuming manual procedures, towards greater automation of the research process. Our hardware and software solutions limit bias and increase the reliability of your experiments.

The five **key steps** of the drug discovery process:

Fig 4.1

1. Research & early development

This first phase of the drug development process is basic research. During this phase researchers try to understand the underlying mechanism or cause of a certain disease. Researchers look for new chemical or molecular entities that display promising activity against a particular biological target thought to be important for the disease. Other properties (including safety, toxicity, etc) and metabolic effects of the identified entities in humans are not focused on at this stage.

2. Preclinical research

Preclinical research must be completed before clinical trials (testing in humans) can start. During this stage important feasibility, iterative testing and safety data is collected. The main goal of preclinical study is to determine a product's ultimate safety profile. Products may be new or iterated medical devices, drugs and gene therapy solutions. Each class of product may undergo different types of preclinical research. For instance, drugs may undergo pharmacodynamic, pharmacokinetic, ADME, and toxicity testing through animal testing. Typically, both in vitro and in vivo tests will be performed.

3. Chemical & pharmaceutical development

During the chemical and pharmaceutical development phase the aim is to design a quality product and its manufacturing process to consistently deliver the intended performance of the product. The information and knowledge gained from the studies and manufacturing experience provide scientific understanding to support the establishment of the design space, specifications, and manufacturing controls.

4. Clinical research

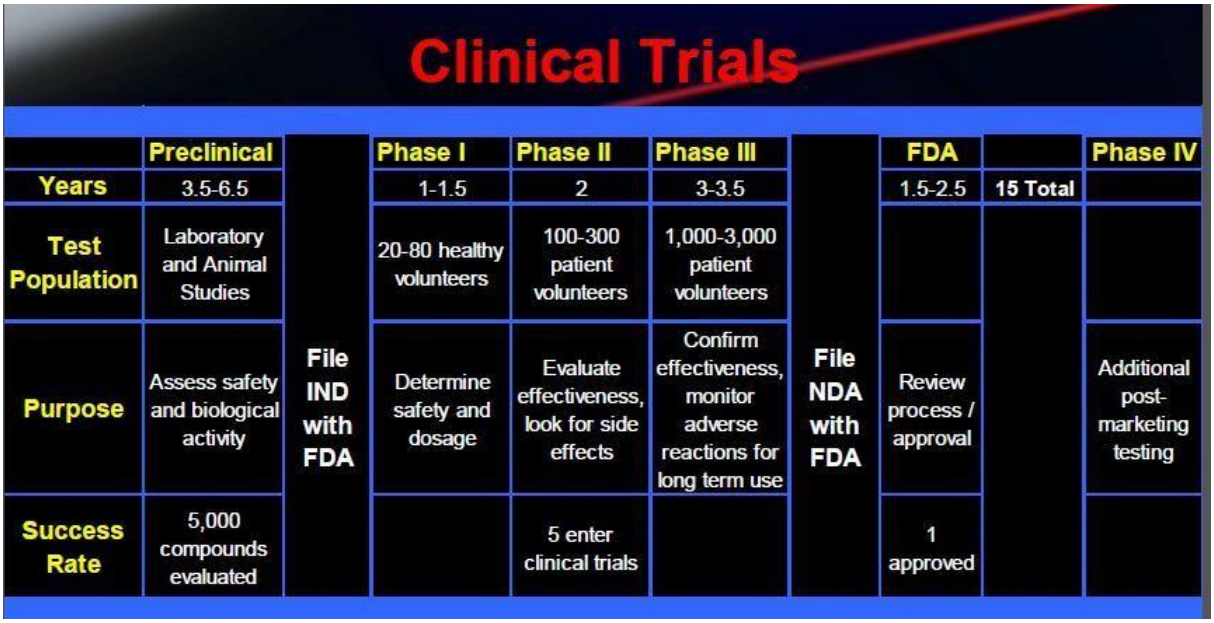
In this phase clinical trials are conducted to collect safety and efficacy data for new drugs. These trials can only take place once there is adequate information about the quality of the product, its non-clinical safety and once health authority approval has been granted. As positive safety and efficacy data are gathered, the number of patients is typically increased.

Clinical trials can vary in size from a single centre in one country to multi-centre trials in multiple countries. After conclusion of the clinical trials the drug will be submitted for regulatory approval, for example with the Food and Drug Administration in the US or with the European Medicines Agency.

5. Chemical & pharmaceutical production

Once a new drug has been approved by the regulatory agencies such as the Food and Drug Administration in the US a full scale manufacturing plant will be built based on the scientific understanding gathered during the chemical and pharmaceutical development phase.

Table 4.1



LEAD DISCOVERY

Approaches to Searching for Hits

- Traditional Library Screening
- Fragment-Based Screening
- Virtual Screening

Filtering Hits to Leads

3. Pharmacodynamics and Pharmacokinetics
4. Biological Assays
5. Lipinski's Rules and Related Indices
6. Final Concerns for Promotion of a Hit to a Lead

Special Cases

7. Serendipity
8. Clinical Observations
9. Natural Products

Once a target, normally an enzyme or receptor, has been established and an assay for activity has been developed, the medicinal chemistry team must discover, find, and make compounds that interact with the target. Through the screening process, some compounds emerge with sufficient activity to warrant further investigation. The active compounds are then examined against a number of criteria, including complexity and anticipated pharmacokinetic behavior. Compounds that satisfy the selection criteria are called leads and advanced for further optimization of activity, selectivity, and biological behavior. Occasionally, leads are found through other methods, such as serendipity or clinical observations. This chapter describes techniques of discovering active compounds through screening and selecting the most promising compounds as leads. The overall process is collectively known as lead discovery.

Approaches to Searching for Hits

The most common tool for discovering hits is library screening. The library may consist of traditional compounds with potentially high activity molecules, smaller fragments of less activity, or even virtual molecules tested through molecular modeling simulations.

Traditional Library Screening

The goal of screening of a library, in whole or in part, is to discover compounds with modest activity against a target. The active compounds discovered through a screen are called *hits*. The threshold for activity varies based on the target, but hit-level activity is typically 1 mM or lower. Targets are normally enzymes or receptors, so the term *activity* refers to an *IC*50 or *EC*50 value.

In-House Libraries

As medicinal chemists synthesize molecules during their day-to-day research, small samples of new compounds are submitted for inclusion in the company's compound library. Over the course of years and decades of research, a compound library steadily grows. A library will reflect the areas of research that have contributed to the collection. A company that has historically been strong in researching b-lactam antibiotics would have a very different library from a company with strength in estrogen receptor binding compounds. Once combinatorial chemistry became recognized in the 1990s as a method for making large numbers of molecules, most pharmaceutical companies hired teams of chemists to create collections of molecules to augment the company's existing library. Samples from natural sources may also be included in a library.

When pharmaceutical companies merge, their libraries merge as well. Through a library, the purchasing company gains a tangible chemical record of the research of the acquired firm. In 1995, Glaxo acquired Wellcome. An area of strength for Wellcome was antiviral research. Wellcome's products at the time included acyclovir and zidovudine, nucleoside analogues with activity against the herpes simplex and human immunodeficiency viruses, respectively. Today, Glaxo, now operating under the name GlaxoSmithKline (GSK), still maintains a strong presence in treatments for viral infections. One recently developed drug is lamivudine, an anti–human immunodeficiency virus nucleoside analogue.

Compound libraries may be bought and sold individually. After the dissolution of the Soviet Union in 1991, laboratories that had been formerly well funded by the Soviet government suddenly became essentially broke.

As a means of generating funds, some research groups began to sell portions of their in-stock compounds. The samples were readily purchased by Western companies, including the pharmaceutical industry. The value of the compounds depends on the novelty of the structures and their purity. Specs, founded in 1987 and based in Delft, The Netherlands, purchases compounds from all over the world, mostly from academic laboratories. These compounds are added to Specs' existing library and in turn sold to interested companies. A company can search the Specs library and purchase promising compounds or the entire collection. Those compounds become the outright property of the purchasing company. If Specs has a sufficient amount of a given compound, the company will sell samples of each compound many times over. The amount sold for each compound may be only 0.5 to 1.0 mg.

Out-Sourced Libraries

Just as a library can be purchased, a library can also be rented. The owner of the library typically enters into an agreement with a drug company. The drug company pays to access and test the compound in a screen. If the compound eventually results in the discovery of a new drug, the owner of the library may receive a bonus. One company built on this type of business model was Pharmacopeia, Inc., of Cranbury, New Jersey.

Pharmacopeia was founded in the early 1990s by W. Clark Still of Columbia University and Michael H. Wigler of Cold Spring Harbor Laboratory. Still and Wigler were early pioneers in the development of combinatorial chemistry for pharmaceutical development purposes. As of 2007, Pharmacopeia claimed to have used the Still and Wigler techniques to prepare a library of over 7.5 million compounds—a massive number that is far greater than the library of a typical major drug company. With a library of this size, Pharmacopeia entered licensing agreements with drug companies. In each partnership, Pharmacopeia brought a large library for discovering hits and an ability to make additional compounds as needed for lead optimization. In turn, the pharmaceutical company provided expertise in screening compounds, performing clinical trials, and marketing a drug. If a compound provided by Pharmacopeia became a marketed drug, Pharmacopeia shared in the revenues from the drug's sales. Over the years, as its own resources grew, Pharmacopeia shifted its business plan more toward developing drugs independently, without the involvement of an outside pharmaceutical company. In 2008, Pharmacopeia was purchased by Ligand Pharmaceuticals of La Jolla, California. Ligand now owns the full chemical library of Pharmacopeia.

Mycosynthetix of Hillsborough, North Carolina, out-sources a large library of fungal broths. The screening of products from fungi is attractive because the number of different species of fungi is astonishingly large and essentially unexplored. Any biological activity discovered through screening compounds from fungi is almost certainly previously unknown and therefore more easily protected through patents.

Fragment-Based Screening

Fragment libraries are no different than traditional compound libraries except molecules in a fragment library are smaller. Fragments have a molecular weight of only 120 to 250 g/mol. Limiting molecular weight dramatically decreases diversity in the library.

Far fewer molecules are required to sample the molecular space of a 250 MW library than a 400 or 500 MW one. Smaller molecules have fewer potential sites for intermolecular binding than larger molecules. Therefore, small molecules rarely bind as strongly as larger compounds. For example, a hit from a typical combinatorial library may show activity (*K*D, *IC*50, *K*i, *EC*50) at concentrations of 1 mM or lower. In contrast, a hit in a fragment library may be selected with activities of around 1 mM, which is a 1,000-fold difference in activity. Remember that a larger *IC*50 value implies weaker enzyme-inhibitor binding.

By itself, a single fragment with 1 mM binding is not very interesting. However, if multiple fragments are known to bind near the same site on a target, then the fragments can sometimes be connected to form a single strongly binding hit. The key to discovering hits through fragment-based screening requires two steps. First binding fragments must be discovered. Second, the fragments must be properly connected and rescreened to discover a hit. Proper connection of the fragments can be a challenge. The tether between the fragments must be the correct length and placed appropriately. Successful examples of fragment-based hit discovery involve targets with two or more Active site pockets, each of which can accommodate a fragment-sized group. Because fragment connection requires a spatial understanding of how a fragment and target interact, fragment-based screening methods need to have a three-dimensional model of the target. Visual models arise from NMR, x-ray crystallography, and molecular modeling data. In x-ray crystallography, x-ray structures of fragments bound to a target provide both the site and position of binding. With quality structural information to

guide the drug discovery group, determining the ideal linker length and position is a much easier task.

Recent research performed in the laboratory of George Whitesides at Harvard University suggests that the only poor choice for a linker is one that is too short to span the distance between fragment binding sites. Linkers that are longer than necessary simply fold upon themselves to bring the fragments closer for binding a target.

This research implies that initial linkers to tether fragments should be longer rather than shorter. The optimal length can be determined in a subsequent study.

The following three Case Studies demonstrate the use of fragment-based screening to discover leads.

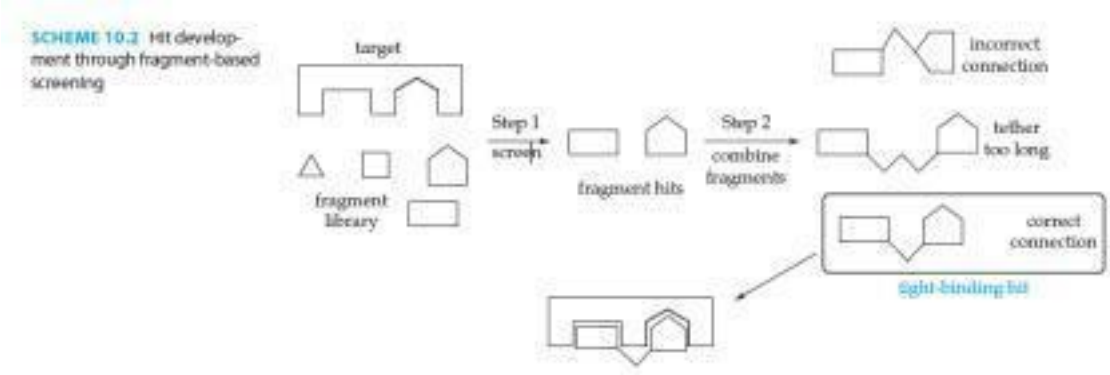


Fig 4.2

Virtual Screening

Virtual screening, sometimes called *in silico screening*, is a relatively new approach to library testing. In a virtual screen, computerized molecular models of both the target and library member are aligned to determine potential complementary intermolecular interactions. Molecules with a high level of complementarity, indicative of potentially strong binding, are flagged for synthesis and testing. A virtual screen requires sufficient knowledge about the target protein structure, likely from x-ray or NMR data. Virtual screening does *not* require an existing compound library. Any molecule imaginable can be modeled in a computer and screened. Of course, the virtual library should consist of realistically synthesizable

compounds. Selected compounds are also normally filtered for those with desired structural elements.

Virtual screening faces a number of challenges. Molecular conformations of both the target and library member are an issue. The target protein will have many low-energy conformations. Some virtual screening methods attempt to accommodate flexibility of the target protein, as complicated as it may be. All in silico approaches try to account for flexibility of the library member, which is a challenge because even small molecules can have many low-energy conformations.

Other factors include tautomers, pH-dependent ionizations, and stereochemistry. Properly handling all these variables is not trivial and quickly complicates the modeling process. Once starting conformations of the target and library members have been established, each library member is virtually brought into contact with the target to determine the likelihood of binding. The process, called *docking*, follows the induced-fit model in which the interacting molecules influence each other's conformations until a minimum energy is reached.

After a compound has been docked to the target, the binding energy is estimated in a process called *scoring*. Standard intermolecular forces, contact forces, dipole interactions, and hydrogen bonding are approximated and totaled. Current scoring methods produce many *false positives*. False positives are compounds with high predicted binding that show little or no activity in validation testing. Often, the number of false positives can be reduced by using several different scoring systems to calculate binding. This approach is called *consensus scoring*. Compounds with high predicted activities from more than one method are selected for further investigation.

High-scoring compounds are then prepared and screened in a biochemical assay. "Preparing" requires either synthesizing or purchasing a sample of the compound. Sample purchase is normally much faster than synthesis. For this reason, compounds in a virtual screen are often limited to structures that can be purchased from any number of commercial suppliers. Databases of available molecules have been prepared for the sole purpose of assisting virtual screening.

If a screening process is successful, a number of hits will be identified. The number of hits

varies depending on the target. It could be as high as 5–10% or as small as 0.1% of the tested library. The cutoff for the required activity of a hit is somewhat arbitrary. The discovery group may select an activity level based on other known active compounds.

The threshold may also be based on the performance of the entire library. For example, the discovery group may count all compounds that are two or three standard deviations more active than the average of the full library.

For a library with normally distributed activity, a cutoff of two standard deviations would give a hit rate of 2.1%.

A cutoff of three standard deviations would give a hit rate of 0.1%. Based on a representative hit rate of 1%, screening a library of 100,000 compounds would generate 1,000 hits. This is too many compounds to follow up each hit individually, so the number of hits needs to be reduced, or *filtered*, to reach a more manageable figure.

