1 The Motivation for Quantum Mechanics

Physicists at the end of the nineteenth century believed that most of the fundamental physical laws had been worked out. They expected only minor refinements to get "an extra decimal place" of accuracy. As it turns out, the field of physics was transformed profoundly in the early twentieth century by Einstein's discovery of relativity and by the development of quantum mechanics. While relativity has had fairly little impact on chemistry, all of theoretical chemistry is founded upon quantum mechanics.

The development of quantum mechanics was initially motivated by two observations which demonstrated the inadeqacy of classical physics. These are the "ultraviolet catastrophe" and the photoelectric effect.

1.1 The Ultraviolet Catastrophe

A *blackbody* is an idealized object which absorbs and emits all frequencies. Classical physics can be used to derive an equation which describes the intensity of blackbody radiation as a function of frequency for a fixed temperature—the result is known as the Rayleigh-Jeans law. Although the Rayleigh-Jeans law works for low frequencies, it diverges as ν^2; this divergence for high frequencies is called the ultraviolet catastrophe.

Max Planck explained the blackbody radiation in 1900 by assuming that the energies of the oscillations of electrons which gave rise to the radiation must be proportional to integral multiples of the frequency, i.e.,

$$E = nh\nu \tag{1}$$

Using statistical mechanics, Planck derived an equation similar to the Rayleigh-Jeans equation, but with the adjustable parameter h. Planck found that for $h = 6.626 \times 10^{-34}$ J s, the experimental data could be reproduced. Nevertheless, Planck could not offer a good justification for his assumption of energy quantization.

Physicicsts did not take this energy quantization idea seriously until Einstein invoked a similar assumption to explain the photoelectric effect.

1.2 The Photoelectric Effect

In 1886 and 1887, Heinrich Hertz discovered that ultraviolet light can cause electrons to be ejected from a metal surface. According to the classical wave theory of light, the intensity of the light determines the amplitude of the wave, and so a greater light intensity should cause the electrons on the metal to oscillate more violently and to be ejected with a greater kinetic energy. In contrast, the experiment showed that the kinetic energy of the ejected electrons depends on the *frequency* of the light. The light intensity affects only the number of ejected electrons and not their kinetic energies.

Einstein tackled the problem of the photoelectric effect in 1905. Instead of assuming that the electronic oscillators had energies given by Planck's formula (1), Einstein assumed that the radiation itself consisted of packets of energy $E = h\nu$, which are now called photons. Einstein successfully explained the photoelectric effect using this assumption, and he calculated a value of h close to that obtained by Planck.

Two years later, Einstein showed that not only is light quantized, but so are atomic vibrations. Classical physics predicts that the molar heat capacity at constant volume (C_v) of a crystal is $3R$, where R is the molar gas constant. This works well for high temperatures, but for low temperatures C_v actually falls to zero. Einstein was able to explain this result by assuming that the oscillations of atoms about their equilibrium positions are quantized according to $E = nh\nu$, Planck's quantization condition for electronic oscillators. This demonstrated that the energy quantization concept was important even for a system of atoms in a crystal, which should be well-modeled by a system of masses and springs (i.e., by classical mechanics).

1.3 Quantization of Electronic Angular Momentum

Rutherford proposed that electrons orbit about the nucleus of an atom. One problem with this model is that, classically, orbiting electrons experience a centripetal acceleration, and accelerating charges lose energy by radiating; a stable electronic orbit is classically forbidden. Bohr nevertheless assumed stable electronic orbits with the electronic angular momentum quantized as

$$l = mvr = n\hbar \tag{2}$$

Quantization of angular momentum means that the radius of the orbit and the energy will be quantized as well. Bohr assumed that the discrete lines seen in the spectrum of the hydrogen atom were due to transitions of an electron from one allowed orbit/energy to another. He further assumed that the energy for a transition is acquired or released in the form of a photon as proposed by Einstein, so that

$$\Delta E = h\nu \tag{3}$$

This is known as the *Bohr frequency condition*. This condition, along with Bohr's expression for the allowed energy levels, gives a good match to the observed hydrogen atom spectrum. However, it works only for atoms with one electron.

1.4 Wave-Particle Duality

Einstein had shown that the momentum of a photon is

$$p = \frac{h}{\lambda} \tag{4}$$

This can be easily shown as follows. Assuming $E = h\nu$ for a photon and $\lambda\nu = c$ for an electromagnetic wave, we obtain

$$E = \frac{hc}{\lambda} \tag{5}$$

Now we use Einstein's relativity result $E = mc^2$ to find

$$\lambda = \frac{h}{mc} \tag{6}$$

which is equivalent to equation (4). Note that m refers to the relativistic mass, not the rest mass, since the rest mass of a photon is zero. Since light can behave both as a wave (it can be diffracted, and it has a wavelength), and as a particle (it contains packets of energy $h\nu$), de Broglie reasoned in 1924 that matter also can exhibit this *wave-particle duality*. He further reasoned that matter would obey the same equation (4) as light. In 1927, Davisson and Germer observed diffraction patterns by bombarding metals with electrons, confirming de Broglie's proposition.

de Broglie's equation offers a justification for Bohr's assumption (2). If we think of an electron as a wave, then for the electron orbit to be stable the wave must complete an integral number of wavelengths during its orbit. Otherwise, it would interfere destructively with itself. This condition may be written as

$$2\pi r = n\lambda \tag{7}$$

If we use the de Broglie relation (4), this can be rewritten as

$$mvr = n\hbar \tag{8}$$

which is identical to Bohr's equation (2).

Although de Broglie's equation justifies Bohr's quantization assumption, it also demonstrates a deficiency of Bohr's model. Heisenberg showed that the wave-particle duality leads to the famous uncertainty principle

$$\Delta x \Delta p \approx h \tag{9}$$

One result of the uncertainty principle is that if the orbital radius of an electron in an atom r is known exactly, then the angular momentum must be completely unknown. The problem with Bohr's model is that it specifies r exactly and it also specifies that the orbital angular momentum must be an integral multiple of $\hbar$. Thus the stage was set for a new quantum theory which was consistent with the uncertainty principle.

2 The Schrödinger Equation

In 1925, Erwin Schrödinger and Werner Heisenberg independently developed the new quantum theory. Schrödinger's method involves partial differential equations, whereas Heisenberg's method employs matrices; however, a year later the two methods were shown to be mathematically equivalent. Most textbooks begin with Schrödinger's equation, since it seems to have a better physical interpretation via the classical wave equation. Indeed, the Schrödinger equation can be viewed as a form of the wave equation applied to matter waves.

2.1 The Time-Independent Schrödinger Equation

Here we follow the treatment of McQuarrie [1], Section 3-1. We start with the one-dimensional classical wave equation,

$$\frac{\partial^2 u}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 u}{\partial t^2} \tag{10}$$

By introducing the separation of variables

$$u(x,t) = \psi(x)f(t) \tag{11}$$

we obtain

$$f(t)\frac{d^2\psi(x)}{dx^2} = \frac{1}{v^2}\psi(x)\frac{d^2 f(t)}{dt^2} \tag{12}$$

If we introduce one of the standard wave equation solutions for $f(t)$ such as $e^{i\omega t}$ (the constant can be taken care of later in the normalization), we obtain

$$\frac{d^2\psi(x)}{dx^2} = \frac{-\omega^2}{v^2}\psi(x) \tag{13}$$

Now we have an ordinary differential equation describing the spatial amplitude of the matter wave as a function of position. The energy of a particle is the sum of kinetic and potential parts

$$E = \frac{p^2}{2m} + V(x) \tag{14}$$

which can be solved for the momentum, p, to obtain

$$p = \{2m[E - V(x)]\}^{1/2} \tag{15}$$

Now we can use the de Broglie formula (4) to get an expression for the wavelength

$$\lambda = \frac{h}{p} = \frac{h}{\{2m[E - V(x)]\}^{1/2}} \tag{16}$$

The term ω^2/v^2 in equation (13) can be rewritten in terms of λ if we recall that $\omega = 2\pi\nu$ and $\nu\lambda = v$.

$$\frac{\omega^2}{v^2} = \frac{4\pi^2\nu^2}{v^2} = \frac{4\pi^2}{\lambda^2} = \frac{2m[E - V(x)]}{\hbar^2} \tag{17}$$

When this result is substituted into equation (13) we obtain the famous *time-independent Schrödinger equation*

$$\frac{d^2\psi(x)}{dx^2} + \frac{2m}{\hbar^2}[E - V(x)]\psi(x) = 0 \tag{18}$$

which is almost always written in the form

$$-\frac{\hbar^2}{2m}\frac{d^2\psi(x)}{dx^2} + V(x)\psi(x) = E\psi(x) \tag{19}$$

This single-particle one-dimensional equation can easily be extended to the case of three dimensions, where it becomes

$$-\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{r}) + V(\mathbf{r})\psi(\mathbf{r}) = E\psi(\mathbf{r}) \tag{20}$$

A two-body problem can also be treated by this equation if the mass m is replaced with a reduced mass μ.

It is important to point out that this analogy with the classical wave equation only goes so far. We cannot, for instance, derive the time-*dependent* Schrödinger equation in an analogous fashion (for instance, that equation involves the partial first derivative with respect to time instead of the partial second derivative). In fact, Schrödinger presented his time-independent equation first, and then went back and postulated the more general time-dependent equation.