Pharmacodynamics and Pharmacokinetics

The most obvious filter would be to select the most potent hits. The threshold for activity of a hit may be 1 to 10 mM. Setting the limit at 100 nM would quickly reduce the number of hits. This approach has its problems.

Activity in a biochemical assay is strictly a measure of how a molecule interacts with a target, that is, pharmacodynamics. Since advancements in biochemical assays have made them the norm, pharmaceutical companies have continually watched compounds with excellent pharmacodynamics fail in clinical trials because of poor pharmacokinetics. Drug companies have now learned to emphasize both pharmacodynamics and pharmacokinetics throughout the lead discovery process. Instead of prioritizing hits based on binding (pharmacodynamics) with a simple activity threshold, initial hits are also screened for a preliminary pharmacokinetic behavior.

Pharmacokinetic properties of a hit in humans can be estimated with cellular assays as well as animal testing. It is important to be able to estimate properties in humans because U.S. Food and Drug Administration (FDA) approval for testing in humans will not have been obtained for hits from a library screen.

Another selection criterion for hits is the structural complexity of the hit. A hit that is advanced in a discovery program must be modified to increase its binding to a suitable level.

If the hit has a complex structure that is difficult to prepare, synthesis of derivatives of the hit will require much time and slow the entire discovery process. Very complex hits are therefore often less attractive for promotion. Biological Assays Preliminary pharmacokinetic behavior can be tested through a number of whole cell assays. Most commercially successful drugs are administered orally, meaning the drug must be able to enter the bloodstream by crossing membranes in the intestines. The most common membrane permeability assay is performed by monitoring the absorption and secretion of a compound by colon carcinoma cells (Caco-2). Diffusion across Caco-2 cell membranes is considered to be a valid model for molecular transport in the small intestines.

Drugs are mostly metabolized by liver enzymes, especially the cytochrome P-450 enzyme family. The ability for cytochrome P-450 enzymes to metabolize a hit is tested with liver microsomes. Liver microsomes consist primarily of endoplasmic reticulum that contains metabolic enzymes. Hits are individually incubated in the presence of the liver microsomes. Monitoring changes in concentrations provides a sense of the rate of metabolism of each hit. Liver microsomes are also used to determine whether the hit inhibits metabolic processes. Hits that inhibit liver metabolism are shunned.

Acceptable hits do not need to show ideal behavior, but problem compounds will be removed from consideration. If all the hits fail initial pharmacokinetic screening, several options are possible. First, the search for hits could start over with screening of a new library. Second, the threshold for selection of hits could be lowered to enlarge the pool of hits, some of which may pass the permeability and metabolic screens. Third, the criteria for passing the Caco-2 and microsome screens may be softened to allow some hits to pass.

Lipinski's Rules and Related Indices Permeability and liver microsome screens are not high throughput. To save time, researchers, have sought simple methods for eliminating compounds that will be poor lead candidates. A common method involves calculated indices. The first and most widely recognized index-based filter was reported by Lipinski in 1997. This filter is called *Lipinski's rules* or the *Rule of 5.*

Lipinski's rules are designed to predict oral availability of compounds that passively diffuse across membranes. Lipinski's rules are based on observations of a database of approximately 2,500 drugs or compounds studied in clinical trials. In general, the compounds could be described structurally with limits on their molecular weight, number of hydrogen-bond donors and acceptors, and lipophilicity (log *P* or clog *P*). Log *P* is an experimental measure of lipophilicity. A higher log *P* value indicates lower water solubility. The form clog *P* (pronounced "see log P") is a computer-estimated version of log *P*. A compound that violates any of Lipinski's rules may not be absorbed well when orally administered.

Drugs that cross membranes by active, facilitated, or other means of transport fall beyond Lipinski's rules. Exceptions include the macrolide antibiotics such as erythromycin. Over time, Lipinski's rules have been criticized as inappropriate for the evaluation of hits and leads. Hits and leads have weaker binding energies than final drugs. The process of optimization increases a lead's binding energy with multiple, successive structural modifications.

These modifications typically increase a molecule's functionality and subsequently raise the molecular weight of the lead. A hit or lead with a molecular weight of 480 may slip under the Lipinski molecular weight requirement, but after going though the optimization process, the molecule may balloon to a molecular weight of 600 or higher. With this logic, Lipinski's rules are perhaps too permissive to be useful as a filter for hits and leads.

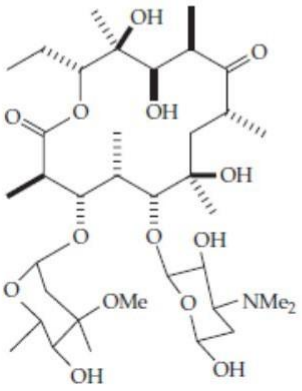
In 1999, Teague distinguished between *lead-like* and *drug-like* hits, and combinatorial library collections. Lead-like hits are characterized as having lower molecular weights 163502, activity 170.1 mM2, and clog *P* values 1632. The lower values give lead-like compounds room to grow into an optimized, high-affinity drug that still satisfies Lipinski's rules. Drug-like hits have higher molecular weights 173502 and clog *P* values 1732 but still modest affinity 1_0.1 mM2. The definition of lead-like has since been used as a preliminary filter for selecting more promising hits from a screen.

Sometimes, simple selection criteria such as Lipinski's rules or lead-like properties are applied to a library before the initial screen is even performed. A recently reported tool for hit evaluation and prioritization is *ligand lipophilicity efficiency* (*LLE*). *LLE* is calculated as the difference between the negative logarithm of a hit's binding affinity, such as -log *IC*50 or -log *K*D, and the logarithm of a hit's partition coefficient, such as log *P* or clog *P*.

TABLE 10.1 Lipinski's rules[19]

1. Molecular weight ≤ 500
2. Lipophilicity (log *P* or clog *P*) ≤ 5
3. Sum of hydrogen-bond donors ≤ 5
4. Sum of hydrogen-bond acceptors ≤ 10

FIGURE 10.11 Erythromycin, a drug that violates Lipinski's rules



erythromycin
10.43

734 g/mol
6 H-bond donors
15 H-bond acceptors

Fig 4.3

$$LLE = -\log IC_{50} - \log P \qquad (10.2)$$

TABLE 10.2 Lead-like and drug-like compounds[20]

Lead-Like	Drug-Like
1. Activity >0.1 μM	1. Activity >0.1 μM
2. MW <350	2. MW >350
3. Clog *P* <3	3. Clog *P* >3

Higher *LLE* values are considered to be better. Consider what this equation says: "*LLE* equals activity less lipophilicity." Without this equation, one might be tempted to say that two hits with the same activity are equally attractive to a drug discovery team. Based on Equation 10.2, the less lipophilic hit (lower log *P*) has a higher *LLE* value and would be more attractive as a hit. Although this may not seem to be an earth-shattering conclusion, Equation 10.1 does show the trade-off between activity and lipophilicity when prioritizing hits.22 Equation 10.2 quantitatively relates some of the central ideas behind the guidelines of Lipinski and Teague. An underlying assumption in Equation 10.2 is that growth of a lead into a drug will increase the compound's lipophilicity. A good hit or lead should therefore start with a lower lipophilicity so that the log *P* of the final drug will not surpass Lipinski's magic value of 5.

Other attempts to refine or improve Lipinski's rule set have appeared in the literature. One notable factor for consideration is the number of rotatable bonds in a hit. Increased molecular flexibility can reduce the ability of a molecule to cross a membrane. The maximum number of rotatable bonds has been suggested as 10. The polar surface area, often abbreviated as PSA, of a molecule is another important factor. Polar surface area is tightly correlated to the number of hydrogen-bond donors and acceptors contained in a molecule.

A maximum polar surface area of 140 Å2 or the equivalent of 12 hydrogen-bond donors/ acceptors has been suggested. This is in line with Lipinski's rules.22

Lipinski terms and related indices exclusively predict oral bioavailability. None addresses metabolism concerns. While the formation of unwanted metabolites is difficult to predict, several functional groups have become recognized as common sources of problems. Examples include quinones and hydroquinones, aryl nitro groups, primary aryl amines, and Michael acceptors. Quinones and Michael acceptors are strong electrophiles that tend to react quickly with and deplete glutathione stores in the liver. Aryl nitro compounds are reduced in the body to aryl amines, which are oxidized to electrophilic species with the same problems as quinones and Michael acceptors. Because the liver enzymes perform a large fraction of a body's metabolism of xenobiotics, the liver is the most commonly damaged organ when metabolites are toxic.

Final Concerns for Promotion of a Hit to a Lead

Only a small number of hits remain after various selection criteria have been applied to the initial hit pool. The surviving hits, sometimes called *compounds of interest* or similar, receive additional scrutiny. Each remaining hit undergoes a handful of structural modifications. Modified hits are often called *analogues*. The analogues allow the discovery team to gain a preliminary understanding of the impact of structural changes on the activity of the hit against

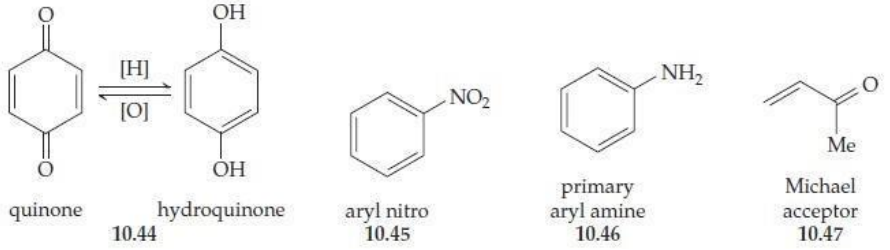


FIGURE 10.12 Common problematic functional groups in drug candidates

its target. If similar targets are known and available, filtered hits are tested for selectivity against the desired target and undesired related targets. These selectivity comparisons can predict the likelihood of side effects. Some in vivo testing may be performed in animals, especially rats. The in vivo tests provide a more accurate and reliable picture of a compound's pharmacokinetic profile. Finally, patent searches are performed to determine the patentability of the hit and later leads that might arise. If a compound cannot be patented, then that compound will certainly not be advanced as a lead.

The outcome of all these selection steps is hopefully one or more leads. Final leads may differ in structure somewhat from their original respective hits. Early structural modifications hopefully generate analogues of higher activity. While hits are often selected at an activity level of 1 mM (*K*D, *IC*50, etc.), structural changes may provide a lead with activity at concentrations of 0.1 mM (100 nM). Ultimately, potency will typically be improved down to the 1–10 nM level during the lead optimization stage.

PHARMACOPHORE

A pharmacophore is an abstract description of molecular features which are necessary for molecular recognition of a ligand by a biological macromolecule. The IUPAC defines a pharmacophore to be "an ensemble of steric and electronic features that is necessary to ensure the optimal supramolecular interactions with a specific biological target and to trigger (or block) its biological response".

A pharmacophore model explains how structurally diverse ligands can bind to a common receptor site. Furthermore pharmacophore models can be used to identify through denovo design or virtual screening novel ligands that will bind to the same receptor

Features

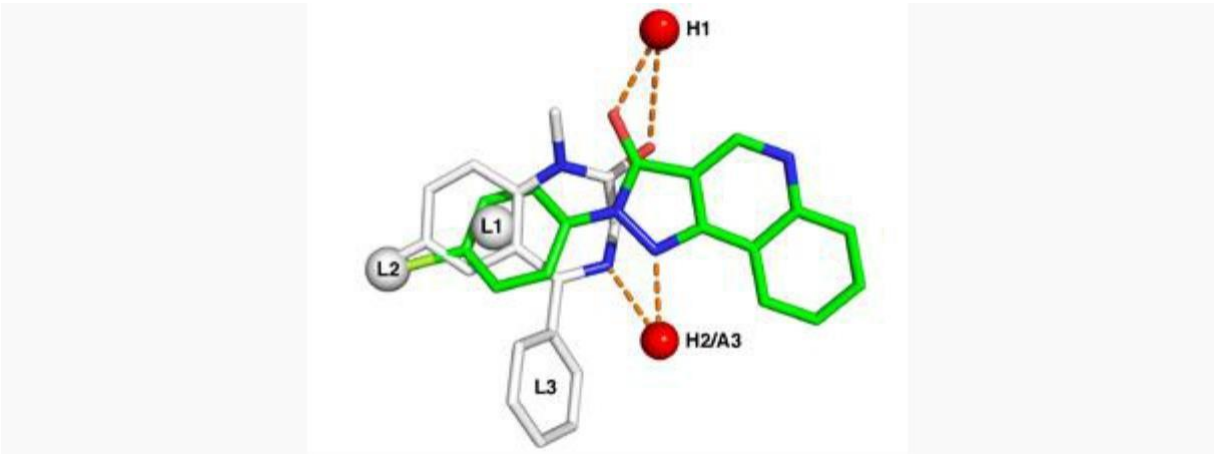


Fig 4.4

An example of a pharmacophore model of the benzodiazepine binding site on the GABAA receptor. White sticks represent the carbon atoms of the benzodiazepine diazepam, while green represents carbon atoms of the nonbenzodiazepine CGS-9896. Red and blue sticks are oxygen and nitrogen atoms that are present in both structures. The red spheres labeled H1 and H2/A3 are, respectively, hydrogen bond donating and accepting sites in the receptor, while L1, L2, and L3 denote lipophilic binding sites.

Typical pharmacophore features include hydrophobic centroids, aromatic rings, hydrogen bond acceptors or donors, cations, and anions. These pharmacophoric points may be located on the ligand itself or may be projected points presumed to be located in the receptor.

The features need to match different chemical groups with similar properties, in order to identify novel ligands. Ligand-receptor interactions are typically “polar positive”, “polar negative” or “hydrophobic”. A well-defined pharmacophore model includes both hydrophobic volumes and hydrogen bond vectors.

Model Development

The process for developing a pharmacophore model generally involves the following steps:

1. Select a training set of ligands – Choose a structurally diverse set of molecules that will be used for developing the pharmacophore model. As a pharmacophore model should be able to discriminate between molecules with and without bioactivity, the set of molecules should include both active and inactive compounds.

2. Conformational analysis – Generate a set of low energy conformations that is likely to contain the bioactive conformation for each of the selected molecules.

3. Molecular superimposition – Superimpose ("fit") all combinations of the low-energy conformations of the molecules. Similar (bioisosteric) functional groups common to all molecules in the set might be fitted (e.g., phenyl rings or carboxylic acid groups). The set of conformations (one conformation from each active molecule) that results in the best fit is presumed to be the active conformation.

4. Abstraction – Transform the superimposed molecules into an abstract representation. For example, superimposed phenyl rings might be referred to more conceptually as an 'aromatic ring' pharmacophore element. Likewise, hydroxy groups could be designated as a 'hydrogen-bond donor/acceptor' pharmacophore element.

5. Validation – A pharmacophore model is a hypothesis accounting for the observed biological activities of a set of molecules that bind to a common biological target. The model is only valid insofar as it is able to account for differences in biological activity of a range of molecules.

As the biological activities of new molecules become available, the pharmacophore model can be updated to further refine it.

Applications

In modern computational chemistry, pharmacophores are used to define the essential features of one or more molecules with the same biological activity. A database of diverse chemical compounds can then be searched for more molecules which share the same features arranged in the same relative orientation. Pharmacophores are also used as the starting point for developing 3D-QSAR models.

Lead

Early drug discovery involves several phases from target identification to preclinical development. The identification of small molecule modulators of protein function and the process of transforming these into high-content lead series are key activities in modern drug discovery. The Hit-to-Lead phase is usually the follow-up of high-throughput screening (HTS). It includes the following steps:

Hit confirmation

The Hit confirmation phase will be performed during several weeks as follows:

- Re-testing: compounds that were found active against the selected target are re-tested using the same assay conditions used during the HTS.

- Dose response curve generation: several compound concentrations are tested using the same assay, an IC50 or EC50 value is then generated. Methods are being developed that may allow the reuse of the compound that generated the hit in the initial HTS step. These molecules are removed from beads and transferred to a microarray for quantitative assessment of binding affinities in a "seamless" approach that could allow for the investigation of more hits and larger libraries.