2.2 The Time-Dependent Schrödinger Equation

We are now ready to consider the time-dependent Schrödinger equation. Although we were able to derive the single-particle time-independent Schrödinger equation starting from the classical wave equation and the de Broglie relation, the time-dependent Schrödinger equation cannot be derived using elementary methods and is generally given as a postulate of quantum mechanics. It is possible to show that the time-dependent equation is at least *reasonable* if not derivable, but the arguments are rather involved (cf. Merzbacher [2], Section 3.2; Levine [3], Section 1.4).

The single-particle three-dimensional time-dependent Schrödinger equation is

$$i\hbar\frac{\partial\psi(\mathbf{r},t)}{\partial t} = -\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{r},t) + V(\mathbf{r})\psi(\mathbf{r},t) \tag{21}$$

where V is assumed to be a real function and represents the potential energy of the system (a complex function V will act as a source or sink for probability, as shown in Merzbacher [2], problem 4.1). *Wave Mechanics* is the branch of quantum mechanics with equation (21) as its dynamical law. Note that equation (21) does not yet account for spin or relativistic effects.

Of course the time-dependent equation can be used to derive the time-independent equation. If we write the wavefunction as a product of spatial and temporal terms, $\psi(\mathbf{r},t) = \psi(\mathbf{r})f(t)$, then equation (21) becomes

$$\psi(\mathbf{r})i\hbar\frac{df(t)}{dt} = f(t)\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r})\right]\psi(\mathbf{r}) \tag{22}$$

or

$$\frac{i\hbar}{f(t)}\frac{df}{dt} = \frac{1}{\psi(\mathbf{r})}\left[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{r})\right]\psi(\mathbf{r}) \tag{23}$$

Since the left-hand side is a function of t only and the right hand side is a function of $\mathbf{r}$ only, the two sides must equal a constant. If we tentatively designate this constant E (since the right-hand side clearly must have the dimensions of energy),

then we extract two ordinary differential equations, namely

$$\frac{1}{f(t)}\frac{df(t)}{dt} = -\frac{iE}{\hbar} \tag{24}$$

and

$$-\frac{\hbar^2}{2m}\nabla^2\psi(\mathbf{r}) + V(\mathbf{r})\psi(\mathbf{r}) = E\psi(\mathbf{r}) \tag{25}$$

The latter equation is once again the time-independent Schrödinger equation. The former equation is easily solved to yield

$$f(t) = e^{-iEt/\hbar} \tag{26}$$

The Hamiltonian in equation (25) is a Hermitian operator, and the eigenvalues of a Hermitian operator must be real, so E is real. This means that the solutions $f(t)$ are purely oscillatory, since $f(t)$ never changes in magnitude (recall Euler's formula $e^{\pm i\theta} = cos\theta \pm i\ sin\theta$). Thus if

$$\psi(\mathbf{r}, t) = \psi(\mathbf{r})e^{-iEt/\hbar} \tag{27}$$

then the total wave function $\psi(\mathbf{r}, t)$ differs from $\psi(\mathbf{r})$ only by a phase factor of constant magnitude. There are some interesting consequences of this. First of all, the quantity $|\psi(\mathbf{r}, t)|^2$ is time independent, as we can easily show:

$$|\psi(\mathbf{r}, t)|^2 = \psi^*(\mathbf{r}, t)\psi(\mathbf{r}, t) = e^{iEt/\hbar}\psi^*(\mathbf{r})e^{-iEt/\hbar}\psi(\mathbf{r}) = \psi^*(\mathbf{r})\psi(\mathbf{r}) \tag{28}$$

Secondly, the expectation value for any time-independent operator is also time-independent, if $\psi(\mathbf{r}, t)$ satisfies equation (27). By the same reasoning applied above,

$$< A >= \int \psi^*(\mathbf{r}, t)\hat{A}\psi(\mathbf{r}, t) = \int \psi^*(\mathbf{r})\hat{A}\psi(\mathbf{r}) \tag{29}$$

For these reasons, wave functions of the form (27) are called *stationary states*. The state $\psi(\mathbf{r}, t)$ is "stationary," but the particle it describes is not!

Of course equation (27) represents a particular solution to equation (21). The general solution to equation (21) will be a linear combination of these particular solutions, i.e.

$$\psi(\mathbf{r}, t) = \sum_i c_i e^{-iE_i t/\hbar}\psi_i(\mathbf{r}) \tag{30}$$

3 Mathematical Background

3.1 Operators

Levine [3] defines an *operator* as "a rule that transforms a given function into another function" (p. 33). The differentation operator d/dx is an example—it transforms a differentiable function $f(x)$ into another function $f'(x)$. Other examples include integration, the square root, and so forth. Numbers can also be considered as operators (they multiply a function). McQuarrie [1] gives an even more general definition for an operator: "An *operator* is a symbol that tells you to do something with whatever follows the symbol" (p. 79). Perhaps this definition is more appropriate if we want to refer to the $\hat{C}_3$ operator acting on NH_3, for example.

3.1.1 Operators and Quantum Mechanics

In quantum mechanics, physical observables (e.g., energy, momentum, position, etc.) are represented mathematically by operators. For instance, the operator corresponding to energy is the Hamiltonian operator

$$\hat{H} = -\frac{\hbar^2}{2}\sum_i \frac{1}{m_i}\nabla_i^2 + V \tag{31}$$

where i is an index over all the particles of the system. We have already encountered the single-particle Hamiltonian in equation (25). The average value of an observable A represented by an operator $\hat{A}$ for a quantum molecular state $\psi(\mathbf{r})$ is given by the "expectation value" formula

$$< A >= \int \psi^*(\mathbf{r})\hat{A}\psi(\mathbf{r})d\mathbf{r} \tag{32}$$

3.1.2 Basic Properties of Operators

Most of the properties of operators are obvious, but they are summarized below for completeness.

- The **sum** and **difference** of two operators $\hat{A}$ and $\hat{B}$ are given by

$$(\hat{A} + \hat{B})f = \hat{A}f + \hat{B}f \tag{33}$$
$$(\hat{A} - \hat{B})f = \hat{A}f - \hat{B}f \tag{34}$$

- The **product** of two operators is defined by

$$\hat{A}\hat{B}f \equiv \hat{A}[\hat{B}f] \tag{35}$$

- Two operators are **equal** if

$$\hat{A}f = \hat{B}f \tag{36}$$

  for all functions f.

- The **identity operator** $\hat{1}$ does nothing (or multiplies by 1)

$$\hat{1}f = f \tag{37}$$

  A common mathematical trick is to write this operator as a sum over a complete set of states (more on this later).

$$\sum_i |i\rangle\langle i|f = f \tag{38}$$

- The **associative law** holds for operators

$$\hat{A}(\hat{B}\hat{C}) = (\hat{A}\hat{B})\hat{C} \tag{39}$$

- The **commutative law** does *not* generally hold for operators. In general, $\hat{A}\hat{B} \neq \hat{B}\hat{A}$. It is convenient to define the quantity

$$[\hat{A}, \hat{B}] \equiv \hat{A}\hat{B} - \hat{B}\hat{A} \tag{40}$$

  which is called the **commutator** of $\hat{A}$ and $\hat{B}$. Note that the order matters, so that $[\hat{A}, \hat{B}] = -[\hat{B}, \hat{A}]$. If $\hat{A}$ and $\hat{B}$ happen to commute, then $[\hat{A}, \hat{B}] = 0$.

- The **n-th power** of an operator $\hat{A}^n$ is defined as n successive applications of the operator, e.g.

$$\hat{A}^2 f = \hat{A}\hat{A}f \tag{41}$$

- The **exponential** of an operator $e^{\hat{A}}$ is defined via the power series

$$e^{\hat{A}} = \hat{1} + \hat{A} + \frac{\hat{A}^2}{2!} + \frac{\hat{A}^3}{3!} + \cdots \tag{42}$$

3.1.3 Linear Operators

Almost all operators encountered in quantum mechanics are *linear operators*. A linear operator is an operator which satisfies the following two conditions:

$$\hat{A}(f+g) = \hat{A}f + \hat{A}g \tag{43}$$
$$\hat{A}(cf) = c\hat{A}f \tag{44}$$

where c is a constant and f and g are functions. As an example, consider the operators d/dx and $()^2$. We can see that d/dx is a linear operator because

$$(d/dx)[f(x)+g(x)] = (d/dx)f(x) + (d/dx)g(x) \tag{45}$$
$$(d/dx)[cf(x)] = c\,(d/dx)f(x) \tag{46}$$

However, $()^2$ is not a linear operator because

$$(f(x)+g(x))^2 \neq (f(x))^2 + (g(x))^2 \tag{47}$$

The only other category of operators relevant to quantum mechanics is the set of *antilinear* operators, for which

$$\hat{A}(\lambda f + \mu g) = \lambda^* \hat{A}f + \mu^* \hat{A}g \tag{48}$$

Time-reversal operators are antilinear (cf. Merzbacher [2], section 16-11).

3.1.4 Eigenfunctions and Eigenvalues

An *eigenfunction* of an operator $\hat{A}$ is a function f such that the application of $\hat{A}$ on f gives f again, times a constant.

$$\hat{A}f = kf \tag{49}$$

where k is a constant called the *eigenvalue*. It is easy to show that if $\hat{A}$ is a linear operator with an eigenfunction g, then any multiple of g is also an eigenfunction of $\hat{A}$.

When a system is in an *eigenstate* of observable A (i.e., when the wavefunction is an eigenfunction of the operator $\hat{A}$) then the expectation value of A is the eigenvalue of the wavefunction. Thus if

$$\hat{A}\psi(\mathbf{r}) = a\psi(\mathbf{r}) \tag{50}$$

then

$$\begin{aligned} < A > &= \int \psi^*(\mathbf{r})\hat{A}\psi(\mathbf{r})d\mathbf{r} \\ &= \int \psi^*(\mathbf{r})a\psi(\mathbf{r})d\mathbf{r} \\ &= a\int \psi^*(\mathbf{r})\psi(\mathbf{r})d\mathbf{r} \\ &= a \end{aligned} \tag{51}$$

assuming that the wavefunction is normalized to 1, as is generally the case. In the event that $\psi(\mathbf{r})$ is not or cannot be normalized (free particle, etc.) then we may use the formula

$$< A >= \frac{\int \psi^*(\mathbf{r})\hat{A}\psi(\mathbf{r})}{d}\mathbf{r}\int \psi^*(\mathbf{r})\psi(\mathbf{r})d\mathbf{r} \tag{52}$$

What if the wavefunction is a combination of eigenstates? Let us assume that we have a wavefunction which is a linear combination of two eigenstates of $\hat{A}$ with eigenvalues a and b.

$$\psi = c_a\psi_a + c_b\psi_b \tag{53}$$

where $\hat{A}\psi_a = a\psi_a$ and $\hat{A}\psi_b = b\psi_b$. Then what is the expectation value of A?

$$\begin{aligned} < A > &= \int \psi^* \hat{A} \psi \\ &= \int \left[c_a\psi_a + c_b\psi_b\right]^* \hat{A} \left[c_a\psi_a + c_b\psi_b\right] \\ &= \int \left[c_a\psi_a + c_b\psi_b\right]^* \left[ac_a\psi_a + bc_b\psi_b\right] \\ &= a|c_a|^2 \int \psi_a^*\psi_a + bc_a^*c_b \int \psi_a^*\psi_b + ac_b^*c_a \int \psi_b^*\psi_a + b|c_b|^2 \int \psi_b^*\psi_b \\ &= a|c_a|^2 + b|c_b|^2 \end{aligned} \tag{54}$$

assuming that ψ_a and ψ_b are orthonormal (shortly we will show that eigenvectors of Hermitian operators are orthogonal). Thus the average value of A is a weighted average of eigenvalues, with the weights being the squares of the coefficients of the eigenvectors in the overall wavefunction.

3.1.5 Hermitian Operators

As mentioned previously, the expectation value of an operator $\hat{A}$ is given by

$$< A >= \int \psi^*(\mathbf{r}) \hat{A} \psi(\mathbf{r}) d\mathbf{r} \tag{55}$$

and all physical observables are represented by such expectation values. Obviously, the value of a physical observable such as energy or density must be real, so we require $< A >$ to be real. This means that we must have $< A >=< A >^*$, or

$$\int \psi^*(\mathbf{r}) \hat{A} \psi(\mathbf{r}) d\mathbf{r} = \int (\hat{A}\psi(\mathbf{r}))^* \psi(\mathbf{r}) d\mathbf{r} \tag{56}$$

Operators $\hat{A}$ which satisfy this condition are called *Hermitian*. One can also show that for a Hermitian operator,

$$\int \psi_1^*(\mathbf{r}) \hat{A} \psi_2(\mathbf{r}) d\mathbf{r} = \int (\hat{A}\psi_1(\mathbf{r}))^* \psi_2(\mathbf{r}) d\mathbf{r} \tag{57}$$

for any two states ψ_1 and ψ_2.

An important property of Hermitian operators is that their eigenvalues are real. We can see this as follows: if we have an eigenfunction of $\hat{A}$ with eigenvalue

a, i.e. $\hat{A}\psi_a = a\psi_a$, then for a Hermitian operator $\hat{A}$

$$\begin{aligned} \int \psi_a^* \hat{A} \psi_a &= \int \psi_a (\hat{A}\psi_a)^* \\ a \int \psi_a^* \psi_a &= a^* \int \psi_a \psi_a^* \\ (a - a^*) \int |\psi_a|^2 &= 0 \end{aligned} \tag{58}$$

Since $|\psi_a|^2$ is never negative, we must have either $a = a^*$ or $\psi_a = 0$. Since $\psi_a = 0$ is not an acceptable wavefunction, $a = a^*$, so a is real.

Another important property of Hermitian operators is that their eigenvectors are orthogonal (or can be chosen to be so). Suppose that ψ_a and ψ_b are eigenfunctions of $\hat{A}$ with eigenvalues a and b, with $a \neq b$. If $\hat{A}$ is Hermitian then

$$\begin{aligned} \int \psi_a^* \hat{A} \psi_b &= \int \psi_b (\hat{A}\psi_a)^* \\ b \int \psi_a^* \psi_b &= a^* \int \psi_b \psi_a^* \\ (b - a) \int \psi_a^* \psi_b &= 0 \end{aligned} \tag{59}$$

since $a = a^*$ as shown above. Because we assumed $b \neq a$, we must have $\int \psi_a^* \psi_b = 0$, i.e. ψ_a and ψ_b are orthogonal. Thus we have shown that eigenfunctions of a Hermitian operator with different eigenvalues are orthogonal. In the case of degeneracy (more than one eigenfunction with the same eigenvalue), we can *choose* the eigenfunctions to be orthogonal. We can easily show this for the case of two eigenfunctions of $\hat{A}$ with the same eigenvalue. Suppose we have

$$\begin{aligned} \hat{A}\psi_j &= j\psi_j \\ \hat{A}\psi_k &= j\psi_k \end{aligned} \tag{60}$$

We now want to take linear combinations of ψ_j and ψ_k to form two new eigenfunctions $\psi_{j'}$ and $\psi_{k'}$, where $\psi_{j'} = \psi_j$ and $\psi_{k'} = \psi_k + c\psi_j$. Now we want $\psi_{j'}$ and $\psi_{k'}$ to be orthogonal, so

$$\begin{aligned} \int \psi_{j'}^* \psi_{k'} &= 0 \\ \int \psi_j^* (\psi_k + c\psi_j) &= 0 \\ \int \psi_j^* \psi_k + c \int \psi_j^* \psi_j &= 0 \end{aligned} \tag{61}$$

Thus we merely need to choose

$$c = -\frac{\int \psi_j^* \psi_k}{\int \psi_j^* \psi_j} \tag{62}$$

and we obtain orthogonal eigenfunctions. This Schmidt-orthogonalization procedure can be extended to the case of n-fold degeneracy, so we have shown that for a Hermitian operator, the eigenvectors can be made orthogonal.

3.1.6 Unitary Operators

A linear operator whose inverse is its adjoint is called *unitary*. These operators can be thought of as generalizations of complex numbers whose absolue value is 1.

$$\begin{aligned} U^{-1} &= U^\dagger \\ UU^\dagger &= U^\dagger U = I \end{aligned} \tag{63}$$

A unitary operator preserves the "lengths" and "angles" between vectors, and it can be considered as a type of rotation operator in abstract vector space. Like Hermitian operators, the eigenvectors of a unitary matrix are orthogonal. However, its eigenvalues are not necessarily real.

3.2 Commutators in Quantum Mechanics

The *commutator*, defined in section 3.1.2, is very important in quantum mechanics. Since a definite value of observable A can be assigned to a system only if the system is in an eigenstate of $\hat{A}$, then we can simultaneously assign definite values to two observables A and B only if the system is in an eigenstate of both $\hat{A}$ and $\hat{B}$. Suppose the system has a value of A_i for observable A and B_j for observable B. The we require

$$\begin{aligned} \hat{A}\psi_{A_i,B_j} &= A_i \psi_{A_i,B_j} \\ \hat{B}\psi_{A_i,B_j} &= B_j \psi_{A_i,B_j} \end{aligned} \tag{64}$$

If we multiply the first equation by $\hat{B}$ and the second by $\hat{A}$ then we obtain

$$\begin{aligned} \hat{B}\hat{A}\psi_{A_i,B_j} &= \hat{B}A_i\psi_{A_i,B_j} \\ \hat{A}\hat{B}\psi_{A_i,B_j} &= \hat{A}B_j\psi_{A_i,B_j} \end{aligned} \tag{65}$$

and, using the fact that ψ_{A_i,B_j} is an eigenfunction of $\hat{A}$ and $\hat{B}$, this becomes

$$\begin{aligned} \hat{B}\hat{A}\psi_{A_i,B_j} &= A_iB_j\psi_{A_i,B_j} \\ \hat{A}\hat{B}\psi_{A_i,B_j} &= B_jA_i\psi_{A_i,B_j} \end{aligned} \tag{66}$$

so that if we subtract the first equation from the second, we obtain

$$(\hat{A}\hat{B} - \hat{B}\hat{A})\psi_{A_i,B_j} = 0 \tag{67}$$

For this to hold for general eigenfunctions, we must have $\hat{A}\hat{B} = \hat{B}\hat{A}$, or $[\hat{A}, \hat{B}] = 0$. That is, for two physical quantities to be simultaneously observable, their operator representations must commute.