- Orthogonal testing: Confirmed hits are assayed using a different assay which is usually closer to the target physiological condition or using a different technology.

- Secondary screening: Confirmed hits are tested in a functional assay or in a cellular environment. Membrane permeability is usually a critical parameter.

- Chemical amenability: Medicinal chemists will evaluate compounds according to their synthesis feasibility and other parameters such as up-scaling or costs

- Intellectual property evaluation: Hit compound structures are quickly checked in specialized databases to define patentability

- Biophysical testing: Nuclear magnetic resonance (NMR), Isothermal Titration Calorimetry, dynamic light scattering, surface plasmon resonance, dual polarisation interferometry, microscale thermophoresis (MST) are commonly used to assess whether the compound binds effectively to the target, the stoïchiometry of binding, any associated conformational change and to identify promiscuous inhibitors.

- Hit ranking and clustering: Confirmed hit compounds are then ranked according to the various hit confirmation experiments.

Hit expansion

Following hit confirmation, several compound clusters will be chosen according to their characteristics in the previously defined tests. An Ideal compound cluster will:

- have compound members that exhibit a high affinity towards the target (less than 1 µM)
- Moderate molecular weight and lipophilicity (usually measured as cLogP). Affinity, molecular weight and lipophilicity can be linked in single parameter such as ligand efficiency and lipophilic efficiency to assess druglikeness
- show chemical tractability
- be free of Intellectual property
- not interfere with the P450 enzymes nor with the P-glycoproteins
- not bind to human serum albumin
- be soluble in water(above 100 µM)
- be stable
- have a good druglikeness
- exhibit cell membrane permeability
- show significant biological activity in a cellular assay
- not exhibit cytotoxicity
- not be metabolized rapidly
- show selectivity versus other related targets

The project team will usually select between three and six compound series to be further explored. The next step will allow to test analogous compounds to define Quantitative structure-activity relationship (QSAR). Analogs can be quickly selected from an internal library or purchased from commercially available sources. Medicinal chemists also synthesize related compounds using different methods such as combinatorial chemistry, high-throughput chemistry or more classical organic chemistry synthesis.

Lead optimization phase

The objective of this drug discovery phase is to synthesize lead compounds, new analogs with improved potency, reduced off-target activities, and physiochemical/metabolic properties suggestive of reasonable in vivo pharmacokinetics. This optimization is accomplished through chemical modification of the hit structure, with modifications chosen by employing structure-activity analysis (SAR) as well as structure-based design if structural information about the target is available.

DRUG DESIGNING

The shortcoming of traditional drug discovery; as well as the allure of a more deterministic approach to combating disease has led to the concept of "Rational drug design" (Kuntz 1992). Nobody could design a drug before knowing more about the disease or infectious process than past. For "rational" design, the first necessary step is the identification of a molecular target critical to a disease process or an infectious pathogen. Then the important prerequisite of "drug design" is the determination of the molecular structure of target, which makes sense of the word "rational".

In fact, the validity of "rational" or "structure-based" drug discovery rests largely on a high-resolution target structure of sufficient molecule detail to allow selectivity in the screening of compounds. Simple flowchart for drug designing shown in the figure:

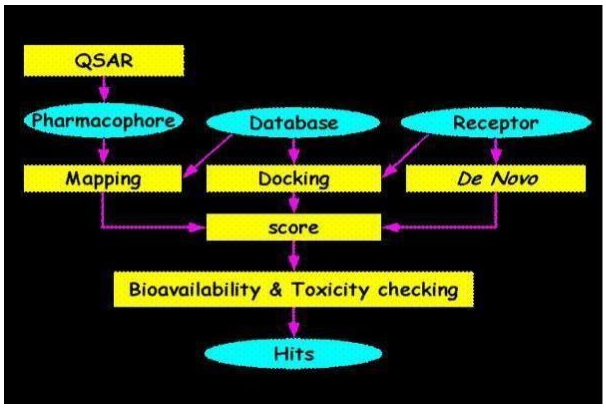


Fig 4.5

Drug designing basically of two types namely ligand based approach or receptor based approach. In both the case the point of centre only differ but requirement of receptor and ligand essential in both the case. By considering these facts, the following steps and online tools are shown below for drug designing.

Drug designing steps were usually divided into steps as follows:
LIGAND PREPARATION
RECEPTOR PREPARATION
DOCKING
BINDING AFFINITY STUDIES

1. LIGAND PREPARATION:

Ligand preparation further divided into different heading namely, ligand retrieval or collection, liand conversion and ligand analysis

a. **Ligands Collection:**

1. DRUGBANK:http://www.drugbank.ca/

The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug (i.e. chemical, pharmacological and pharmaceutical) data with comprehensive drug target (i.e. sequence, structure, and pathway) information.

The database contains nearly 4800 drug entries including >1,480 FDA-approved small molecule drugs, 128 FDA-approved biotech (protein/peptide) drugs, 71 nutraceuticals and >3,200 experimental drugs. Additionally, more than 2,500 non-redundant protein (i.e. drug target) sequences are linked to these FDA approved drug entries. Each DrugCard entry contains more than 100 data fields with half of the information being devoted to drug/chemical data and the other half devoted to drug target or protein data.

2. PUBCHEM: http://pubchem.ncbi.nlm.nih.gov/

PubChem provides information on the biological activities of small molecules. It is a component of NIH's Molecular Libraries Roadmap Initiative.

 CHEMBANK : http://chembank.broad.harvard.edu/welcome.htmIt contains part of an electronic structure collection donated by Tudor Oprea. This set originates from a compilation of ~4.5 million compounds commercially available in August 2002.

These were collected from CDs offered by 10 vendors. The structures were processed into a standardized format using OpenEye's FILTER software (http://www.eyesopen.com/products/applications/filter.html). Compliance with Lipinski's Rule-of-5 was enforced (no violations allowed), and several "undesirable" chemical substructures were removed. A low-value for drug-like scores (scores > 0.2) was implemented in order to further remove chemicals that were very different from the then-accepted medicinal chemistry space.

Approximately ~2.5 million compounds passed these filters, and these were subsequently subjected to diversity selection using D-optimal design and a 2D-based descriptor system (mostly topological indices, atom counts, and LogP-type descriptors), in order to realize the final collection of ~800K compound structures.

4. LIGAND EXPO: http://ligand-expo.rcsb.org/ld-search.html

Ligand Expo (formerly Ligand Depot) provides chemical and structural information about small molecules within the structure entries of the Protein Data Bank. Tools are provided to search the PDB dictionary for chemical components, to identify structure entries containing particular small molecules, and to download the 3D structures of the small molecule components in the PDB entry. A sketch tool is also provided for building new chemical definitions from reported PDB chemical components.

5. Small molecules search by descriptors: http://www.scfbio-iitd.res.in/software/nrdbsm/drugsearch.jsp

NRDBSM database is aimed specifically at virtual high throughput screening of small molecules and their further optimization into successful lead-like candidates. The NRDBSM besides facilitating focused searches in larger databases once a hit is identified should also help in finding a small number of hits for further optimization. A Search engine is available for querying NRDBSM based on the properties mentioned.

b. Ligand conversion (format conversion):

1. Smileconvertor: http://cactus.nci.nih.gov/services/translate/

This tool was used to convert smiles into PDB, SDF, Mol formats and both in 2D and 3D formats.

2. 2Dto3D convertor: http://www.molsoft.com/2dto3d.html

ICM 2D to 3D converter conversion functionality allows construction of icm molecular objects from smiles-strings and creates optimized 3D structures using MMFF atom type assignment and force-field optimization. You can use this page to test convert your chemical structure in smiles-format to 3D and view it using our Java applet-based viewer. A simplified 3-D graphical object will be created at the Molsoft server using the ICM software, and your browser won't have to download too much data.

- CORNIA: http://www.molecular-networks.com/online_demos/index.html This tool was used to generate 3D structures in SDF format.

- CONVERT: http://www.molecular-networks.com/online_demos/convert_demo.html

This online tool might be utilized for the generation of 3d structures in different formats like SDF, Mol, smiles etc.,

5. PRODRG: http://davapc1.bioch.dundee.ac.uk/prodrg/index.html

This WWW PRODRG server will convert coordinates for small molecules in PDB format to the following topology formats: GROMOS, GROMACS, WHAT IF, REFMAC5, CNS, O, SHELX, HEX and MOL2. In addition, coordinates for hydrogen atoms are generated. You can now also sketch your small molecule in a simple text editor, and paste this into the window below. You will be returned all of the above topology files + a GROMOS energy

c. **Ligand Analysis:**

i)Molecular Descriptors:

Molecular descriptors plays crucial role in the drug identification area. So it is essential to know and have the molecular descriptors values regarding ligands. Molecular descriptors predicted through QSAR and QPAR models.

- **MOLINSPIRATION: http://www.molinspiration.com/cgi-bin/properties**

Draw your molecule and press the [Calculate Properties] or [Predict Bioactivity] button.
When the JME input is not working on your computer, try our NEW WebME Ajax editor NEW or paste a raw SMILES here. You may wish to check also our Property Prediction FAQ or more information about calculated properties and drug likeness.

2. **EDRAGAON:** http://www.vcclab.org/lab/edragon/start.html

E-DRAGON is the electronic remote version of the well known software DRAGON, which is an application for the calculation of molecular descriptors developed by the Milano Chemometrics and QSAR Research Group of Prof. R. Todeschini. These descriptors can be used to evaluate molecular structure-activity or structure-property relationships, as well as for similarity analysis and highthroughput screening of molecule databases. DRAGON provides more than 1,600 molecular descriptors that are divided into 20 logical blocks.

The user can calculate not only the simplest atom type, functional group and fragment counts, but also several topological and geometrical descriptors. The first release of DRAGON dates back to 1997. Updates and inclusions of new molecular descriptors are regularly made in order to advance research in QSAR.

3. MOLEDB: http://michem.disat.unimib.it/mole_db/

Molecular descriptors data base was used to get different molecular descriptors for different ligands and drugs which already stored in databases.

4. Model: http://jing.cz3.nus.edu.sg/cgi-bin/model/model.cgi

MODEL - **Molecular Descriptor Lab** for Computing structural and physichemical properties of molecules from their 3D structures.

 Lipinski's rule prediction: http://www.scfbio-iitd.res.in/utility/LipinskiFilters.jsp

Lipinski rule of 5 helps in distinguishing between drug like and non drug like molecules. It predicts high probability of success or failure due to drug likeness for molecules complying with 2 or more of the following rules:

Molecular mass less than 500 Dalton
High lipophilicity (expressed as LogP less than 5)
Less than 5 hydrogen bond donors
Less than 10 hydrogen bond acceptors
Molar refractivity should be between 40-130.

► **ADME Prediction and Drugliklіness prediction:**

The success of a drug mainly depends upon its ability to enter into the host and not producing any adverse effect on it. These properties were tested by ADME (Absorption, Distribution and Metabolism and Excretion) prediction tools and lead and druglikeness of the chemicals also determined in this phase of designing.

► **ADMETools:**

http://www.simcyp.com/ProductServices/FreeADMETools/

ADME DATABASE: http://modem.ucsd.edu/adme/databases/databases_extend.htm

► **ADMETox online Tool:** http://www.pharma-algorithms.com/webboxes/

4. **Drug Likliness: http://www.molsoft.com/mprop/**

5. **Lead finding :** http://leadfinding.com/

II **RECEPTOR PREPARATION:**

For drug to act, target is necessary which might be of receptor or enzyme or hormone type. Initially, the receptor should be prepared by structure modeling method or it might be download from the structure databases. After modeling or downloading from corresponding sources, it should be prepared properly for docking which might be achieved by using binding site analysis tools and target determination tools.

i) Bindingsite Analysis:

- **CASTp:** http://sts.bioengr.uic.edu/castp/index.php

- **Protein Pocket:** http://sts.bioengr.uic.edu/pni/

- **Protein Cavities:** http://luna.bioc.columbia.edu/honiglab/mark-us/cgi-bin/submit.pl

MEDOCK: http://bioinfo.mc.ntu.edu.tw/medock/step1.html

The MEDock (**M**aximum-**E**ntropy based **Dock**ing) web server is aimed at providing an efficient utility for prediction of ligand binding site. A major distinction in the design of MEDock is that its global search mechanism is based on a novel optimization algorithm that exploits the maximum entropy property of the Gaussian distribution.

5**. ODA :** http://www.molsoft.com/oda.cgi

ODA (Optimal Docking Areas) is a new method to predict protein-protein interaction sites on protein surfaces. It identifies optimal surface patches with the lowest docking desolvation energy values as calculated by atomic solvation parameters (ASP) derived from octanol/water transfer experiments and adjusted for protein-protein docking. The predictor has been benchmarked on 66 non- homologous unbound structures, and the identified interactions points (top 10 ODA hot-spots) are correctly located in 70% of the cases (80% if we disregard NMR structures).

ii) Target Determination:

1. **TarFisDock :** http://www.dddc.ac.cn/tarfisdock/index.php

TarFisDock is a web server for identifying drug targets with docking approach. Given a small molecule which can be drug, drug candidate, natural product, or new synthetic compound, TarFisDock docks it into the protein targets in PDTD (Potential Drug Target Database), and outputs the top 2%, 5% or 10% candidates ranked by the energy score, including their binding conformations and a table of the related target information. The server is freely accessible for anonymous user. And one user's result is protected from being retrieved by another. However users are encouraged to fill in a very simple registration form for better safety and convenience. Now submit your molecular structure (in mol2 format) by clicking

2. **TTD:** http://bidd.nus.edu.sg/group/cjttd/ttd.asp

A database to provide information about the known and explored therapeutic protein and nucleic acid targets, the targeted disease, pathway information and the corresponding drugs/ligands directed at each of these targets. Also included in this database are links to relevant databases that contain information about the function, sequence, 3D structure, ligand binding properties, enzyme nomenclature and related literatures of each target. This database currently contains 1535 targets and 2107 drugs/ligands.

SUPERTARGET SEARCH: http://insilico.charite.de/supertarget/main.html#Home

Binding Target Determination: http://www.bindingdb.org/bind/vsOverview.jsp

II. DOCKING:

After collecting, preparing ligands and receptors, they should be assessed for their interaction ability with docking procedure. There were many docking tools are available online. Docking studied under two heads namely, protein-ligand docking and protein-protein docking.

i) Protein-Ligand Docking

β **Patch dock:** http://bioinfo3d.cs.tau.ac.il/PatchDock/

β **ParDock:** http://www.scfbio-iitd.res.in/dock/pardock.jsp

ParDOCK is an all-atom energy based Monte Carlo,rigid protein ligand docking, implemented in a fully automated, parallel processing mode which predicts the binding mode of the ligand in receptor target site.

ii) Protein-Protein Docking:

1. **CLUSPRO**: http://nrc.bu.edu/cluster/

2.**FIREDOCK:** http://bioinfo3d.cs.tau.ac.il/FireDock/

3. **CLUSPRO:** http://cluspro.bu.edu/~rb/cluspro/login/main.php

Studying protein-protein interactions is crucial for a better understanding of processes such as metabolic control, signal transduction, and gene regulation, whereas the ability to dock small ligands to proteins is the key to rational drug and vaccine design strategies. Both problems become much more difficult if no x-ray structure of the protein is available. Accordingly, our main research areas are (1) the development of efficient protein docking algorithms, (2) docking of small ligands to proteins, primarily for the characterization of binding sites, and (3) homology modeling of proteins.

4. **Vakser Lab:** http://vakser.bioinformatics.ku.edu/resources/gramm/grammx/

This is the Web interface to our current protein docking software made available to the public. This software is different from the original GRAMM, except that both packages use FFT for the global search of the best rigid body conformations.

5. **3DGarden:** http://www.sbg.bio.ic.ac.uk/3dgarden/index.cgi

3DGarden is an integrated software suite for performing protein-protein docking. For any pair of protein structures specified by the user, 3DGarden's primary function is to generate an ensemble of putative complexed structures and rank them. The highest-ranking candidates constitute predictions for the structure of the complex. 3DGarden cannot be used to decide whether or not a particular pair of proteins interacts. 3DGarden cannot currently be used for docking DNA/RNA structures with proteins or with other DNA/RNA.