Section 8.8 of Merzbacher [2] contains some useful rules for evaluating commutators. They are summarized below.

$$[\hat{A}, \hat{B}] + [\hat{B}, \hat{A}] = 0 \tag{68}$$

$$[\hat{A}, \hat{A}] = 0 \tag{69}$$

$$[\hat{A}, \hat{B} + \hat{C}] = [\hat{A}, \hat{B}] + [\hat{A}, \hat{C}] \tag{70}$$

$$[\hat{A} + \hat{B}, \hat{C}] = [\hat{A}, \hat{C}] + [\hat{B}, \hat{C}] \tag{71}$$

$$[\hat{A}, \hat{B}\hat{C}] = [\hat{A}, \hat{B}]\hat{C} + \hat{B}[\hat{A}, \hat{C}] \tag{72}$$

$$[\hat{A}\hat{B}, \hat{C}] = [\hat{A}, \hat{C}]\hat{B} + \hat{A}[\hat{B}, \hat{C}] \tag{73}$$

$$[\hat{A}, [\hat{B}, \hat{C}]] + [\hat{C}, [\hat{A}, \hat{B}]] + [\hat{B}, [\hat{C}, \hat{A}]] = 0 \tag{74}$$

If $\hat{A}$ and $\hat{B}$ are two operators which commute with their commutator, then

$$[\hat{A}, \hat{B}^n] = n\hat{B}^{n-1}[\hat{A}, \hat{B}] \tag{75}$$

$$[\hat{A}^n, \hat{B}] = n\hat{A}^{n-1}[\hat{A}, \hat{B}] \tag{76}$$

We also have the identity (useful for coupled-cluster theory)

$$e^{\hat{A}}\hat{B}e^{-\hat{A}} = \hat{B} + [\hat{A},\hat{B}] + \frac{1}{2!}[\hat{A},[\hat{A},\hat{B}]] + \frac{1}{3!}[\hat{A},[\hat{A},[\hat{A},\hat{B}]]] + \cdots \tag{77}$$

Finally, if $[\hat{A},\hat{B}] = i\hat{C}$ then the uncertainties in A and B, defined as $\Delta A^2 = < A^2 > - < A >^2$, obey the relation[1]

$$(\Delta A)(\Delta B) \geq \frac{1}{2}| < C > | \tag{78}$$

This is the famous Heisenberg uncertainty principle. It is easy to derive the well-known relation

$$(\Delta x)(\Delta p_x) \geq \frac{\hbar}{2} \tag{79}$$

from this generalized rule.

3.3 Linear Vector Spaces in Quantum Mechanics

We have observed that most operators in quantum mechanics are linear operators. This is fortunate because it allows us to represent quantum mechanical operators as matrices and wavefunctions as vectors in some linear vector space. Since computers are particularly good at performing operations common in linear algebra (multiplication of a matrix times a vector, etc.), this is quite advantageous from a practical standpoint.

In an n-dimensional space we may expand any vector Ψ as a linear combination of basis vectors

$$\Psi = \sum_{i=1}^{n} a_i \Psi_i \tag{80}$$

For a general vector space, the coefficients a_i may be complex; thus one should not be too quick to draw parallels to the expansion of vectors in three-dimensional Euclidean space. The coefficients a_i are referred to as the "components" of the state vector Ψ, and for a given basis, the components of a vector specify it completely.

[1] Assuming that the quantum covariance $< (\hat{A}\hat{B} + \hat{B}\hat{A})/2 - < \hat{A} >< \hat{B} >>$ is zero.

The components of the sum of two vectors are the sums of the components. If $\Psi_a = \sum a_i \Psi_i$ and $\Psi_b = \sum b_i \Psi_i$ then

$$\Psi_a + \Psi_b = \sum_i (a_i + b_i)\Psi_i \tag{81}$$

and similarly

$$\lambda \Psi_a = \sum_i (\lambda a_i)\Psi_i \tag{82}$$

The *scalar product* of two vectors is a complex number denoted by

$$(\Psi_b, \Psi_a) = (\Psi_a, \Psi_b)^* \tag{83}$$

where we have used the standard linear-algebra notation. If we also require that

$$(\Psi_a, \lambda \Psi_b) = \lambda (\Psi_a, \Psi_b) \tag{84}$$

then it follows that

$$(\lambda \Psi_a, \Psi_b) = \lambda^* (\Psi_a, \Psi_b) \tag{85}$$

We also require that

$$(\Psi_a, \Psi_b + \Psi_c) = (\Psi_a, \Psi_b) + (\Psi_a, \Psi_c) \tag{86}$$

If the scalar product vanishes (and if neither vector in the product is the null vector) then the two vectors are orthogonal.

Generally the basis is chosen to be orthonormal, such that

$$(\hat{\Psi}_i, \hat{\Psi}_j) = \delta_{ij} \tag{87}$$

In this case, we can write the scalar product of two arbitrary vectors as

$$\begin{aligned} (\Psi_a, \Psi_b) &= (\sum_i a_i \hat{\Psi}_i, \sum_j b_j \hat{\Psi}_j) \\ &= \sum_i \sum_j a_i^* b_j (\hat{\Psi}_i, \hat{\Psi}_j) \\ &= \sum_i a_i^* b_i \end{aligned} \tag{88}$$

This can also be written in vector notation as

$$(\Psi_a, \Psi_b) = (a_1^* a_2^* \cdots a_n^*) \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{89}$$

It is useful at this point to introduce Dirac's bra-ket notation. We define a "bra" as

$$\langle \Psi_a | = (a_1^* a_2^* \cdots a_n^*) \tag{90}$$

and a "ket" as

$$|\Psi_a\rangle = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \tag{91}$$

A bra to the left of a ket implies a scalar product, so

$$\langle \Psi_a | \Psi_b \rangle = (\Psi_a, \Psi_b) \tag{92}$$

Sometimes in superficial treatments of Dirac notation, the symbol $\langle \Psi_a | \Psi_b \rangle$ is defined alternatively as

$$\langle \Psi_a | \Psi_b \rangle = \int \Psi_a^*(x) \Psi_b(x) dx \tag{93}$$

This is equivalent to the above definition if we make the connections $a_i = \Psi_a(x)$ and $b_i = \Psi_b(x)$. This means that our basis vectors are *every possible value of x.* Since x is continuous, the sum is replaced by an integral (see Szabo and Ostlund [4] , exercise 1.17). Often only the subscript of the vector is used to denote a bra or ket; we may have written the above equation as

$$\langle a | b \rangle = \int \Psi_a^*(x) \Psi_b(x) dx \tag{94}$$

Now we turn our attention to matrix representations of operators. An operator $\hat{A}$ can be characterized by its effect on the basis vectors. The action of $\hat{A}$ on a

basis vector $\hat{\Psi}_j$ yields some new vector Ψ'_j which can be expanded in terms of the basis vectors so long as we have a complete basis set.

$$\hat{A}\hat{\Psi}_j = \Psi'_j = \sum_i^n \hat{\Psi}_i A_{ij} \tag{95}$$

If we know the effect of $\hat{A}$ on the basis vectors, then we know the effect of $\hat{A}$ on any arbitrary vector because of the linearity of $\hat{A}$.

$$\begin{aligned} \Psi_b = \hat{A}\Psi_a = \hat{A}\sum_j a_j\hat{\Psi}_j &= \sum_j a_j\hat{A}\hat{\Psi}_j = \sum_j\sum_i a_j\hat{\Psi}_i A_{ij} \\ &= \sum_i \hat{\Psi}_i(\sum_j A_{ij}a_j) \end{aligned} \tag{96}$$

or

$$b_i = \sum_j A_{ij}a_j \tag{97}$$

This may be written in matrix notation as

$$\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \tag{98}$$

We can obtain the coefficients A_{ij} by taking the inner product of both sides of equation 95 with $\hat{\Psi}_i$, yielding

$$\begin{aligned} (\hat{\Psi}_i, \hat{A}\hat{\Psi}_j) &= (\hat{\Psi}_i, \sum_k^n \hat{\Psi}_k A_{kj}) \\ &= \sum_k^n A_{kj}(\hat{\Psi}_i, \hat{\Psi}_k) \\ &= A_{ij} \end{aligned} \tag{99}$$

since $(\hat{\Psi}_i, \hat{\Psi}_k) = \delta_{ik}$ due to the orthonormality of the basis. In bra-ket notation, we may write

$$A_{ij} = \langle i|\hat{A}|j\rangle \tag{100}$$

where i and j denote two basis vectors. This use of bra-ket notation is consistent with its earlier use if we realize that $\hat{A}|j\rangle$ is just another vector $|j'\rangle$.