III. BINDING AFFINITY STUDIES:

1. DRUGSCORE: http://pc1664.pharmazie.uni-marburg.de/drugscore/

DrugScore^{ONLINE} is a web-based user interface for the knowledge-based scoring functions DrugScore^{CSD} and DrugScore^{PDB}. DrugScore^{ONLINE} enables you to score protein-ligand complexes of your interest and to visualize the per-atom score contributions as illustrated in the figures shown below. Blue spheres denote favorable interactions whereas red spheres stand for disfavorable ones. The sizes of the spheres correlate with the contributing per-atom scores.

2. BAPPL : http://www.scfbio-iitd.res.in/software/drugdesign/bappl.jsp

Binding Affinity Prediction of Protein-Ligand (BAPPL) server computes the binding free energy of a non-metallo protein-ligand complex using an all atom energy based empirical scoring function.

3.AFFINITY DB: http://www.agklebe.de/affinity

AffinDB is a database of affinity data for structurally resolved protein–ligand complexes from the Protein Data Bank (PDB). It is freely accessible at http://www.agklebe.de/affinity. Affinity data are collected from the scientific literature, both from primary sources describing the original experimental work of affinity determination and from secondary references which report affinity values determined by others. AffinDB currently contains over 730 affinity entries covering more than 450 different protein–ligand complexes. Besides the affinity value, PDB summary information and additional data are provided, including the experimental conditions of the affinity measurement (if available in the corresponding reference); 2D drawing, SMILES code and molecular weight of the ligand; links to other databases, and bibliographic information. AffinDB can be queried by PDB code or by any combination of affinity range, temperature and pH value of the measurement, ligand molecular weight, and publication data (author, journal and year).

Search results can be saved as tabular reports in text files. The database is supposed to be a valuable resource for researchers interested in biomolecular recognition and the development of tools for correlating structural data with affinities, as needed, for example, in structure-based drug design.

4. **LIGAND-PROTEIN DB**:
5. http://lpdb.chem.lsa.umich.edu/
6. **LIGAND-PROTEIN DB**:
7. http://lpdb.chem.lsa.umich.edu

In computational structure-based drug design, the scoring functions are the cornerstones to the success of design/discovery. Many approaches have been explored to improve their reliability and accuracy, leading to three families of scoring functions: force-field-based, knowledge-based, and empirical. The last family is the most widely used in association with docking algorithms because of its speed, even though such empirical scoring functions produce far too many false positives to be fully reliable. In this work, we describe a World Wide Web accessible database that gathers the structural information from known complexes of the PDB with experimental binding data. This database, the Ligand-Protein DataBase (LPDB), is designed to allow the selection of complexes based on various properties of receptors and ligands for the design and parametrization of new scoring functions or to assess and improve existing ones. Moreover, for each complex, a continuum of ligand positions ranging from the crystallographic position to points on the surface of the protein receptor allows an assessment of the energetic behavior of particular scoring functions.

5. **PEARLS: http://ang.cz3.nus.edu.sg/cgi-bin/prog/rune.pl**

OTHER SITES

DENOVO DOCKING-GWIDD http://gwidd.bioinformatics.ku.edu/home

GWIDD is a comprehensive resource for genome-wide structural modeling of protein-protein interactions. It contains interaction information for multiple organisms. The structures of the participating proteins are modeled or crystallographic coordinates are retrieved, if available, and docked by GRAMM-X. The resource is not restricted to interactions in the GWIDD database - other sequences or structures may be entered at various stages.

CLINICAL TRIALS: http://www.centerwatch.com/ **Screening**

Strategy for Lead Optimization Introduction

The goal of small molecule drug discovery for pharmaceutical companies is to turn an organic compound into a highly valued drug candidate in a timely fashion. This is an extremely complex endeavor with numerous hurdles that have to be overcome. The late stage attrition of drug candidates in development and beyond is costly, therefore such failures must be kept to a minimum by setting in place a rigorous assessment of drug candidates at the earliest stage of drug discovery: Lead optimization. During this stage, the initial hits are optimized into lead series by knowledge-driven decisions. It is important to increase the knowledge applied to the design of compounds through each synthesis cycle by providing early, complete, and parallel structure-activity-relationship (SAR) data.

Lead Optimization Process

During lead optimization, there are many unanticipated scientific, medical, and business challenges to every drug discovery program. Therefore, we must increase our understanding of the properties of each leading compound and set up criteria for selection of viable leading series.

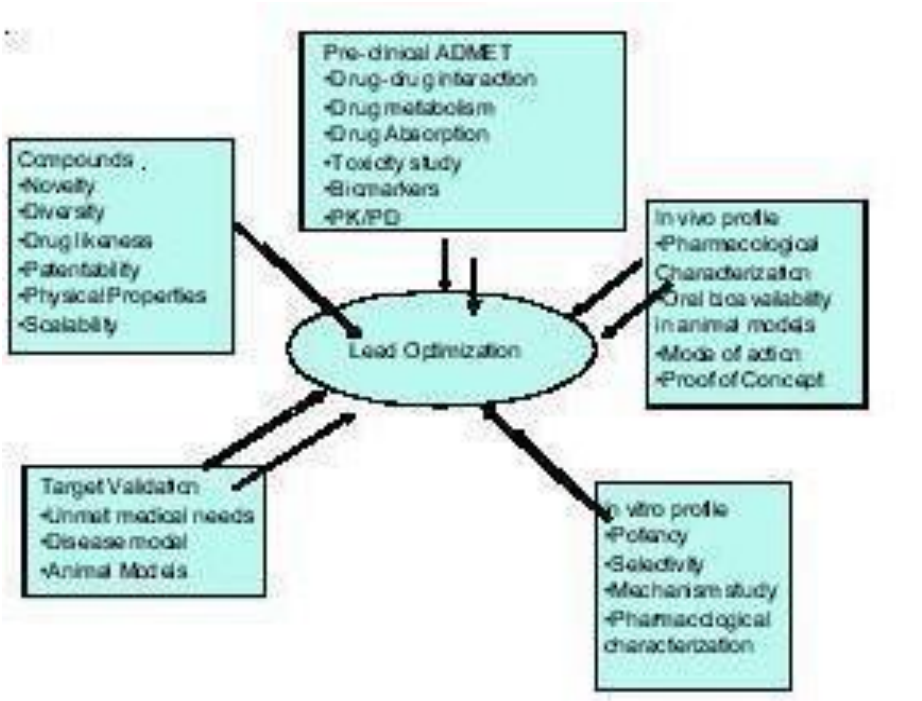


Figure 1. Criteria for selection of lead compounds

An Integrated Screening Approach for Lead Optimization

The screening challenges for lead optimization are significantly different from the hit identification. In hit identification, one bioassay tests many hundreds of thousands of compounds in a short space of time. During lead optimization, smaller numbers of compounds, typically 10 to 100, are put through an array of assays over a longer time-frame. We commonly modify in vitro screening paradigms to respond and solve issues that occur during the lead optimization. A series of advances in bioassay technologies, automation, and miniaturization have enabled the efficiency of hit identification during the past decade. The assays that are used for lead optimization demands connectivity between in vitro and in vivo studies. Therefore, the factors that dictate the selection of the assays for hit identification and lead optimization are different. Results from lead optimization assays need to direct chemists and biologists to make more informed decisions on which compound to synthesize and which compound to move into in vivo efficacy tests. Providing a full array of safety and selectivity profiles on each and every project compound, as quickly as possible after synthesis or discovery, allow scientists to rank and compare compounds against a multiparameter matrix, not just potency. In our experiences, the application of appropriate HTS technologies and selection of suitable biologically relevant in vitro assays for lead optimization will enable project teams to track a full package of information throughout the life of drug discovery programs. Here, we propose an integrated screening approach for lead optimization, which is an efficient process from compound preparation, reagent selection, assay execution, and data reporting for each drug discovery program (Figure 2).

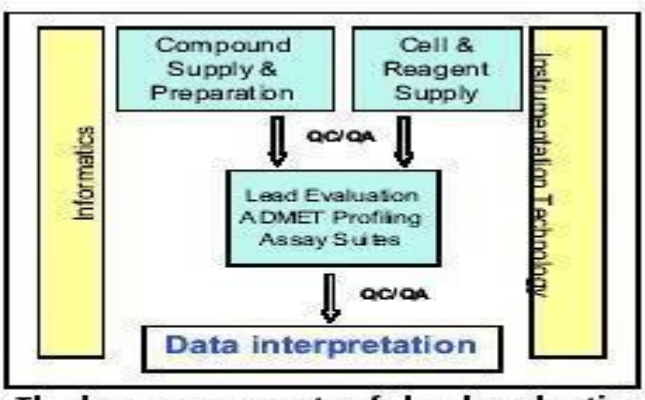


Figure 2. The key components of lead evaluation and lead profiling process: This integrated process allows rapidly delivering high quality, multiparameter data on every compound produced within a discovery program

Lead Evaluation and Lead Profiling

The in vitro bioassays that are used in lead optimization can be divided into lead evaluation and lead profiling assays (see Figure 2). The lead evaluation assays assess potency and selectivity for a given lead compound. The assays used for lead profiling is to provide ADMET liability data that gives the researcher confidence in chemical series as they progress toward drug development. Together, these data direct a project team to identify the most promising preclinical drug candidates.

Lead Evaluation Assays

Lead evaluation assays for lead optimization use both cellfree and cell-based assays. Leading compounds with balanced property profiles will be chosen for further in vivo tests. Here, we refer a group of lead evaluation assays as an assay panel. The assay panel screening approach is determined by 1) target class families, 2) relevant in vitro assays to reflect disease models, and 3) animal models. In this paper, we will exemplify the assay panel screening strategies for target families including G-couple receptors (GPCR), nuclear hormone receptors (NHR), and kinases.

- GPCRs mediate the majority of transmembrane signal transduction in living cells and many marketed drugs target these receptors. About 600 GPCR genes [3] have been identified from genomic sequencing, and 200 GPCRs have known ligands.Three criteria are critical to

  determine the screening strategy for this target family - Ligand Specificity: when molecules are developed against one GPCR, they often show a cross-reactivity with other members of the same family. For example, there are 17 known chemokine receptors. Some chemokine receptors are closely related, i.e., CR2/CCR5 and CCR1/CCR3. Specificity for a selected chemokine target is a common hurdle that needs to be overcome for the design of highly selective antagonists (unpublished internal observation). Consequently, we assay a number of closely related targets concurrently during lead optimization. There are several benefits of using this strategy, including identifying selective leading series, assessing a number of potential drug targets simultaneously, encompassing large range of disease-relevant biology, and serendipitous discovery drug-like compounds for another target.

- Target-Specific Liability: the leading compounds for a given GPCR target may also have cross- reactivity with members of different GPCR families. For example, in our experience, GPCR antagonists often show cross-reactivity with the monoamine GPCR receptors including dopamine, serotonin, and adrenergic receptors. This cross-activity can limit the potential therapeutic use of these antagonists. Therefore, emphasis on the early detection of liability issues related with the monoamine GPCR leads to increased probability of success for drug candidates.

- - GPCR signal transduction pathway: GPCRs are modulated by a three-component system including receptors, G proteins, and downstream effectors. Together, these determine which assay technology will be used for drug discovery. The fluorescent-based and radio-labeled, ligand binding assays have led to the study of GPCR-ligand interactions. These assay techniques have also been enormously useful in defining basic properties of GPCR systems. The cell-based technologies including Ca2+, cAMP, and IP3 quantitation, etc. have been used as functional readouts for GPCRs. These assays require GPCRs to be expressed in multiple cellular backgrounds for pharmacological analysis. During lead optimization, we need both cell-free and cell-based GPCR assays because the binding assays are based on the displacement of labeled ligands and can not differentiate between agonists, antagonists, or inverse agonists. For cell-based GPCR assays, we must keep in mind that the receptor expression levels may have a huge impact on the efficacies of partial agonists and inverse agonists. GPCR signaling in vivo is influenced by tissue location and signaling complexes can be cell type specific. A GPCR can interact with different proteins when over expressed in a cell system compared with its interaction partners in native tissue.

- Nuclear hormone receptors (NRs) constitute a large super family of intracellular ligand- dependent transcription factors. The mode of action of nuclear receptors consists of three steps, including: 1) repression, where NRs recruit a co-repressor complex; 2)de-repression, where the NR- corepressor complex binds ligand, resulting in chromatin decondensation, which is believed to be necessary but not sufficient for activation of the target gene

3) transcription, where a second coactivator complex is assembled that is able to establish contact with the basal transcription machinery, resulting in transcription activation of the target gene. This mechanism is not ubiquitous, since some NRs may act as activators without a ligand, whereas others are unable to interact with target gene promoter in the absence of ligands. The panel screening strategy is often defined by the following criteria:

- Ligand Pharmacology: Agonist and antagonists of NRs often describe ligands that either activate or repress transcription in certain tissues. Cell-free binding and Cell based transcription reporter assays are important methods for lead evaluation.
- Gene profile: Depending on the disease indications, discovery programs may require the monitoring of gene profiles in mutiple cell types through transcriptional profiling of compounds.
- Species selectivity: Commonly, there are efficacy differences between rodent and human NRs. Therefore, species selectivity NR panels are necessary to progress leading compounds and to interpret in vivo data.

Protein kinases regulate significant aspects of cell life. There are about 518 identified protein kinases in the human genome. Therefore, it is not surprising that selectivity is one major obstacle for the kinase drug discovery. Several protein kinase inhibitors that have been approved or entered human clinical trials are also not very specific for a single kinase. This leads to the concept that it is probably impossible to design completely selective ATP site-based inhibitors. It is perhaps more practical to design an inhibitor with a preferred kinase profile rather than a specific kinase inhibition profile. Protein kinase panel screening strategies are critical to the development of protein-kinase inhibitors. The following two factors have impact on establishing the kianse panel for lead optimization.

- Protein kinases are bisubstrate enzymes, binding of ATP to protein kinases will affect binding of protein substrates (or peptides) and vice versa. Therefore, it is important to define the appropriate kinetic parameters and understand the limitations of assay technologies.

.

-Discovery of kinase inhibitors has been focused on three different types of inhibitors; ATP-competitive inhibitors, non-ATP or allosteric inhibitors, and inhibitors that bind inactivated kinases, and therefore the assays need to reflect these mechanistic approaches.

Lead Profiling Assays

The type of lead profiling assays used is very dependent on the stage of the compounds in the drug discovery process. In the early stages of drug discovery there are usually multiple chemotypes, and the in vitro ADMET assays need to be able to assay relatively large numbers of compounds to help drive the early SAR. Under these circumstances, high throughput methods are critical to provide timely progression. Later in the process where single lead series has been identified, the ADMET assays are more focused on predicting in vivo effects and potential clinical liabilities.

Molecular docking

Molecular docking is a well established computational technique which predicts the interaction energy between two molecules. Molecular docking studies are used to determine the interaction of two molecules and to find the best orientation of ligand which would form a complex with overall minimum energy. The small molecule, known as ligand usually fits within protein's cavity which is predicted by the search algorithm. These protein cavities become active when come in contact with any external compounds and are thus called as active sites.

The results are analyzed by a statistical scoring function which converts interacting energy into numerical values called as the docking score; and also the interacting energy is calculated. The 3D pose of the bound ligand can be visualized using different visualizing tools like Pymol, Rasmol etc which could help in inference of the best fit of ligand. Predicting the mode of protein-ligand interaction can assume the active site of the protein molecule and further help in protein annotation. Moreover molecular docking has major application in drug discovery and designing.

This technique mainly incorporates algorithms like molecular dynamics, Monte Carlo stimulation, fragment based search methods.

Different types of Interactions

Interactions between particles can be defined as a consequence of forces between the molecules contained by the particles. These forces are divided into four categories:

• **Electrostatic forces** - Forces with electrostatic origin due to the charges residing in the matter. The most common interactions are charge-charge, charge-dipole and dipole-dipole.

• **Electrodynamics forces**-The most widely known is the Van der Waals interactions.