It is easy to show that for a linear operator $\hat{A}$, the inner product $(\Psi_a, \hat{A}\Psi_b)$ for two general vectors (not necessarily basis vectors) Ψ_a and Ψ_b is given by

$$(\Psi_a, \hat{A}\Psi_b) = \sum_i \sum_j a_i^* A_{ij} b_j \tag{101}$$

or in matrix notation

$$(\Psi_a, \hat{A}\Psi_b) = (a_1^* a_2^* \cdots a_n^*) \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{102}$$

By analogy to equation (93), we may generally write this inner product in the form

$$(\Psi_a, \hat{A}\Psi_b) = \langle a|\hat{A}|b\rangle = \int \Psi_a^*(x)\hat{A}\Psi_b(x)dx \tag{103}$$

Previously, we noted that $(\Psi_a, \Psi_b) = (\Psi_b, \Psi_a)^*$, or $\langle a|b\rangle = \langle b|a\rangle^*$. Thus we can see also that

$$(\Psi_a, \hat{A}\Psi_b) = (\hat{A}\Psi_b, \Psi_a)^* \tag{104}$$

We now define the *adjoint* of an operator $\hat{A}$, denoted by $\hat{A}^\dagger$, as that linear operator for which

$$(\Psi_a, \hat{A}\Psi_b) = (\hat{A}^\dagger\Psi_a, \Psi_b) \tag{105}$$

That is, we can make an operator act *backwards* into "bra" space if we take it's adjoint. With this definition, we can further see that

$$(\Psi_a, \hat{A}\Psi_b) = (\hat{A}\Psi_b, \Psi_a)^* = (\Psi_b, \hat{A}^\dagger\Psi_a)^* = (\hat{A}^\dagger\Psi_a, \Psi_b) \tag{106}$$

or, in bra-ket notation,

$$\langle a|\hat{A}|b\rangle = \langle \hat{A}b|a\rangle^* = \langle b|\hat{A}^\dagger|a\rangle^* = \langle \hat{A}^\dagger a|b\rangle \tag{107}$$

If we pick $\Psi_a = \hat{\Psi}_i$ and $\Psi_b = \hat{\Psi}_j$ (i.e., if we pick two basis vectors), then we obtain

$$\begin{aligned} (\hat{A}\hat{\Psi}_i, \hat{\Psi}_j) &= (\hat{\Psi}_i, \hat{A}^\dagger \hat{\Psi}_j) \\ (\hat{\Psi}_j, \hat{A}\hat{\Psi}_i)^* &= (\hat{\Psi}_i, \hat{A}^\dagger \hat{\Psi}_j) \\ A^*_{ji} &= A^\dagger_{ij} \end{aligned} \tag{108}$$

But this is precisely the condition for the elements of a matrix and its adjoint! Thus the adjoint of the matrix representation of $\hat{A}$ is the same as the matrix representation of $\hat{A}^\dagger$.

This correspondence between operators and their matrix representations goes quite far, although of course the specific matrix representation depends on the choice of basis. For instance, we know from linear algebra that if a matrix and its adjoint are the same, then the matrix is called Hermitian. The same is true of the operators; if

$$\hat{A} = \hat{A}^\dagger \tag{109}$$

then $\hat{A}$ is a Hermitian operator, and all of the special properties of Hermitian operators apply to $\hat{A}$ or its matrix representation.

4 Postulates of Quantum Mechanics

In this section, we will present six postulates of quantum mechanics. Again, we follow the presentation of McQuarrie [1], with the exception of postulate 6, which McQuarrie does not include. A few of the postulates have already been discussed in section 3.

> **Postulate 1**. The state of a quantum mechanical system is completely specified by a function $\Psi(\mathbf{r}, t)$ that depends on the coordinates of the particle(s) and on time. This function, called the wave function or state function, has the important property that $\Psi^*(\mathbf{r}, t)\Psi(\mathbf{r}, t)d\tau$ is the probability that the particle lies in the volume element $d\tau$ located at $\mathbf{r}$ at time t.

The wavefunction must satisfy certain mathematical conditions because of this probabilistic interpretation. For the case of a single particle, the probability of finding it *somewhere* is 1, so that we have the normalization condition

$$\int_{-\infty}^{\infty} \Psi^*(\mathbf{r}, t)\Psi(\mathbf{r}, t)d\tau = 1 \tag{110}$$

It is customary to also normalize many-particle wavefunctions to 1.[2] The wavefunction must also be single-valued, continuous, and finite.

> **Postulate 2**. To every observable in classical mechanics there corresponds a linear, Hermitian operator in quantum mechanics.

This postulate comes about because of the considerations raised in section 3.1.5: if we require that the expectation value of an operator $\hat{A}$ is real, then $\hat{A}$ must be a Hermitian operator. Some common operators occuring in quantum mechanics are collected in Table 1.

[2] In some cases, such as the free-particle, one must use special tricks to normalize the wavefunction. See Merzbacher [2], section 8.1.

Table 1: Physical observables and their corresponding quantum operators (single particle)

Observable Name	Observable Symbol	Operator Symbol	Operator Operation
Position	$\mathbf{r}$	$\hat{\mathbf{r}}$	Multiply by $\mathbf{r}$
Momentum	$\mathbf{p}$	$\hat{\mathbf{p}}$	$-i\hbar\left(\hat{i}\frac{\partial}{\partial x}+\hat{j}\frac{\partial}{\partial y}+\hat{k}\frac{\partial}{\partial z}\right)$
Kinetic energy	T	$\hat{T}$	$-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)$
Potential energy	$V(\mathbf{r})$	$\hat{V}(\mathbf{r})$	Multiply by $V(\mathbf{r})$
Total energy	E	$\hat{H}$	$-\frac{\hbar^2}{2m}\left(\frac{\partial^2}{\partial x^2}+\frac{\partial^2}{\partial y^2}+\frac{\partial^2}{\partial z^2}\right)+V(\mathbf{r})$
Angular momentum	l_x	$\hat{l}_x$	$-i\hbar\left(y\frac{\partial}{\partial z}-z\frac{\partial}{\partial y}\right)$
	l_y	$\hat{l}_y$	$-i\hbar\left(z\frac{\partial}{\partial x}-x\frac{\partial}{\partial z}\right)$
	l_z	$\hat{l}_z$	$-i\hbar\left(x\frac{\partial}{\partial y}-y\frac{\partial}{\partial x}\right)$

Postulate 3. In any measurement of the observable associated with operator $\hat{A}$, the only values that will ever be observed are the eigenvalues a, which satisfy the eigenvalue equation

$$\hat{A}\Psi = a\Psi \tag{111}$$

This postulate captures the central point of quantum mechanics—the values of dynamical variables can be quantized (although it is still possible to have a continuum of eigenvalues in the case of unbound states). If the system is in an eigenstate of $\hat{A}$ with eigenvalue a, then any measurement of the quantity A will yield a.

Although measurements must always yield an eigenvalue, the state does not have to be an eigenstate of $\hat{A}$ *initially*. An arbitrary state can be expanded in the complete set of eigenvectors of $\hat{A}$ ($\hat{A}\Psi_i = a_i\Psi_i$) as

$$\Psi = \sum_i^n c_i\Psi_i \tag{112}$$

where n may go to infinity. In this case we only know that the measurement of A will yield *one* of the values a_i, but we don't know which one. However, we do know the *probability* that eigenvalue a_i will occur—it is the absolute value squared of the coefficient, $|c_i|^2$ (cf. section 3.1.4), leading to the fourth postulate below.

An important second half of the third postulate is that, after measurement of Ψ yields some eigenvalue a_i, the wavefunction immediately "collapses" into the corresponding eigenstate Ψ_i (in the case that a_i is degenerate, then Ψ becomes the projection of Ψ onto the degenerate subspace). Thus, measurement affects the state of the system. This fact is used in many elaborate experimental tests of quantum mechanics.

> **Postulate 4**. If a system is in a state described by a normalized wave function Ψ, then the average value of the observable corresponding to $\hat{A}$ is given by

$$< A >= \int_{-\infty}^{\infty} \Psi^* \hat{A} \Psi d\tau \tag{113}$$

> **Postulate 5**. The wavefunction or state function of a system evolves in time according to the time-dependent Schrödinger equation

$$\hat{H}\Psi(\mathbf{r}, t) = i\hbar \frac{\partial \Psi}{\partial t} \tag{114}$$

The central equation of quantum mechanics must be accepted as a postulate, as discussed in section 2.2.

> **Postulate 6**. The total wavefunction must be antisymmetric with respect to the interchange of all coordinates of one fermion with those of another. Electronic spin must be included in this set of coordinates.

The Pauli exclusion principle is a direct result of this *antisymmetry principle*. We will later see that Slater determinants provide a convenient means of enforcing this property on electronic wavefunctions.

5 Some Analytically Soluble Problems

Quantum chemists are generally concerned with solving the time-independent Schrödinger equation (25). This equation can be solved analytically only in a few special cases. In this section we review the results of some of these analytically soluble problems.

5.1 The Particle in a Box

Consider a particle constrained to move in a single dimension, under the influence of a potential $V(x)$ which is zero for $0 \leq x \leq a$ and infinite elsewhere. Since the wavefunction is not allowed to become infinite, it must have a value of zero where $V(x)$ is infinite, so $\psi(x)$ is nonzero only within $[0, a]$. The Schrödinger equation is thus

$$-\frac{\hbar^2}{2m}\frac{d^2\psi}{dx^2} = E\psi(x) \quad 0 \leq x \leq a \tag{115}$$

It is easy to show that the eigenvectors and eigenvalues of this problem are

$$\psi_n(x) = \sqrt{\frac{2}{a}} sin\left(\frac{n\pi x}{a}\right) \qquad 0 \leq x \leq a \qquad n = 1, 2, 3, \ldots \tag{116}$$

$$E_n = \frac{h^2 n^2}{8ma^2} \qquad n = 1, 2, \ldots \tag{117}$$

Extending the problem to three dimensions is rather straightforward; see McQuarrie [1], section 6.1.