• **Steric forces** - Steric forces are generated when atoms in different molecules come into very close contact with one another and start affecting the reactivity of each other. The resulting forces can affect chemical reactions and the free energy of a system.

• **Solvent-related forces** - These are forces generated due to chemical reactions between the solvent and the protein or ligand. Examples are Hydrogen bonds (hydrophilic interactions) and hydrophobic interactions.

• A common characteristic of all these forces is their **electromagnetic** nature.

• Other physical factors - **Conformational changes** in the protein and the ligand are often necessary for successful docking.

Molecular docking

Molecular docking can be divided into two separate sections.

1) **Search algorithm** – These algorithms determine all possible optimal conformations for a given complex (protein-protein, protein-ligand) in a environment i.e. the position and orientation of both molecules relative to each other. They can also calculate the energy of the resulting complex and of each individual interaction.

The different types of algorithms that can be used for docking analysis are given below.

- Molecular dynamics
- Monte Carlo methods
- Genetic algorithms
- Fragment-based methods
- Point complementary methods
- Distance geometry methods
- Systematic searches

2) **Scoring function** – These are mathematical methods used to predict the strength of the non-covalent interaction called as binding affinity, between two molecules after they have been docked. Scoring functions have also been developed to predict the strength of other types of intermolecular interactions, for example between two proteins or between protein and DNA or protein and drug. These configurations are evaluated using scoring functions to distinguish the experimental binding modes from all other modes explored through the searching algorithm.

For example:

• Binding Energy

$\Delta G_{bind} = \Delta G_{vdw} + \Delta G_{hbond} + \Delta G_{elect} + \Delta G_{conform} + \Delta G_{tor} + \Delta G_{sol}$

General concept of the algorithm:

1) A 'negative' image of the binding site is made - a collection of spheres of varying radii, each touching the molecular surface at just 2 points.

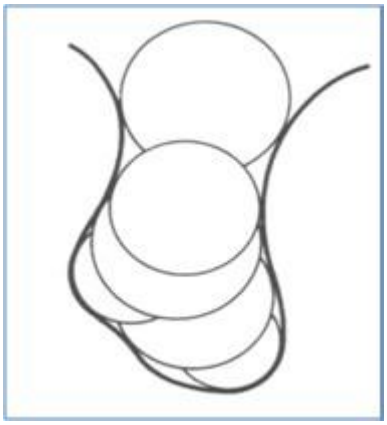


Fig 4.6

2) Ligand atoms are then matched to sphere centers where at least four distances between ligand atoms are matched to sphere center distances.

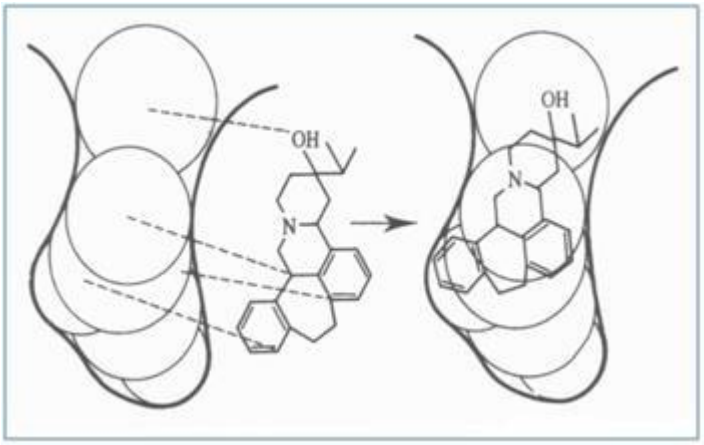


Fig 4.7

3) Proper orientation is achieved by a least squares fit of ligand atoms to the sphere centers.

4) Orientation is checked for any steric clashes between ligand and receptor.

5) If acceptable, then interaction energy is computed as a 'score' for that binding mode

6) New orientations are obtained by matching different sets of atoms and sphere centers

7) Top-scoring orientations are retained for subsequent analysis

Types of Docking -

The following are majorly used type of docking are-

- **Lock and Key or Rigid Docking** – In rigid docking, both the internal geometry of the receptor and ligand is kept fixed during docking.
- **Induced fit or Flexible Docking -** In this model, the ligand is kept flexible and the energy for different conformations of the ligand fitting into the protein is calculated. Though more time consuming, this method can evaluate many different possible conformations which make it more reliable.

Major steps in molecular docking:

Step I – Building the Receptor

In this step the 3D structure of the receptor should be downloaded from PDB; and modified. This should include removal of the water molecules from the cavity, stabilizing charges, filling in the missing residues, generation the side chains etc according to the parameters available. After modification the receptor should be biological active and stable.

Step II – Identification of the Active Site

After the receptor is built, the active site within the receptor should be identified. The receptor may have many active sites but the one of the interest should be selected. Most of the water molecules and heteroatoms if present should be removed.

Step III – Ligand Preparation

Ligands can be obtained from various databases like ZINC, PubChem or can be sketched using tools like Chemsketch. While selecting the ligand, the LIPINSKY'S RULE OF 5 should be applied. The rule is important for drug development where a pharmacologically active lead structure is optimized stepwise for increased activity and selectivity, as well as drug-like properties, as described.

For the selection of a ligand using LIPINSKY'S RULE:

- Not more than 5 –H bond donors.
- Molecular Weight NOT more than 500 Da.
- Log P not over 5 for octanol water partition coefficient.
- NOT more than 10 H bond acceptors.

Step IV- Docking

This is the last step, where the ligand is docked onto the receptor and the interactions are checked. The scoring function generates scores depending on which the ligand with the best fit is selected.

Software available for Molecular Docking:

SCHRODINGER, DOCK, AUTOLOCK TOOLS, DISCOVERY STUDIO, iGemDock

Quantitative structure–activity relationship

Quantitative structure–activity relationship models (**QSAR** models) are regression or classification models used in the chemical and biological sciences and engineering. Like other regression models, QSAR regression models relate a set of "predictor" variables (X) to the potency of the response variable (Y), while classification QSAR models relate the predictor variables to a categorical value of the response variable.

In QSAR modeling, the predictors consist of physico-chemical properties or theoretical molecular descriptors of chemicals; the QSAR response-variable could be abiological activity of the chemicals. QSAR models first summarize a supposed relationship between chemical structures and biological activity in a data-set of chemicals. Second, QSAR models predict the activities of new chemicals.

Related terms include *quantitative structure–property relationships* (*QSPR*) when a chemical property is modeled as the response variable.

As an example, biological activity can be expressed quantitatively as the concentration of a substance required to give a certain biological response. Additionally, when physicochemical properties or structures are expressed by numbers, one can find a mathematical relationship, or quantitative structure-activity relationship, between the two. The mathematical expression, if carefully validated can then be used to predict the modeled response of other chemical structures.

A QSAR has the form of a mathematical model:

- Activity = *f*(physiochemical properties and/or structural properties) + error

The error includes model error (bias) and observational variability, that is, the variability in observations even on a correct model.

$$\text{Activity} = n_1x_1 + n_2x_2 + n_3x_3 + ... + \text{constant}$$

where nx = molecular descriptors

Examples of **biological activity** that can be used for QSAR studies include:

- Enzyme activity
- Minimum effective dose
- Toxicity

Possible **molecular descriptors** that can be used for building QSAR models may include:

- Dipole moment
- Atomic volume
- Number of carbons
- Number of aromatic moieties
- Molar volume
- Wang octanol-water partition coefficient
- Molecular weight
- Quantum chemical descriptors such as molecular orbital energies (HOMO & LUMO) and atomic net charge

Advantages and Disadvantages of QSAR

Advantages of predicting biological activity with quantitative structure-activity relationships modelling include:

- Able to predict activities of a large number of compounds with little to no prior experimental data on activity.
- Can reveal which molecular properties may be worth investigating further.
- Regarded as a "green chemistry" approach since chemical waste is not generated when performing in silico predictions.
- In vivo and in vitro experimentation can be very expensive and time-consuming. QSAR modelling reduces the need for testing on animals and/or on cell cultures and saves time.

Disadvantages of predicting biological activity with QSAR modelling include:

- Does not provide an in-depth insight on the mechanism of biological action.
- Some risk of highly inaccurate predictions of pharmacological or biological activity.

Applications of QSAR in Pharmacology and Medicinal Chemistry

- Quantitative structure-activity relationships (QSAR) can be used during the drug design and drug discovery process. QSAR models can be used as a screening tool to test a large set of compounds or for eliminating test compounds which do not show promise in terms of predicted biological activity.
- Toxicity endpoints of compounds towards organisms can be predicted using QSAR-based methodologies. For instance, the oral rat 50% lethal dose (LD_{50})

Molecular Descriptors used in QSAR

Molecular descriptors can be defined as a numerical representation of chemical information encoded within a molecular structure via mathematical procedure.267 This mathematical representation has to be invariant to the molecule's size and number of atoms to allow model building with statistical methods. The information content of structure descriptors depends on two major factors:

(1) The molecular representation of compounds.

(2) The algorithm which is used for the calculation of the descriptor.

The three major types of parameters initially suggested are, (1) Hydrophobic (2) Electronic (3) Steric

Table 11: Molecular Descriptors used in QSAR

Type	Descriptors
Hydrophobic Parameters	Partition coefficient ; log P
	Hansch's substitution constant; π
	Hydrophobic fragmental constant; f, f'
	Distribution coefficient; log D
	Apparent log P
	Capacity factor in HPLC; log k' , log k'_W
	Solubility parameter; log S
Electronic Parameters	Hammett constant; σ, σ^+, σ^-
	Taft's inductive (polar) constant; σ*
	Swain and Lupton field parameter
	Ionization constant; pK_a, ΔpK_a
	Chemical shifts: IR, NMR
Steric Parameters	Taft's steric parameter; Es
	Molar volume; MV
	Van der waals radius
	Van der waals volume
	Molar refractivity: MR
	Parachor
	Sterimol
Quantum chemical descriptors	Atomic net charge; Q^σ, Q^π
	Superdelocalizability
	Energy of highest occupied molecular orbital; E_{HOMO}
	Energy of lowest unoccupied molecular orbital; E_{LUMO}
Spatial Descriptor	Jurs descriptors, Shadow indices, Radius of Gyration, Principle moment of inertia

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 BIOTECHNOLOGY

UNIT – V – Computational representation of molecules – SBI1310

V. Computational representation of molecules

Chemical database

A **chemical database** is a database specifically designed to store chemical information. This information is about chemical and crystal structures, spectra, reactions and syntheses, and thermophysical data.

Tpes of chemical databases

Chemical structures

Chemical structures are traditionally represented using lines indicating chemical bonds between atoms and drawn on paper (2D structural formulae). While these are ideal visual representations for the chemist, they are unsuitable for computational use and especially for search and storage. Small molecules (also called ligands in drug design applications), are usually represented using lists of atoms and their connections. Large molecules such as proteins are however more compactly represented using the sequences of their amino acid building blocks. Large chemical databases for structures are expected to handle the storage and searching of information on millions of molecules taking terabytes of physical memory...

Literature database

Chemical literature databases correlate structures or other chemical information to relevant references such as academic papers or patents. This type of database includes STN, Scifinder, and Reaxys. Links to literature are also included in many databases that focus on chemical characterization.

Crystallographic database

Crystallographic databases store X-ray crystal structure data. Common examples include Protein Data Bank and Cambridge Structural Database.

NMR spectra database

NMR spectra databases correlate chemical structure with NMR data. These databases often include other characterization data such as FTIR and mass spectrometry.

Reactions database

Most chemical databases store information on stable molecules but in databases for reactions also intermediates and temporarily created unstable molecules are stored. Reaction databases contain information about products, educts, and reaction mechanisms.

Thermophysical database

Thermophysical data are information about phase equilibria including vapor–liquid equilibrium, solubility of gases in liquids, liquids in solids (SLE), heats of mixing, vaporization, and fusion. caloric data like heat capacity, heat of formation and combustion, transport properties like viscosity and thermal conductivity

Chemical structure representation

There are two principal techniques for representing chemical structures in digital databases

- As connection tables / adjacency matrices / lists with additional information on bond (edges) and atom attributes (nodes), such as:
  MDL Molfile, PDB, CML

- As a linear string notation based on depth first or breadth first traversal,
  such as: SMILES/SMARTS, SLN, WLN, InChI

These approaches have been refined to allow representation of stereochemical differences and charges as well as special kinds of bonding such as those seen inorgano-metallic compounds. The principal advantage of a computer representation is the possibility for increased storage and fast, flexible search.

Chemists can search databases using parts of structures, parts of their IUPAC names as well as based on constraints on properties. Chemical databases are particularly different from other general purpose databases in their support for sub-structure search. This kind of search is achieved by looking for subgraph isomorphism (sometimes also called a monomorphism) and is a widely studied application of Graph theory.

The algorithms for searching are computationally intensive, often of O (n^3) or O (n^4) time complexity (where n is the number of atoms involved).

The intensive component of search is called atom-by-atom-searching (ABAS), in which a mapping of the search substructure atoms and bonds with the target molecule is sought. ABAS searching usually makes use of the Ullman algorithm or variations of it (*i.e.* **SMSD**). Speedups are achieved by time amortization, that is, some of the time on search tasks is saved by using precomputed information. This pre-computation typically involves creation of bit strings representing presence or absence of molecular fragments. By looking at the fragments present in a search structure it is possible to eliminate the need for ABAS comparison with target molecules that do not possess the fragments that are present in the search structure. This elimination is called screening (not to be confused with the screening procedures used in drug-discovery). The bit-strings used for these applications are also called structural-keys. The performance of such keys depends on the choice of the fragments used for constructing the keys and the probability of their presence in the database molecules. Another kind of key makes use of hash-codes based on fragments derived computationally. These are called 'fingerprints' although the term is sometimes used synonymously with structural-keys. The amount of memory needed to store these structural-keys and fingerprints can be reduced by 'folding', which is achieved by combining parts of the key using bitwise-operations and thereby reducing the overall length.

Conformation

Search by matching 3D conformation of molecules or by specifying spatial constraints is another feature that is particularly of use in drug design. Searches of this kind can be computationally very expensive. Many approximate methods have been proposed, for instance BCUTS, special function representations, moments of inertia, ray-tracing histograms, maximum distance histograms, shape multipoles to name a few.

Descriptors

All properties of molecules beyond their structure can be split up into either physico-chemical or pharmacological attributes also called descriptors. On top of that, there exist various artificial and more or less standardized naming systems for molecules that supply more or less ambiguous names and synonyms. The IUPAC name is usually a good choice for representing a molecule's structure in a both human-readable and unique string although it becomes unwieldy for larger molecules.

Trivial names on the other hand abound with homonyms and synonyms and are therefore a bad choice as a defining database key. While physico-chemical descriptors like molecular weight, (partial) charge, solubility, etc. can mostly be computed directly based on the molecule's structure, pharmacological descriptors can be derived only indirectly using involved multivariate statistics or experimental (screening, bioassay) results. All of those descriptors can for reasons of computational effort be stored along with the molecule's representation and usually are.

Similarity

There is no single definition of molecular similarity, however the concept may be defined according to the application and is often described as an inverse of a measure of distance in descriptor space. Two molecules might be considered more similar for instance if their difference in molecular weights is lower than when compared with others. A variety of other measures could be combined to produce a multi-variate distance measure. Distance measures are often classified into Euclidean measures and non-Euclidean measures depending on whether the triangle inequality holds. Maximum Common Subgraph (MCS) based substructure search (similarity or distance measure) is also very common. MCS is also used for screening drug like compounds by hitting molecules, which share common subgraph (substructure).

Chemicals in the databases may be clustered into groups of 'similar' molecules based on similarities. Both hierarchical and non-hierarchical clustering approaches can be applied to chemical entities with multiple attributes. These attributes or molecular properties may either be determined empirically or computationally derived descriptors. One of the most popular clustering approaches is the Jarvis-Patrick algorithm.

In pharmacologically oriented chemical repositories, similarity is usually defined in terms of the biological effects of compounds (ADME/tox) that can in turn be semiautomatically inferred from similar combinations of physico-chemical descriptors using QSAR methods.