5.2 The Harmonic Oscillator

Now consider a particle subject to a restoring force $F = -kx$, as might arise for a mass-spring system obeying Hooke's Law. The potential is then

$$\begin{aligned} V(x) &= -\int_{-\infty}^{\infty}(-kx)dx \\ &= V_0 + \frac{1}{2}kx^2 \end{aligned} \tag{118}$$

If we choose the energy scale such that $V_0 = 0$ then $V(x) = (1/2)kx^2$. This potential is also appropriate for describing the interaction of two masses connected by an ideal spring. In this case, we let x be the distance between the masses, and for the mass m we substitute the reduced mass μ. Thus the harmonic oscillator is the simplest model for the vibrational motion of the atoms in a diatomic molecule, if we consider the two atoms as point masses and the bond between them as a spring. The one-dimensional Schrödinger equation becomes

$$-\frac{\hbar^2}{2\mu}\frac{d^2\psi}{dx^2} + \frac{1}{2}kx^2\psi(x) = E\psi(x) \tag{119}$$

After some effort, the eigenfunctions are

$$\psi_n(x) = N_n H_n(\alpha^{1/2}x)e^{-\alpha x^2/2} \qquad n = 0, 1, 2, \ldots \tag{120}$$

where H_n is the Hermite polynomial of degree n, and α and N_n are defined by

$$\alpha = \sqrt{\frac{k\mu}{\hbar^2}} \qquad N_n = \frac{1}{\sqrt{2^n n!}}\left(\frac{\alpha}{\pi}\right)^{1/4} \tag{121}$$

The eigenvalues are

$$E_n = \hbar\omega(n + 1/2) \tag{122}$$

with $\omega = \sqrt{k/\mu}$.

5.3 The Rigid Rotor

The rigid rotor is a simple model of a rotating diatomic molecule. We consider the diatomic to consist of two point masses at a fixed internuclear distance. We then reduce the model to a one-dimensional system by considering the rigid rotor to have one mass fixed at the origin, which is orbited by the reduced mass μ, at a distance r. The Schrödinger equation is (cf. McQuarrie [1], section 6.4 for a clear explanation)

$$-\frac{\hbar^2}{2I}\left[\frac{1}{sin\theta}\frac{\partial}{\partial\theta}\left(sin\theta\frac{\partial}{\partial\theta}\right) + \frac{1}{sin^2\theta}\frac{\partial^2}{\partial\phi^2}\right]\psi(r) = E\psi(r) \tag{123}$$

After a little effort, the eigenfunctions can be shown to be the spherical harmonics $Y_J^M(\theta, \phi)$, defined by

$$Y_J^M(\theta,\phi) = \left[\frac{(2J+1)}{4\pi}\frac{(J-|M|)!}{(J+|M|)!}\right]^{1/2} P_J^{|M|}(cos\theta)e^{iM\phi} \tag{124}$$

where $P_J^{|M|}(x)$ are the associated Legendre functions. The eigenvalues are simply

$$E_J = \frac{\hbar^2}{2I}J(J+1) \tag{125}$$

Each energy level E_J is $2J+1$-fold degenerate in M, since M can have values $-J, -J+1, \ldots, J-1, J$.

5.4 The Hydrogen Atom

Finally, consider the hydrogen atom as a proton fixed at the origin, orbited by an electron of reduced mass μ. The potential due to electrostatic attraction is

$$V(r) = -\frac{e^2}{4\pi\epsilon_0 r} \tag{126}$$

in SI units. The kinetic energy term in the Hamiltonian is

$$\hat{T} = -\frac{\hbar^2}{2\mu}\nabla^2 \tag{127}$$

so we write out the Schrödinger equation in spherical polar coordinates as

$$-\frac{\hbar^2}{2\mu}\left[\frac{1}{r^2}\frac{\partial}{\partial r}\left(r^2\frac{\partial\psi}{\partial r}\right)\frac{1}{r^2 sin\theta}\frac{\partial}{\partial\theta}\left(sin\theta\frac{\partial\psi}{\partial\theta}\right)+\frac{1}{r^2 sin^2\theta}\frac{\partial^2\psi}{\partial\phi^2}\right]-\frac{e^2}{4\pi\epsilon_0 r}\psi(r,\theta,\phi) = E\psi(r,\theta,\phi) \tag{128}$$

It happens that we can factor $\psi(r,\theta,\phi)$ into $R(r)_{nl}Y_l^m(\theta,\phi)$, where $Y_l^m(\theta,\phi)$ are again the spherical harmonics. The radial part $R(r)$ then can be shown to obey the equation

$$-\frac{\hbar^2}{2\mu r^2}\frac{d}{dr}\left(r^2\frac{dR}{dr}\right)+\left[\frac{\hbar^2 l(l+1)}{2\mu r^2}+V(r)-E\right]R(r) = 0 \tag{129}$$

which is called the *radial equation* for the hydrogen atom. Its (messy) solutions are

$$R_{nl}(r) = -\left[\frac{(n-l-1)!}{2n[(n+l)!]^3}\right]^{1/2}\left(\frac{2}{na_0}\right)^{l+3/2} r^l e^{-r/na_0} L_{n+l}^{2l+1}\left(\frac{2r}{na_0}\right) \tag{130}$$

where $0 \leq l \leq n-1$, and a_0 is the Bohr radius, $\epsilon_0 h^2/\pi\mu e^2$. The functions $L_{n+l}^{2l+1}(2r/na_0)$ are the associated Laguerre functions. The hydrogen atom eigenvalues are

$$E_n = -\frac{e^2}{8\pi\epsilon_0 a_0 n^2} \qquad n = 1, 2, \ldots \tag{131}$$

There are relatively few other interesting problems that can be solved analytically. For molecular systems, one must resort to approximate solutions.

6 Approximate Methods

The problems discussed in the previous section (harmonic oscillator, rigid rotator, etc.) are some of the few quantum mechanics problems which can be solved analytically. For the vast majority of chemical applications, the Schrödinger equation must be solved by approximate methods. The two primary approximation techniques are the variational method and perturbation theory.

6.1 Perturbation Theory

The basic idea of perturbation theory is very simple: we split the Hamiltonian into a piece we know how to solve (the "reference" or "unperturbed" Hamiltonian) and a piece we don't know how to solve (the "perturbation"). As long as the perburbation is small compared to the unperturbed Hamiltonian, perturbation theory tells us how to correct the solutions to the unperturbed problem to approximately account for the influence of the perturbation. For example, perturbation theory can be used to approximately solve an anharmonic oscillator problem with the Hamiltonian

$$\hat{H} = -\frac{\hbar^2}{2\mu}\frac{d^2}{dx^2} + \frac{1}{2}kx^2 + \frac{1}{6}\gamma x^3. \tag{132}$$

Here, since we know how to solve the harmonic oscillator problem (see 5.2), we make that part the unperturbed Hamiltonian (denoted $\hat{H}^{(0)}$), and the new, anharmonic term is the perturbation (denoted $\hat{H}^{(1)}$):

$$\hat{H}^{(0)} = -\frac{\hbar^2}{2\mu}\frac{d^2}{dx^2} + \frac{1}{2}kx^2, \tag{133}$$

$$\hat{H}^{(1)} = +\frac{1}{6}\gamma x^3. \tag{134}$$

Perturbation theory solves such a problem in two steps. First, obtain the eigenfunctions and eigenvalues of the unperturbed Hamiltonian, $\hat{H}^{(0)}$:

$$\hat{H}^{(0)}\Psi_n^{(0)} = E_n^{(0)}\Psi_n^{(0)}. \tag{135}$$

Second, correct these eigenvalues and/or eigenfunctions to account for the perturbation's influence. Perturbation theory gives these corrections as an infinite series of terms, which become smaller and smaller for well-behaved systems:

$$E_n = E_n^{(0)} + E_n^{(1)} + E_n^{(2)} + \cdots \tag{136}$$

$$\Psi_n = \Psi_n^{(0)} + \Psi_n^{(1)} + \Psi_n^{(2)} + \cdots \tag{137}$$

Quite frequently, the corrections are only taken through first or second order (i.e., superscripts (1) or (2)). According to perturbation theory, the first-order correction to the energy is

$$E_n^{(1)} = \int \Psi_n^{(0)*} \hat{H}^{(1)} \Psi_n^{(0)}, \tag{138}$$

and the second-order correction is

$$E_n^{(2)} = \int \Psi_n^{(0)*} \hat{H}^{(1)} \Psi_n^{(1)}. \tag{139}$$

One can see that the first-order correction to the wavefunction, $\Psi_n^{(1)}$, seems to be needed to compute the second-order energy correction. However, it turns out that the correction $\Psi_n^{(1)}$ can be written in terms of the zeroth-order wavefunction as

$$\Psi_n^{(1)} = \sum_{i \neq n} \Psi_i^{(0)} \frac{\int \Psi_i^{(0)*} \hat{H}^{(1)} \Psi_n^{(0)}}{E_n^{(0)} - E_i^{(0)}}. \tag{140}$$

Substituting this in the expression for $E_n^{(2)}$, we obtain

$$E_n^{(2)} = \sum_{i \neq n} \frac{|\int \Psi_n^{(0)*} \hat{H}^{(1)} \Psi_i^{(0)}|^2}{E_n^{(0)} - E_i^{(0)}}. \tag{141}$$

Going back to the anharmonic oscillator example, the ground state wavefunction for the unperturbed problem is just (from section 5.2)

$$E_0^{(0)} = \frac{1}{2}\hbar\omega, \tag{142}$$

$$\Psi_0^{(0)}(x) = N_0 H_0(\alpha^{1/2} x) e^{-\alpha x^2/2} \tag{143}$$

$$= \left(\frac{\alpha}{\pi}\right)^{1/4} e^{-\alpha x^2/2}. \tag{144}$$

The first-order correction to the ground state energy would be

$$E_0^{(1)} = \left(\frac{\alpha}{\pi}\right)^{1/2} \int_{-\infty}^{\infty} \frac{1}{6}\gamma x^3 e^{-\alpha x^2} dx. \tag{145}$$

It turns out in this case that $E_0^{(1)} = 0$, since the integrand is odd. Does this mean that the anharmonic energy levels are the same as for the harmonic oscillator? No, because there are higher-order corrections such as $E_0^{(2)}$ which are not necessarily zero.