Registration system

Databases systems for maintaining unique records on chemical compounds are termed as Registration systems. These are often used for chemical indexing, patent systems and industrial databases.

Registration systems usually enforce uniqueness of the chemical represented in the database through the use of unique representations. By applying rules of precedence for the generation of stringified notations, one can obtain unique/'canonical' string representations such as 'canonical SMILES'. Some registration systems such as the CAS system make use of algorithms to generate unique hash codes to achieve the same objective.

A key difference between a registration system and a simple chemical database is the ability to accurately represent that which is known, unknown, and partially known. For example, a chemical database might store a molecule with stereochemistry unspecified, whereas a chemical registry system requires the registrar to specify whether the stereo configuration is unknown, a specific (known) mixture, or racemic. Each of these would be considered a different record in a chemical registry system.

Registration systems also preprocess molecules to avoid considering trivial differences such as differences in halogen ions in chemicals. An example is the Chemical Abstracts Service (CAS) registration system. See also CAS registry number.

PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures

Jae-Min Shin* and Doo-Ho Cho

Research and Development, IDR Tech. B-3003 Tripolis, 210 KumGok-Dong, BunDang-Ku, SungNam-Shi, KyungKi-Do, Republic of Korea 463-805

Received June 30, 2004; Revised and Accepted October 4, 2004

ABSTRACT

PDB-Ligand (http://www.idrtech.com/PDB-Ligand/) is a three-dimensional structure database of small molecular ligands that are bound to larger biomolecules deposited in the Protein Data Bank (PDB). It is also a database tool that allows one to browse, classify, superimpose and visualize these structures. As of May 2004, there are about 4870 types of small molecular ligands, experimentally determined as a complex with protein or DNA in the PDB. The proteins that a given ligand binds are often homologous and present the same binding structure to the ligand. However, there are also many instances wherein a given ligand binds to two or more unrelated proteins, or to the same or homologous protein in different binding environments. PDB-Ligand serves as an interactive structural analysis and clustering tool for all the ligand-binding structures in the PDB. PDB-Ligand also provides an easier way to obtain a number of different structure alignments of many related ligand-binding structures based on a simple and flexible ligand clustering method. PDB-Ligand will be a good resource for both a better interpretation of ligand-binding structures and the development of better scoring functions to be used in many drug discovery applications.

INTRODUCTION

Understanding the interaction between protein and small molecular ligand is very important in post-genomics life science because many important proteins require small molecular ligands or cofactors such as ATP or NAD, in order to function properly. In addition, there is a huge need to design small molecular inhibitors for new drug discovery, based on the analysis of protein–ligand interaction.

The first step for understanding protein–ligand interaction would be to analyze the known protein–ligand complex structures in the Protein Data Bank (PDB) (1) (http://www.rcsb.org/). When analyzing protein–ligand structures, it is often necessary to cluster related ligand-binding structures, according to the ligand conformation, the three-dimensional (3D) ligand-binding structures, and the relative position and orientation of any important residues at the ligand-binding sites.

There are already many protein cluster databases. Protein structure classification databases such as SCOP (2), FSSP (3) and CATH (4) are based on the clustering of the whole 3D structures of protein domains. Other databases such as Pfam (5), Swiss-Model (6) and CDD (7) are primarily based on sequence similarities. With the structural genomics initiatives, these databases have been greatly expanded in size and the structure and function of many experimentally undetermined proteins are now readily inferred using these databases. However, these databases are more focused on the protein structure and function rather than on the structures of ligand or ligand-binding sites. These ligand-binding structures are probably more important in many post-genomics applications such as small molecular inhibitor design for new drug discovery.

There are also many web-based databases of the ligand-binding structure of PDB, including PDBSum (8), Relibase (9) (http://relibase.ebi.ac.uk/), Hic-Up (http://xray.bmc.uu.se/hicup/) and PLD (10). Although these ligand databases provide very useful information on the ligand–protein binding structures, they cannot easily be used to compare or to classify the ligand-binding structures in 3D. Therefore, there is a need for a convenient tool to analyze and classify the ligand-binding structures based on the clustering of the relevant 3D-structures using all the PDB data.

PDB-Ligand is a 3D ligand-binding structure database, derived from the PDB. It is also a database tool that can be used to build such a database and for conveniently browsing through these databases. One novel feature of PDB-Ligand is

that it allows an interactive clustering of ligand-binding structures based on user-specific clustering criteria such as root-mean-square deviation (RMSD) using flexible combinations of the atoms at the ligand-binding sites.

DATABASE CONTENTS AND FEATURES

Figure 1 shows the scheme used in PDB-Ligand database construction. Currently, PDB-Ligand holds 4870 different types of ligands, extracted from 116,019 ligand-binding structures derived from about 25 000 PDB entries. In PDB-Ligand, a ligand-binding structure is defined by the ligands and all the residues and other atoms that are within 6.5 Å around the ligand. Thus, every ligand-binding structure in PDB-Ligand database is surrounded by the residues of the protein, DNA, RNA, solvent or even other ligands. PDB-Ligand uses Chime Plug-in (http://www.mdli.com) as a web-based molecular graphics interface for visualization. It also provides a URL-link to the original PDB file for each ligand-binding structure so that one can easily view the whole ligand–protein structure with other related ligand-binding structures.

One of the most useful features of PDB-Ligand is the interactive clustering of ligand-binding structures, based on the RMSD between different ligand-binding structures. When analyzing the ligand-binding structures for many biologically important ligands such as ATP or FAD, one wants to know how many are in a similar binding environment, and how similar they are in 3D conformation. PDB-Ligand database and its clustering tool allow fast structural classification of the similar ligand-binding structures from all the ligand-binding structures in the PDB. The structure-based clustering feature of PDB-Ligand may be more effectively used with other ligand-binding analysis tools such as LIGPLOT (11) and LPC (12), or with other ligand databases such as Relibase (9) and Ligand-Depot (http://ligand-depot-i.rutgers.edu/).

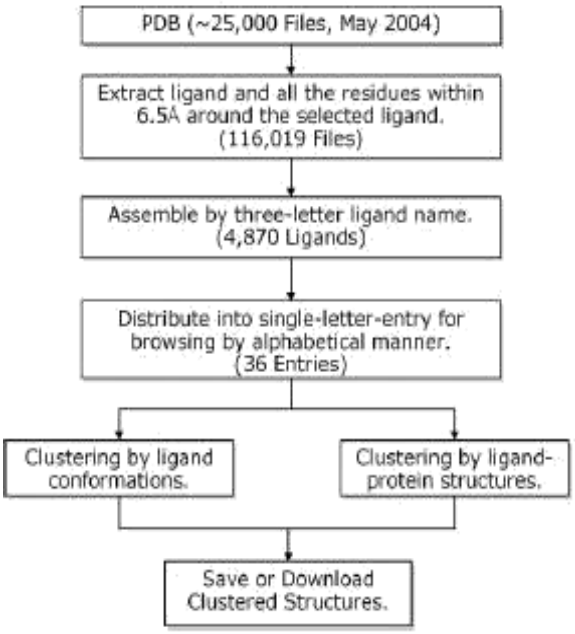


Figure 1. Scheme used in the PDB-Ligand database construction.

In addition, PDB-Ligand allows more flexible clustering based on both the ligand and the protein residues at the ligand-binding sites. This feature is useful, for example, when analyzing the same ligand-binding structures of a structurally related protein family.

CLUSTERING AND STRUCTURE ALIGNMENT

Since PDB-Ligand aims to be a 3D ligand-binding structure database with an interactive clustering feature, it only uses the ligand and the residues within 6.5 Å around the ligand in RMSD-based clustering by structure–structure alignment. Thus, using only the selected number of atoms at the ligand-binding site can greatly speed up the structure-alignment operation for RMSD calculations, while including all the important residues at the ligand-binding sites.

In PDB-Ligand, the clustering of ligand-binding structures is based on the RMSD value between all the corresponding atoms in the ligand after 3D structure superposition by Kabsch method (13). By default, all atoms of the ligand are considered in the superimposition for clustering. Therefore, in this case, every ligand in each cluster will have an overall structural similarity defined by the RMSD cut-off value (default is 0.5 Å).

However, if the ligand shows several different binding modes, it is more important to consider a part of ligand atoms and/or any critical residues at the ligand-binding site in the clustering process. In order to provide users with more convenient atom-selection, PDB-Ligand uses 'copy-and-paste' mechanism, based on chime script utilities (e.g. see E. Martz, http://www.umass.edu/microbio/chime/). For example, if a user selects main chain atoms of the residues at the ligand-binding site in the graphics window, these atoms are listed in the chime-log window, then they can be used in the clustering by 'copy-and-paste' into the selected atoms window. The user can copy any set of atoms shown in this window and paste them into the 'Selected Atoms' window. These atoms are then used to compute the superposition matrix. The simplicity and flexibility of the atom selection mechanism allow the users to perform a more precise clustering of ligand-binding structures. Currently, all the atoms in ligand and protein main chain atoms (N, CA, C and O), are allowed in the clustering.

As a clustering method, a simple greedy algorithm similar to that used by Hobohm and Sander (14) is used. In PDB-Ligand, a reference ligand-binding structure is always the one at the top of the list. Based on a given RMSD cut-off, all the structures similar to the reference structure are clustered together and removed from the list. The clustering is complete if no structure remains in the list.

AN EXAMPLE: ATP-BINDING STRUCTURES

In the current release of PDB-Ligand, there are 321 ATP-binding structures derived from 161 PDB entries. The ATP is the 46th most abundant ligand. If these 321 ATP-binding structures are clustered using 0.5 Å RMSD cut-off, we obtain 165 clusters (see Table 1). It means that there are 165 different conformations of ATP, each one of which is different from all

others, at least, by 0.5 Å in RMSD. If 1.0 Å RMSD is used, we obtain 91 different structural clusters for ATP.

Figure 2 shows a sample cluster of ATP-binding structures using 0.5 Å RMSD cut-off. One can easily see in this figure the common 3D structure of the amino acids surrounding the ligand. Interestingly, based on SCOP 1.65 protein family classification (2), the ATP-binding structures shown in Figure 2 are classified as Actin/Hsp70 protein family. This result may be useful for the users who want to investigate further the ATP-binding structures of such protein family.

FUTURE DIRECTIONS AND APPLICATIONS

The ligand-binding structures in PDB-Ligand will be updated, at least, every four months. In addition, the methods and algorithms for ligand-binding structure clustering will be improved for speed and convenience. Substructure search among ligand structures will also be included in the future. This feature will be useful in analyzing binding structures of various functional groups in many important ligands. Protein sequence and structure information based on the clustered ligand-binding structures will be also useful because it provides more complete information about ligand-binding structures. We also believe that the methods and the strategies used in PDB-Ligand, based on the clustering of ligand-binding structures, will be very useful in many applications for new drug discovery. For an example, based on the classification of similar ligand-binding structures, we have a plan to derive more accurate scoring functions for ligand-docking, virtual screening and lead-optimization for specific target proteins.

AVAILABILITY

PDB-Ligand is freely accessible through the URL at http://www.idrtech.com/PDB-Ligand/.

ACKNOWLEDGEMENTS

We thank B. K. Lee for useful discussions and for valuable suggestions on the manuscript. We also thank H. C. Shin, S. M. Kim, C. K. Han, J. H. Yoon, Y. H. In and N. D. Kim for useful discussions and comments. We also thank M. R. Roh and Y. W. Kim for maintaining the website. This study was supported by a grant of Korean Health 21 R&D Project, Ministry of Health and Welfare, Republic of Korea (Grant ID: 03-PJ2-PG4-BD02-0001).

REFERENCES

1. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
2. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
3. Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–602.
4. Pearl,F.M.G., Lee,D., Bray,J.E., Sillitoe,I., Todd,A.E., Harrison,A.P., Thornton,J.M. and Orengo,C.A. (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res.*, **28**, 277–282.
5. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
6. Kopp,J. and Schwede,T. (2004) The SWISS-MODEL Repository of annotated three-dimensional protein structure homology models. *Nucleic Acids Res.*, **32**, D230–D234.
7. Marchler-Bauer,A., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al.* (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.*, **31**, 383–387.
8. Laskowski,R.A. (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.*, **29**, 221–222.
9. Hendlich,M., Bergner,A., Gunther,J. and Klebe,G. (2003) Relibase—design and development of a database for comprehensive analysis of protein–ligand interactions. *J. Mol. Biol.*, **326**, 607–620.
10. Puvanendrampillai,D. and Mitchell,J.B.O. (2003) Protein Ligand Database (PLD): additional understanding of the nature and specificity of protein–ligand complexes. *Bioinformatics*, **19**, 1856–1857.
11. Wallace,A.C., Laskowski,R.A. and Thornton,J.M. (1995) LIGPLOT: a program to generate schematic diagrams of protein–ligand interactions. *Protein Eng.*, **8**, 127–134.
12. Sobolev,V., Sorokine,A., Prilusky,J., Abola,E.E. and Edelman,M. (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.
13. Kabsch,W. (1978) A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallogr.*, **A34**, 827–828.
14. Hobohm,J. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.

PubChem

PubChem is a database of chemical molecules and their activities against biological assays. The system is maintained by the National Center for Biotechnology Information (NCBI), a component of the National Library of Medicine, which is part of the United States National Institutes of Health (NIH). PubChem can be accessed for free through a web user interface. Millions of compound structures and descriptive datasets can be freely downloaded via FTP. PubChem contains substance descriptions and small molecules with fewer than 1000 atoms and 1000 bonds. More than 80 database vendors contribute to the growing PubChem database.

PubChem

PubChem is designed to provide information on biological activities of small molecules, generally those with molecular weight less than 500 daltons. PubChem's integration with NCBI's Entrez (3) information retrieval system provides sub/structure, similarity structure, bioactivity data as well as links to biological property information in PubMed and NCBI's Protein 3D Structure Resource.

PubChem Databases

PubChem is comprised of three linked
databases -- PubChem Compound,
PubChem
Substance and
PubChem Bioassay

PubChem Compound (unique structures with computed properties)

PubChem Compound is a searchable database of chemical structures with validated chemical depiction information provided to describe substances in PubChem Substance. Structures stored within PubChem Compounds are pre-clustered and cross-referenced by identity and similarity groups. PubChem Compound includes over 5M compounds.

- Molecular Name Searches (e.g., Tylenol, Benzene) allow searching with a variety of chemical synonyms,
- Chemical Property Range Searches (e.g., Molecular Weight between 100 and 200, Hydrogen Bond Acceptor Count between 3 and 5) allow searching for compounds with a variety of physical/chemical properties, and descriptors.

- Simple Elemental Searches (all compounds containing Gallium) allow searching with specific element restrictions.

PubChem Substance (deposited structures)

PubChem Substance is a searchable database containing descriptions of chemical samples, from a variety of sources, and links to PubMed citations, protein 3D structures, and biological screening results available in PubChem BioAssay. PubChem Substance includes over 8M records. Substances with known content are linked to PubChem Compound.

- Molecule Synonym Searches (e.g. all substances with 'deoxythymidine' as a name fragment, or substances that contain 3'-Azido-3'-deoxythymidine).

- Biology Links Search (e.g. substances with tested, active or inactive bioassays).

- Combined Searches (e.g. substances that are 'Active in any BioAssay' and contain the element Ruthenium).

PubChem BioAssay

PubChem BioAssay is a searchable database containing bioactivity screens of chemical substances described in PubChem Substance. PubChem BioAssay includes over 180 bioassays. Searchable descriptions of each bioassay are provided that include descriptions of screening procedural conditions and readouts.

- To Search for BioAssay Data Sets (e.g. HIV growth inhibition).

- To Browse or Download PubChem BioAssay Results (NCI AIDS Antiviral Assay)

PubChem Substance (deposited structures)

PubChem Substance is a searchable database containing descriptions of chemical samples, from a variety of sources, and links to PubMed citations, protein 3D structures, and biological screening results available in PubChem BioAssay. PubChem Substance includes over 8M records. Substances with known content are linked to PubChem Compound.

- Molecule Synonym Searches (e.g. all substances with 'deoxythymidine' as a name fragment, or substances that contain 3'-Azido-3'-deoxythymidine).

- Biology Links Search (e.g. substances with tested, active or inactive bioassays).

- Combined Searches (e.g. substances that are 'Active in any BioAssay' and contain the element Ruthenium).

PubChem BioAssay

PubChem BioAssay is a searchable database containing bioactivity screens of chemical substances described in PubChem Substance. PubChem BioAssay includes over 180 bioassays. Searchable descriptions of each bioassay are provided that include descriptions of screening procedural conditions and readouts.

- To Search for BioAssay Data Sets (e.g. HIV growth inhibition).

- To Browse or Download PubChem BioAssay Results (NCI AIDS Antiviral Assay)

Searching PubChem

PubChem Text Search for searching compound name, synonym or ID that defaults to PubChem Compound. The search results page offers a pull down 'databases' menu that allows searching in PubChem Substance, PubChem BioAssay and a variety of other Entrez databases.

PubChem Chemical Structure Search

PubChem Chemical Structure Search has the following options: Search SMILES (including SMARTS or InChI) or Formula which includes a 'Sketch' link to a drawing program that converts structural diagrams to SMILES(exact), SMARTS(substructure) or InChI(exact) strings for searching.

Clicking 'Done' on the 'structure editor' converts the structural diagram to the appropriate string and transfers it to the search box.

Select Structure File allows importation of standard and common chemical file formats.

Specify Search Type allows restriction to: same compound, similar compounds, formula or substructure.

PubChem Indexes and Index Search

PubChem Indexes and Index Search allows fielded/range searching from either the PubChem homepage or Entrez search page. A extensive list of field aliases and examples of range searching is provided

PubChem Search Results

PubChem Compound

PubChem Compound results are derived from PubChem Substance records that provide structures. Since compounds are structurally unique, one compound may link to multiple substances. The default display is a compound summary with thumbnails with cross links to each PubChem database, other NCBI databases, and depositor's databases.

Clicking either the structure or SID link gives the full display which includes the compound's property data, description, related substance information, neighboring structures, and cross links.

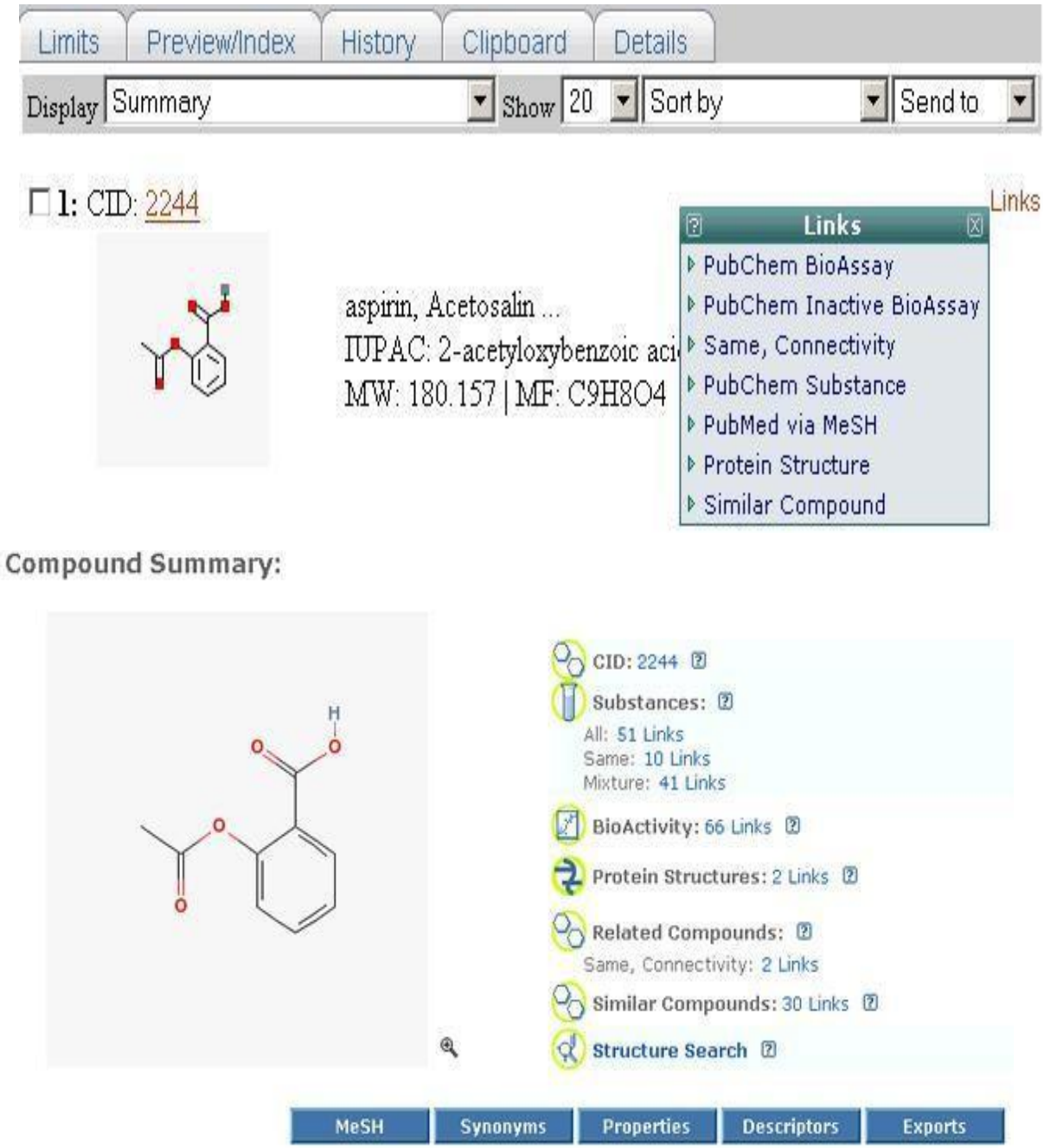

Limits
Preview/Index
History
Clipboard
Details
Display
Summary
Show
20
Sort by
Send to
1: CID: 2244
Links
aspirin, Acetosalin ...
IUPAC: 2-acetyloxybenzoic aci
MW: 180.157 | MF: C9H8O4
Links
PubChem BioAssay
PubChem Inactive BioAssay
Same, Connectivity
PubChem Substance
PubMed via MeSH
Protein Structure
Similar Compound
Compound Summary:
CID: 2244
Substances:
All: 51 Links
Same: 10 Links
Mixture: 41 Links
BioActivity: 66 Links
Protein Structures: 2 Links
Related Compounds:
Same, Connectivity: 2 Links
Similar Compounds: 30 Links
Structure Search
MeSH
Synonyms
Properties
Descriptors
Exports
Medical Subject Annotations: (Total:6)
Display: Next 1 | All
Aspirin
Show MeSH Tree Structure
The prototypical analgesic used in the treatment of mild to moderate pain. It has anti-inflammatory and antipyretic properties and acts as an inhibitor of cyclooxygenase which results in the inhibition of the biosynthesis of prostaglandins. Aspirin also inhibits platelet aggregation and is used in the prevention of arterial and venous thrombosis. (From Martindale, The Extra Pharmacopoeia, 30th ed, p5)

Fig 5.1

PubChem Substance has unique records if the structure is not known or supplied. For example, Sulfated polymannuroguluronate, a novel anti-acquired immune deficiency syndrome (AIDS) drug candidate, and other natural products.

The PubChem Substance Summary Record, is linked to the full record by clicking on the SID number (PubChem's substance identifier). This displays the full substance record, that includes links: to PubMed and the source; the Medical Subject Annotation (MESH Substance Name) and a MESH PubMed search link; and depositor supplied synonyms and comments.

PubChem BioAssay

The PubChem BioAssay Summary Record,is linked to the full record by clicking on the AID number (PubChem's assay (protocol) identifier). This displays the full bioassay record, that includes: links to the substances tested (all, active, inactive, inconclusive) and related PubMed, Protein, Taxonomy, OMIM and related BioAssay records; and a description of the assay possibly with protocols and comments.

1. PubChem Substance has unique records if the structure is not known or supplied. For example, Sulfated polymannuroguluronate, a novel anti-acquired immune deficiency syndrome (AIDS) drug candidate, and other natural products.
2. The PubChem Substance Summary Record, is linked to the full record by clicking on the SID number (PubChem's substance identifier). This displays the full substance record, that includes links: to PubMed and the source; the Medical Subject Annotation (MESH Substance Name) and a MESH PubMed search link; and depositor supplied synonyms and comments.
3. PubChem BioAssay
4. The PubChem BioAssay Summary Record,is linked to the full record by clicking on the AID number (PubChem's assay (protocol) identifier). This displays the full bioassay record, that includes: links to the substances tested (all, active, inactive, inconclusive) and related PubMed, Protein, Taxonomy, OMIM and related BioAssay records; and a description of the assay possibly with protocols and comments.
5. Protein Data Bank
6. The Protein Data Bank (PDB) is a crystallographic database for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallography,NMR spectroscopy, or, increasingly, cryo-electron microscopy, and submitted by biologists and biochemists from around the world, are freely.

PubChem Substance

PubChem Substance has unique records if the structure is not known or supplied. For example, Sulfated polymannuroguluronate, a novel anti-acquired immune deficiency syndrome (AIDS) drug candidate, and other natural products.

The PubChem Substance Summary Record, is linked to the full record by clicking on the SID number (PubChem's substance identifier). This displays the full substance record, that includes links: to PubMed and the source; the Medical Subject Annotation (MESH Substance Name) and a MESH PubMed search link; and depositor supplied synonyms and comments.

PubChem BioAssay

The PubChem BioAssay Summary Record,is linked to the full record by clicking on the AID number (PubChem's assay (protocol) identifier). This displays the full bioassay record, that includes: links to the substances tested (all, active, inactive, inconclusive) and related PubMed, Protein, Taxonomy, OMIM and related BioAssay records; and a description of the assay possibly with protocols and comments.

Protein Data Bank

The **Protein Data Bank** (**PDB**) is a crystallographic database for the three-dimensional structural data of large biological molecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallography,NMR spectroscopy, or, increasingly, cryo-electron microscopy, and submitted by biologists and biochemists from around the world, are freely accessible on the Internet via the websites of its member organisations (PDBe,PDBj, and RCSB). The PDB is overseen by an organization called the Worldwide Protein Data Bank, wwPDB.

The PDB is a key resource in areas of structural biology, such as structural genomics. Most major scientific journals, and some funding agencies, now require scientists to submit their structure data to the PDB. Many other databases use protein structures deposited in the PDB. For example, SCOP and CATH classify protein structures, while PDBsum provides a graphic overview of PDB entries using information from other sources, such as Gene ontology

Two forces converged to initiate the PDB: 1) a small but growing collection of sets of protein structure data determined by X-ray diffraction; and 2) the newly available (1968) molecular graphics display, the Brookhaven RAster Display (BRAD), to visualize these protein structures in 3-D. In 1969, with the sponsorship of Walter Hamilton at the Brookhaven National Laboratory, Edgar Meyer (Texas A&M University) began to write software to store atomic coordinate files in a common format to make them available for geometric and graphical evaluation. By 1971, one of Meyer's programs, SEARCH, enabled researchers to remotely access information from the database to study protein structures offline. SEARCH was instrumental in enabling networking, thus marking the functional beginning of the PDB.

Upon Hamilton's death in 1973, Tom Koeztle took over direction of the PDB for the subsequent 20 years. In January 1994, Joel Sussman of Israel's Weizmann Institute of Science. was appointed head of the PDB. In October 1998, the PDB was transferred to the Research Collaboratory for Structural Bioinformatics (RCSB); the transfer was completed in June 1999.

The new director was Helen M. Berman of Rutgers University (one of the member institutions of the RCSB). In 2003, with the formation of the wwPDB, the PDB became an international organization. The founding members are PDBe (Europe), RCSB (USA), and PDBj (Japan). The BMRB joined in 2006. Each of the four members of wwPDB can act as deposition, data processing and distribution centers for PDB data.

Table 5.1

Experimental Method	Proteins	Nucleic Acids	Protein/Nucleic Acid complexes	Other	Total
X-ray diffraction	95636	1694	4817	4	102151
NMR	9840	1135	231	8	11214
Electron microscopy	666	29	227	0	922
Hybrid	83	3	2	1	89
Other	170	4	6	13	193
Total:	106293	2865	5283	26	**114569**

91,748 structures in the PDB have a structure factor file.

The data processing refers to the fact that wwPDB staff review and annotate each submitted entry. The data are then automatically checked for plausibility (the source code for this validation software has been made available to the public at no charge).

8,531 structures have an NMR restraint file.

2,289 structures in the PDB have a chemical shifts file.

901 structures in the PDB have a 3DEM map file deposited in EM Data Bank

These data show that most structures are determined by X-ray diffraction, but about 10% of structures are now determined by protein NMR. When using X-ray diffraction, approximations of the coordinates of the atoms of the protein are obtained, whereas estimations of the distances between pairs of atoms of the protein are found through NMR experiments. Therefore, the final conformation of the protein is obtained, in the latter case, by solving a distance geometry problem. A few proteins are determined by cryo-electron microscopy.

The significance of the structure factor files, mentioned above, is that, for PDB structures determined by X-ray diffraction that have a structure file, the electron density map may be viewed. The data of such structures is stored on the "electron density server".

In the past, the number of structures in the PDB has grown at an approximately exponential rate, passing the 100 registered structures milestone in 1982, the 1,000 in 1993, the 10,000 in 1999 and the 100,000 in 2014. However, since 2007, the rate of accumulation of new protein structures appears to have plateaued.

Viewing the data

The structure files may be viewed using one of several free and open source computer programs, including Jmol, Pymol, and Rasmol. Other non-free, sharewareprograms include ICM-Browser, VMD, MDL Chime, UCSF Chimera, Swiss-PDB Viewer, StarBiochem (a Java-based interactive molecular viewer with integrated search of protein databank), Sirius, and VisProt3DS (a tool for Protein Visualization in 3D stereoscopic view in anaglyth and other modes), and Discovery Studio. The RCSB PDB website contains an extensive list of both free and commercial molecule visualization programs and web browser plugins.

PDBsum

PDBsum is database that provides an overview of the contents of each 3D macromolecular structure deposited in the Protein Data Bank. The original version of the database was developed around 1995 by Roman Laskowski and collaborators at University College London. As of 2014, PDBsum is maintained by Laskowski and collaborators in the laboratory of Janet Thornton at the European Bioinformatics Institute (EBI).

Each structure in the PDBsum database includes an image of structure (main view, Bottom view and right view), molecular components contained in the complex(structure), enzyme reaction diagram if appropriate, Gene Ontology functional assignments, a 1D sequence annotated by Pfam and InterPro domain assignments, description of bound molecules and graphic showing interactions between protein and secondary structure, schematic diagrams of protein-protein interactions, analysis of clefts contained within the structure and links to external. The RasMol and Jmol molecular graphics software are used to provide a 3D view databases of molecules and their interactions within PDBsum.

Since the release of the 1000 Genomes Project in October 2012, all single amino acid variants identified by the project have been mapped to the corresponding protein sequences in the Protein Data Bank. These variants are also displayed within PDBsum, cross-referenced to the relevant UniProt identifier. PDBsum contains a number of protein structures which may be of interest in structure-based drug design. One branch of PDBsum, known as DrugPort, focuses on these models and is linked with the DrugBank drug target database.

PDBsum: summaries and analyses of PDB structures

Roman A. Laskowski*

Department of Crystallography, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK

Received August 31, 2000; Accepted October 4, 2000

ABSTRACT

PDBsum is a web-based database providing a largely pictorial summary of the key information on each macromolecular structure deposited at the Protein Data Bank (PDB). It includes images of the structure, annotated plots of each protein chain's secondary structure, detailed structural analyses generated by the PROMOTIF program, summary PROCHECK results and schematic diagrams of protein–ligand and protein–DNA interactions. RasMol scripts highlight key aspects of the structure, such as the protein's domains, PROSITE patterns and protein–ligand interactions, for interactive viewing in 3D. Numerous links take the user to related sites. PDBsum is updated whenever any new structures are released by the PDB and is freely accessible via http://www.biochem.ucl.ac.uk/bsm/pdbsum.