6.2 The Variational Method

The variational method is the other main approximate method used in quantum mechanics. Compared to perturbation theory, the variational method can be more robust in situations where it's hard to determine a good unperturbed Hamiltonian (i.e., one which makes the perturbation small but is still solvable). On the other hand, in cases where there is a good unperturbed Hamiltonian, perturbation theory can be more efficient than the variational method.

The basic idea of the variational method is to guess a "trial" wavefunction for the problem, which consists of some adjustable parameters called "variational parameters." These parameters are adjusted until the energy of the trial wavefunction is minimized. The resulting trial wavefunction and its corresponding energy are variational method approximations to the exact wavefunction and energy.

Why would it make sense that the best approximate trial wavefunction is the one with the lowest energy? This results from the Variational Theorem, which states that the energy of any trial wavefunction E is always an upper bound to the exact ground state energy $\mathcal{E}_0$. This can be proven easily. Let the trial wavefunction be denoted Φ. Any trial function can formally be expanded as a linear combination of the exact eigenfunctions Ψ_i. Of course, in practice, we don't know the Ψ_i, since we're assuming that we're applying the variational method to a problem we can't solve analytically. Nevertheless, that doesn't prevent us from

using the exact eigenfunctions in our proof, since they certainly exist and form a complete set, even if we don't happen to know them. So, the trial wavefunction can be written

$$\Phi = \sum_i c_i \Psi_i, \tag{146}$$

and the approximate energy corresponding to this wavefunction is

$$E[\Phi] = \frac{\int \Phi^* \hat{H} \Phi}{\int \Phi^* \Phi}. \tag{147}$$

Substituting the expansion over the exact wavefuntions,

$$E[\Phi] = \frac{\sum_{ij} c_i^* c_j \int \Psi_i^* \hat{H} \Psi_j}{\sum_{ij} c_i^* c_j \int \Psi_i^* \Psi_j}. \tag{148}$$

Since the functions Ψ_j are the exact eigenfunctions of $\hat{H}$, we can use $\hat{H}\Psi_j = \mathcal{E}_j \Psi_j$ to obtain

$$E[\Phi] = \frac{\sum_{ij} c_i^* c_j \mathcal{E}_j \int \Psi_i^* \Psi_j}{\sum_{ij} c_i^* c_j \int \Psi_i^* \Psi_j}. \tag{149}$$

Now using the fact that eigenfunctions of a Hermitian operator form an orthonormal set (or can be made to do so),

$$E[\Phi] = \frac{\sum_i c_i^* c_i \mathcal{E}_i}{\sum_i c_i^* c_i}. \tag{150}$$

We now subtract the exact ground state energy $\mathcal{E}_0$ from both sides to obtain

$$E[\Phi] - \mathcal{E}_0 = \frac{\sum_i c_i^* c_i (\mathcal{E}_i - \mathcal{E}_0)}{\sum_i c_i^* c_i}. \tag{151}$$

Since every term on the right-hand side is greater than or equal to zero, the left-hand side must also be greater than or equal to zero, or

$$E[\Phi] \geq \mathcal{E}_0. \tag{152}$$

In other words, the energy of any approximate wavefunction is always greater than or equal to the exact ground state energy $\mathcal{E}_0$. This explains the strategy of the

variational method: since the energy of any approximate trial function is always above the true energy, then any variations in the trial function which lower its energy are necessarily making the approximate energy closer to the exact answer. (The trial wavefunction is also a better approximation to the true ground state wavefunction as the energy is lowered, although not necessarily in every possible sense unless the limit $\Phi = \Psi_0$ is reached).

One example of the variational method would be using the Gaussian function $\phi(r) = e^{-\alpha r^2}$ as a trial function for the hydrogen atom ground state. This problem could be solved by the variational method by obtaining the energy of $\phi(r)$ as a function of the variational parameter α, and then minimizing $E(\alpha)$ to find the optimum value α_{min}. The variational theorem's approximate wavefunction and energy for the hydrogen atom would then be $\phi(r) = e^{-\alpha_{min} r^2}$ and $E(\alpha_{min})$.

Frequently, the trial function is written as a linear combination of basis functions, such as

$$\Phi = \sum_i c_i \phi_i. \tag{153}$$

This leads to the *linear variation method*, and the variational parameters are the expansion coefficients c_i. The energy for this approximate wavefunction is just

$$E[\Phi] = \frac{\sum_{ij} c_i^* c_j \int \phi_i^* \hat{H} \phi_j}{\sum_{ij} c_i^* c_j \int \phi_i^* \phi_j}, \tag{154}$$

which can be simplified using the notation

$$H_{ij} = \int \phi_i^* \hat{H} \phi_j, \tag{155}$$

$$S_{ij} = \int \phi_i^* \phi_j, \tag{156}$$

to yield

$$E[\Phi] = \frac{\sum_{ij} c_i^* c_j H_{ij}}{\sum_{ij} c_i^* c_j S_{ij}}. \tag{157}$$

Differentiating this energy with respect to the expansion coefficients c_i yields a

non-trivial solution only if the following "secular determinant" equals 0.

$$\begin{vmatrix} H_{11} - ES_{11} & H_{12} - ES_{12} & \cdots & H_{1N} - ES_{1N} \\ H_{21} - ES_{21} & H_{22} - ES_{22} & \cdots & H_{2N} - ES_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ H_{N1} - ES_{N1} & H_{N2} - ES_{N2} & \cdots & H_{NN} - ES_{NN} \end{vmatrix} = 0. \tag{158}$$

If an orthonormal basis is used, the secular equation is greatly simplified because S_{ij} is 1 for $i = j$ and 0 for $i \neq j$. In this case, the secular determinant is

$$\begin{vmatrix} H_{11} - E & H_{12} & \cdots & H_{1N} \\ H_{21} & H_{22} - E & \cdots & H_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ H_{N1} & H_{N2} & \cdots & H_{NN} - E \end{vmatrix} = 0. \tag{159}$$

In either case, the secular determinant for N basis functions gives an N-th order polynomial in E which is solved for N different roots, each of which approximates a different eigenvalue.

The variational method lies behind Hartree-Fock theory and the configuration interaction method for the electronic structure of atoms and molecules.

7 Molecular Quantum Mechanics

In this section, we discuss the quantum mechanics of atomic and molecular systems. We begin by writing the Hamiltonian for a collection of nuclei and electrons, and then we introduce the Born-Oppenheimer approximation, which allows us to separate the nuclear and electronic degrees of freedom.

7.1 The Molecular Hamiltonian

We have noted before that the kinetic energy for a system of particles is

$$\hat{T} = -\frac{\hbar^2}{2}\sum_i \frac{1}{m_i}\nabla^2 \tag{160}$$

The potential energy for a system of charged particles is

$$\hat{V}(\mathbf{r}) = \sum_{i>j} \frac{Z_i Z_j e^2}{4\pi\epsilon_0}\frac{1}{|\mathbf{r}_i - \mathbf{r}_j|} \tag{161}$$

For a molecule, it is reasonable to split the kinetic energy into two summations—one over electrons, and one over nuclei. Similarly, we can split the potential energy into terms representing interactions between nuclei, between electrons, or between electrons and nuclei. Using i and j to index electrons, and A and B to index nuclei, we have (in atomic units)

$$\hat{H} = -\sum_A \frac{1}{2M_A}\nabla_A^2 - \sum_i \frac{1}{2}\nabla_i^2 + \sum_{A>B}\frac{Z_A Z_B}{R_{AB}} - \sum_{Ai}\frac{Z_A}{r_{Ai}} + \sum_{i>j}\frac{1}{r_{ij}} \tag{162}$$

where $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$, $R_{Ai} = |\mathbf{r}_A - \mathbf{r}_i|$, and $R_{AB} = |\mathbf{r}_A - \mathbf{r}_B|$. This is known as the "exact" nonrelativistic Hamiltonian in field-free space. However, it is important to remember that this Hamiltonian neglects at least two effects. Firstly, although the speed of an electron in a hydrogen atom is less than 1% of the speed of light, relativistic mass corrections can become appreciable for the inner electrons of heavier atoms. Secondly, we have neglected the spin-orbit effects. From the point

of view of an electron, it is being orbited by a nucleus which produces a magnetic field (proportional to L); this field interacts with the electron's magnetic moment (proportional to S), giving rise to a spin-orbit interaction (proportional to $\mathbf{L} \cdot \mathbf{S}$ for a diatomic.) Although spin-orbit effects can be important, they are generally neglected in quantum chemical calculations.

7.2 The Born-Oppenheimer Approximation

We know that if a Hamiltonian is separable into two or more terms, then the total eigenfunctions are products of the individual eigenfunctions of the separated Hamiltonian terms, and the total eigenvalues are sums of individual eigenvalues of the separated Hamiltonian terms.