INTRODUCTION

To date, the 3D structures of over 13 000 biological macromolecules have been determined experimentally, principally by X-ray crystallography and NMR spectroscopy. The majority of these are protein structures, including protein–DNA and protein–ligand complexes. Together with sequence, physicochemical and functional annotations they provide a wealth of information crucial for the understanding of biological processes.

Each new structure is deposited in the Protein Data Bank (PDB) (1), which is currently run by the Research Collaboratory in Structural Biology (RCSB) (2). The structures can be downloaded from the RCSB's PDB web server, which also provides additional information about each one. Further information, some of it focusing on specific types of molecules or specific aspects of the molecules, can be obtained from a large number of other structural databases (3) on the Web. One such database is PDBsum, which is the subject of this paper.

DESCRIPTION

The PDBsum database at http://www.biochem.ucl.ac.uk/bsm/pdbsum was created in 1995 (4). Its aim was to provide an at-a-glance summary of the molecules contained in each PDB entry (i.e. protein and DNA/RNA chains, small-molecule ligands, metal ions and waters), together with annotations and analyses of their key structural features. Thus, for each PDB entry there is a corresponding summary web page in PDBsum, accessible by the four-character PDB identifier.

The original PDBsum paper (4) described the basic contents of each entry, namely a block of 'header' information, relating to the entry as a whole, followed by a list of the molecules making up the structure, together with any relevant structural analyses of each. The header details start with a thumbnail image of the molecule(s) in question plus buttons for viewing the whole structure in 3D using RasMol (5) or VRML (Virtual Reality Modelling Language). These are followed by information extracted directly from the header records of the PDB file, summary PROCHECK (6) analyses (including a Ramachandran plot) giving an indication of the stereochemical 'quality' of all the protein chains in the structure, and links to related databases. In the list of molecules that follows, each protein chain is shown schematically by a 'wiring diagram' depicting its secondary structural motifs, primary sequence, structural domains and highlighting active site residues and residues that interact with ligands, metals or DNA/RNA molecules. The secondary structural motifs are computed by the PROMOTIF (7) program, whose detailed outputs are available via hyperlinks, while the domain definitions come from the CATH protein structural classification database (8,9). For each ligand molecule a LIGPLOT (10) diagram gives a schematic depiction of the hydrogen bonds and non-bonded interactions between it and the residues of the protein with which it interacts.

In the time since the original paper was published, a number of new analyses, links and functions have been added, and these are described in the remainder of this paper.

NEW FEATURES

The first of the additions relates only to protein–DNA and DNA–ligand complexes. The interactions between the DNA chains and any other molecules in the complex are shown schematically in a diagram generated by the NUCPLOT (11) program. Like the LIGPLOT diagrams of protein–ligand interactions, the NUCPLOT diagrams show all the hydrogen bonds and non-bonded interactions between the molecules, as calculated by HBPLUS (12). The diagrams are output in PostScript format (see, for example, the PDBsum entry for PDB code 2OR1).

Next, each protein chain now has a direct link to the SAS (Sequence Annotated by Structure) (13) database. Clicking on the link initiates a FASTA search that scans the given chain's sequence of amino acid residues against a database of all sequences in the PDB. The net result is a list of all other chains in the PDB that are similar at the sequence level to the one of interest. The SAS database provides a variety of different

annotations of the resultant multiple-sequence alignment, as well as enabling the user to view the superposed structures in 3D in RasMol.

Also new is the identification of any PROSITE (14) patterns present in each protein chain. These are patterns of residues that are found in regions that are highly conserved across all members of a given protein family and consequently characterise both the family itself and the biologically significant sites in its member proteins. In PDBsum the matching residues are coloured according to their conservation (and hence importance): from red for highly conserved, to blue for highly variable. Not all matching PROSITE patterns are shown; only those that appear to be true positives are included (15). The residues matching the PROSITE pattern can be viewed in RasMol to see where they lie in relation to the rest of the protein structure. A RasMol script renders the residues as thick sticks, coloured as on the PDBsum page, while showing the rest of the protein as a white backbone trace and any nearby ligands in spacefill. This often gives a clear indication of the structural and functional significance of the PROSITE pattern residues. See, for example, the entry for 1AAW, an aspartate aminotransferase, which contains the PROSITE pattern AA_TRANSFER_CLASS_1 corresponding to the Class 1 aminotransferases.

The RasMol scripts that display the PROSITE residues are generated on the fly by a program called RomLas (the name being a carefully chosen anagram of RasMol). The program is used throughout PDBsum to generate RasMol scripts for highlighting specific structural features. For example, below each LIGPLOT diagram there is a button for generating a RasMol script that displays the given ligand in the 3D context of the protein residues with which it interacts; the ligand is shown in thick sticks, while the protein residues are shown in wireframe and are labelled with the residue name and number.

Other new features include a simple text search facility on the home page and full listings of all the ligands and hetero groups found in the database. Links to a number of useful new databases have been added.

ACKNOWLEDGEMENTS

PDBsum is maintained at University College, London. The authors of the programs used in generating and running the PDBsum database include David Smith, Gail Hutchinson, Alex Michie, Andrew Martin, Ian McDonald, Andrew Wallace, Nick Luscombe, Duncan Milburn and Atsushi Kasuya. I would like to thank Martin Jones and John Bouquiere for their contribution to the database's development and running. Thanks also to Frances Pearl, Malcolm MacArthur, Edith Chan and, most of all, Janet Thornton.

REFERENCES

1. Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rogers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
2. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 214–218.
3. Berman,H.M. (1999) The past and future of structure databases. *Curr. Opin. Struct. Biol.*, **10**, 76–80.
4. Laskowski,R.A., Hutchinson,E.G., Michie,A.D., Wallace,A.C., Jones,M.L. and Thornton,J.M. (1997). PDBsum: a Web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.*, **22**, 488–490.
5. Sayle,R.A. and Milner-White,E.J. (1995) RASMOL: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374–376.
6. Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK - a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.*, **26**, 283–291.
7. Hutchinson,E.G. and Thornton,J.M. (1996) PROMOTIF – a program to identify and analyze structural motifs in proteins. *Protein Sci.*, **5**, 212–220.
8. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH: a hierarchic classification of protein domain structures, *Structure*, **5**, 1093–1108.
9. Pearl,F.M.G., Lee,D., Bray,J.E., Sillitoe,I., Todd,A.E., Harrison,A.P., Thornton,J.M. and Orengo,C.A. (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res.*, **28**, 277–282. Updated article in this issue: *Nucleic Acids Res.* (2001), **29**, 223–227.
10. Wallace,A.C., Laskowski,R.A. and Thornton,J.M. (1995) LIGPLOT: A program to generate schematic diagrams of protein–ligand interactions. *Protein Eng.*, **8**, 127–134.
11. Luscombe,N.M., Laskowski,R.A. and Thornton,J.M. (1997) NUCPLOT: a program to generate schematic diagrams of protein–nucleic acid interactions. *Nucleic Acids Res.*, **25**, 4940–4945.
12. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen-bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
13. Milburn,D., Laskowski,R.A. and Thornton,J.M. (1998) Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. *Protein Eng.*, **11**, 855–859.
14. Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
15. Kasuya,A. and Thornton,J.M. (1999) Three-dimensional structure analysis of PROSITE patterns. *J. Mol. Biol.*, **286**, 1673–1691.

SMILES

The simplified molecular-input line-entry system (SMILES) is a specification in form of a line notation for describing the structure of chemical species using short ASCII strings. SMILES strings can be imported by most molecule editors for conversion back into two-dimensional drawings or three-dimensional models of the molecules.

The original SMILES specification was initiated by David Weininger at the USEPA Mid- Continent Ecology Division Laboratory in Duluth in the 1980s. Acknowledged for their parts in the early development were "Gilman Veith and Rose Russo (USEPA) and Albert Leo and Corwin Hansch (Pomona College) for supporting the work, and Arthur Weininger (Pomona; Daylight CIS) and Jeremy Scofield (Cedar River Software, Renton, WA) for assistance in programming the system." The Environmental Protection Agency funded the initial project to develop SMILES.

It has since been modified and extended by others, most notably by Daylight Chemical Information Systems. In 2007, anopen standard called "OpenSMILES" was developed by the Blue Obelisk open-source chemistry community. Other 'linear' notations include the Wiswesser Line Notation (WLN), ROSDAL and SLN (Tripos Inc).

In July 2006, the IUPAC introduced the InChI as a standard for formula representation. SMILES is generally considered to have the advantage of being slightly more human-readable than InChI; it also has a wide base of software support with extensive theoretical (e.g., graph theory) backing

Terminology

The term SMILES refers to a line notation for encoding molecular structures and specific instances should strictly be called SMILES strings. However, the term SMILES is also commonly used to refer to both a single SMILES string and a number of SMILES strings; the exact meaning is usually apparent from the context. The terms "canonical" and "isomeric" can lead to some confusion when applied to SMILES. The terms describe different attributes of SMILES strings and are not mutually exclusive.

Typically, a number of equally valid SMILES strings can be written for a molecule. For example, CCO, OCC and C(O)C all specify the structure of ethanol. Algorithms have been developed to generate the same SMILES string for a given molecule; of the many possible strings, these algorithms choose only one of them. This SMILES is unique for each structure, although dependent on the canonicalization algorithm used to generate it, and is termed the canonical SMILES. These algorithms first convert the SMILES to an internal representation of the molecular structure; an algorithm then examines that structure and produces a unique SMILES string. Various algorithms for generating canonical SMILES have been developed and include those by Daylight Chemical Information Systems, OpenEye Scientific Software, MEDIT, Chemical Computing Group, MolSoft LLC, and the Chemistry Development Kit. A common application of canonical SMILES is indexing and ensuring uniqueness of molecules in a database.

The original paper that described the CANGEN[2] algorithm claimed to generate unique SMILES strings for graphs representing molecules, but the algorithm fails for a number of simple cases (e.g. cuneane, 1,2-dicyclopropylethane) and cannot be considered a correct method for representing a graph canonically. There is currently no systematic comparison across commercial software to test if such flaws exist in those packages.

SMILES notation allows the specification of configuration at tetrahedral centers, and double bond geometry. These are structural features that cannot be specified by connectivity alone and SMILES which encode this information are termed isomeric SMILES. A notable feature of these rules is that they allow rigorous partial specification of chirality. The term isomeric SMILES is also applied to SMILES in which isotopes are specified.

Graph-based definition

In terms of a graph-based computational procedure, SMILES is a string obtained by printing the symbol nodes encountered in a depth-first tree traversal of a chemical graph. The chemical graph is first trimmed to remove hydrogen atoms and cycles are broken to turn it into a spanning tree. Where cycles have been broken, numeric suffix labels are included to indicate the connected nodes. Parentheses are used to indicate points of branching on the tree.

Atoms

Atoms are represented by the standard abbreviation of the chemical elements, in square brackets, such as [Au] for gold. Brackets can be omitted for the "organic subset" of B, C, N, O, P, S, F, Cl, Br, and I. All other elements must be enclosed in brackets. If the brackets are omitted, the proper number of implicit hydrogen atoms is assumed; for instance the SMILES for water is simply O.

An atom holding one or more electrical charges is enclosed in brackets, followed by the symbol H if it is bonded to one or more atoms of hydrogen, followed by the number of hydrogen atoms (as usual one is omitted example: NH4 for ammonium), then by the sign '+' for a positive charge or by '-' for a negative charge. The number of charges is specified after the sign (except if there is one only); however, it is also possible write the sign as many times as the ion has charges: instead of "Ti+4", one can also write "Ti++++" (Titanium IV, Ti4+). Thus, the hydroxide anion is represented by [OH-], the oxonium cation is [OH3+] and the cobalt III cation (Co3+) is either [Co+3] or [Co+++].

Bonds

Bonds between aliphatic atoms are assumed to be single unless specified otherwise and are implied by adjacency in the SMILES string. For example, the SMILES for ethanolcan be written as CCO. Ring closure labels are used to indicate connectivity between non-adjacent atoms in the SMILES string, which for cyclohexane and dioxane can be written as C1CCCCC1 and O1CCOCC1 respectively. For a second ring, the label will be 2 (naphthalene: c1cccc2c1cccc2 (note the lower case for aromatic compounds)), and so on. After reaching 9, the label must be preceded by a '%', in order to differentiate it from two different labels bonded to the same atom (~C12~ will mean the atom of carbon holds the ring closure labels 1 and 2, whereas ~C%12~ will indicate one label only, 12). Double, triple, and quadruple bonds are represented by the symbols '=', '#', and '$' respectively as illustrated by the SMILES O=C=O (carbon dioxide), C#N (hydrogen cyanide) and [Ga-]$[As+] (gallium arsenide).

Aromaticity

Aromatic C, O, S and N atoms are shown in their lower case 'c', 'o', 's' and 'n' respectively. Benzene, pyridine and furan can be represented respectively by the SMILES c1ccccc1, n1ccccc1 and o1cccc1. Bonds between aromatic atoms are, by default, aromatic although these can be specified explicitly using the ':' symbol. Aromatic atoms can be singly bonded to each other and biphenyl can be represented by c1ccccc1-c2ccccc2. Aromatic nitrogen bonded to hydrogen, as found in pyrrole must be represented as [nH] and imidazole is written in SMILES notation as n1c[nH]cc1.

The Daylight and OpenEye algorithms for generating canonical SMILES differ in their treatment of aromaticity.

Visualization of 3-cyanoanisole as COc(c1)cccc1C#N.

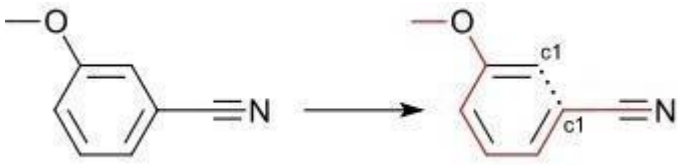


Fig 5.2

Branching

Branches are described with parentheses, as in CCC(=O)O for propionic acid and C(F)(F)F for fluoroform. Substituted rings can be written with the branching point in the ring as illustrated by the SMILES COc(c1)cccc1C#N (see depiction) and COc(cc1)ccc1C#N (see depiction) which encode the 3 and 4-cyanoanisole isomers. Writing SMILES for substituted rings in this way can make them more human-readable.

Stereochemistry

Configuration around double bonds is specified using the characters "/" and "\". For example, F/C=C/F (see depiction) is one representation of trans-difluoroethene, in which the fluorine atoms are on opposite sides of the double bond, whereas F/C=C\F (see depiction) is one possible representation of cis-difluoroethene, in which the Fs are on the same side of the double bond, as shown in the figure.

Configuration at tetrahedral carbon is specified by @ or @@. L-Alanine, the more common enantiomer of the amino acid alanine can be written as N[C@@H](C)C(=O)O (see depiction). The @@ specifier indicates that, when viewed from nitrogen along the bond to the chiral center, the sequence of substituents hydrogen (H), methyl (C) and carboxylate (C(=O)O) appear clockwise. D-Alanine can be written as N[C@H](C)C(=O)O (see depiction). The order of the substituents in the SMILES string is very important and D-alanine can also be encoded as N[C@@H](C(=O)O)C (see depiction).

Isotopes

Isotopes are specified with a number equal to the integer isotopic mass preceding the atomic symbol. Benzene in which one atom is carbon-14 is written as [14c]1ccccc1 and deuterochloroform is [2H]C(Cl)(Cl)Cl.

Examples

Table 5.2

Molecule	Structure	SMILES Formula
Dinitrogen	N≡N	N#N
Methyl isocyanate (MIC)	CH3–N=C=O	CN=C=O
Copper(II) sulfate	Cu^{2+} SO_4^{2-}	[Cu+2].[O-]S(=O)(=O)[O-]
Oenanthotoxin (C17H22O2)		CCC[C@@H](O)CC\C=C\C=C\C#CC#C\C=C\CO

SMILES can be converted back to 2-dimensional representations using Structure Diagram Generation algorithms (Helson, 1999). This conversion is not always unambiguous. Conversion to 3-dimensional representation is achieved by energy minimization approaches. There are many downloadable and web-based conversion utilities.