Consider, for example, a Hamiltonian which is separable into two terms, one involving coordinate q_1 and the other involving coordinate q_2.

$$\hat{H} = \hat{H}_1(q_1) + \hat{H}_2(q_2) \tag{163}$$

with the overall Schrödinger equation being

$$\hat{H}\psi(q_1, q_2) = E\psi(q_1, q_2) \tag{164}$$

If we assume that the total wavefunction can be written in the form $\psi(q_1, q_2) = \psi_1(q_1)\psi_2(q_2)$, where $\psi_1(q_1)$ and $\psi_2(q_2)$ are eigenfunctions of $\hat{H}_1$ and $\hat{H}_2$ with eigenvalues E_1 and E_2, then

$$\begin{aligned} \hat{H}\psi(q_1, q_2) &= (\hat{H}_1 + \hat{H}_2)\psi_1(q_1)\psi_2(q_2) \\ &= \hat{H}_1\psi_1(q_1)\psi_2(q_2) + \hat{H}_2\psi_1(q_1)\psi_2(q_2) \\ &= E_1\psi_1(q_1)\psi_2(q_2) + E_2\psi_1(q_1)\psi_2(q_2) \\ &= (E_1 + E_2)\psi_1(q_1)\psi_2(q_2) \\ &= E\psi(q_1, q_2) \end{aligned} \tag{165}$$

Thus the eigenfunctions of $\hat{H}$ are products of the eigenfunctions of $\hat{H}_1$ and $\hat{H}_2$, and the eigenvalues are the sums of eigenvalues of $\hat{H}_1$ and $\hat{H}_2$.

If we examine the nonrelativistic Hamiltonian (162), we see that the term

$$\sum_{Ai} \frac{Z_A}{r_{Ai}} \tag{166}$$

prevents us from cleanly separating the electronic and nuclear coordinates and writing the total wavefunction as $\psi(\mathbf{r}, \mathbf{R}) = \psi_e(\mathbf{r})\psi_N(\mathbf{R})$, where $\mathbf{r}$ represents the set of all electronic coordinates, and $\mathbf{R}$ represents the set of all nuclear coordinates. The Born-Oppenheimer approximation is to assume that this separation is nevertheless *approximately* correct.

Qualitatively, the Born-Oppenheimer approximation rests on the fact that the nuclei are much more massive than the electrons. This allows us to say that the nuclei are nearly fixed with respect to electron motion. We can fix $\mathbf{R}$, the nuclear configuration, at some value $\mathbf{R_a}$, and solve for $\psi_e(\mathbf{r}; \mathbf{R_a})$; the electronic wavefunction depends only parametrically on $\mathbf{R}$. If we do this for a range of $\mathbf{R}$, we obtain the potential energy curve along which the nuclei move.

We now show the mathematical details. Let us abbreviate the molecular Hamiltonian as

$$\hat{H} = \hat{T}_N(\mathbf{R}) + \hat{T}_e(\mathbf{r}) + \hat{V}_{NN}(\mathbf{R}) + \hat{V}_{eN}(\mathbf{r}, \mathbf{R}) + \hat{V}_{ee}(\mathbf{r}) \tag{167}$$

where the meaning of the individual terms should be obvious. Initially, $\hat{T}_N(\mathbf{R})$ can be neglected since $\hat{T}_N$ is smaller than $\hat{T}_e$ by a factor of M_A/m_e, where m_e is the mass of an electron. Thus for a *fixed* nuclear configuration, we have

$$\hat{H}_{el} = \hat{T}_e(\mathbf{r}) + \hat{V}_{eN}(\mathbf{r}; \mathbf{R}) + \hat{V}_{NN}(\mathbf{R}) + \hat{V}_{ee}(\mathbf{r}) \tag{168}$$

such that

$$\hat{H}_{el}\phi_e(\mathbf{r}; \mathbf{R}) = E_{el}\phi_e(\mathbf{r}; \mathbf{R}) \tag{169}$$

This is the "clamped-nuclei" Schrödinger equation. Quite frequently $\hat{V}_{NN}(\mathbf{R})$ is neglected in the above equation, which is justified since in this case $\mathbf{R}$ is just a parameter so that $\hat{V}_{NN}(\mathbf{R})$ is just a constant and shifts the eigenvalues only by some constant amount. Leaving $\hat{V}_{NN}(\mathbf{R})$ out of the electronic Schrödinger equation leads to a similar equation,

$$\hat{H}_e = \hat{T}_e(\mathbf{r}) + \hat{V}_{eN}(\mathbf{r}; \mathbf{R}) + \hat{V}_{ee}(\mathbf{r}) \tag{170}$$

$$\hat{H}_e\phi_e(\mathbf{r};\mathbf{R}) = E_e\phi_e(\mathbf{r};\mathbf{R}) \tag{171}$$

where we have used a new subscript "e" on the electronic Hamiltonian and energy to distinguish from the case where $\hat{V}_{NN}$ is included.

We now consider again the original Hamiltonian (167). If we insert a wavefunction of the form $\phi_T(\mathbf{r},\mathbf{R}) = \phi_e(\mathbf{r};\mathbf{R})\phi_N(\mathbf{R})$, we obtain

$$\hat{H}\phi_e(\mathbf{r};\mathbf{R})\phi_N(\mathbf{R}) = E_{tot}\phi_e(\mathbf{r};\mathbf{R})\phi_N(\mathbf{R}) \tag{172}$$

$$\{\hat{T}_N(\mathbf{R})+\hat{T}_e(\mathbf{r})+\hat{V}_{eN}(\mathbf{r},\mathbf{R})+\hat{V}_{NN}(\mathbf{R})+\hat{V}_{ee}(\mathbf{r})\}\phi_e(\mathbf{r};\mathbf{R})\phi_N(\mathbf{R}) = E_{tot}\phi_e(\mathbf{r};\mathbf{R})\phi_N(\mathbf{R}) \tag{173}$$

Since $\hat{T}_e$ contains no $\mathbf{R}$ dependence,

$$\hat{T}_e\phi_e(\mathbf{r};\mathbf{R})\phi_N(\mathbf{R}) = \phi_N(\mathbf{R})\hat{T}_e\phi_e(\mathbf{r};\mathbf{R}) \tag{174}$$

However, we may not immediately assume

$$\hat{T}_N\phi_e(\mathbf{r};\mathbf{R})\phi_N(\mathbf{R}) = \phi_e(\mathbf{r};\mathbf{R})\hat{T}_N\phi_N(\mathbf{R}) \tag{175}$$

(this point is tacitly assumed by most introductory textbooks). By the chain rule,

$$\nabla_A^2\phi_e(\mathbf{r};\mathbf{R})\phi_N(\mathbf{R}) = \phi_e(\mathbf{r};\mathbf{R})\nabla_A^2\phi_N(\mathbf{R})+2\nabla_A\phi_e(\mathbf{r};\mathbf{R})\nabla_A\phi_N(\mathbf{R})+\phi_N(\mathbf{R})\nabla_A^2\phi_e(\mathbf{r};\mathbf{R}) \tag{176}$$

Using these facts, along with the electronic Schrödinger equation,

$$\{\hat{T}_e + \hat{V}_{eN}(\mathbf{r};\mathbf{R}) + \hat{V}_{ee}(\mathbf{r})\}\phi_e(\mathbf{r};\mathbf{R}) = \hat{H}_e\phi_e(\mathbf{r};\mathbf{R}) = E_e\phi_e(\mathbf{r};\mathbf{R}) \tag{177}$$

we simplify (173) to

$$\begin{aligned} &\phi_e(\mathbf{r};\mathbf{R})\hat{T}_N\phi_N(\mathbf{R}) + \phi_N(\mathbf{R})\phi_e(\mathbf{r};\mathbf{R})(E_e + \hat{V}_{NN}) \\ -&\left\{\sum_A \frac{1}{2M_A}(2\nabla_A\phi_e(\mathbf{r};\mathbf{R})\nabla_A\phi_N(\mathbf{R}) + \phi_N(\mathbf{R})\nabla_A^2\phi_e(\mathbf{r};\mathbf{R}))\right\} \\ &= E_{tot}\phi_e(\mathbf{r};\mathbf{R})\phi_N(\mathbf{R}) \end{aligned} \tag{178}$$

We must now estimate the magnitude of the last term in brackets. Following Steinfeld [5], a typical contribution has the form $1/(2M_A)\nabla_A^2\phi_e(\mathbf{r};\mathbf{R})$, but

$\nabla_A \phi_e(\mathbf{r};\mathbf{R})$ is of the same order as $\nabla_i \phi_e(\mathbf{r};\mathbf{R})$ since the derivatives operate over approximately the same dimensions. The latter is $\phi_e(\mathbf{r};\mathbf{R})p_e$, with p_e the momentum of an electron. Therefore $1/(2M_A)\nabla_A^2 \phi_e(\mathbf{r};\mathbf{R}) \approx p_e^2/(2M_A) = (m/M_A)E_e$. Since $m/M_A \sim 1/10000$, the term in brackets can be dropped, giving

$$\phi_e(\mathbf{r};\mathbf{R})\hat{T}_N\phi_N(\mathbf{R}) + \phi_N(\mathbf{R})E_e\phi_e(\mathbf{r};\mathbf{R}) + \phi_N(\mathbf{R})\hat{V}_{NN}\phi_e(\mathbf{r};\mathbf{R}) = E_{tot}\phi_e(\mathbf{r};\mathbf{R})\phi_N(\mathbf{R}) \tag{179}$$

$$\{\hat{T}_N + E_e + \hat{V}_{NN}\}\phi_N(\mathbf{R}) = E_{tot}\phi_N(\mathbf{R}) \tag{180}$$

This is the nuclear Shrodinger equation we anticipated—the nuclei move in a potential set up by the electrons.

To summarize, the large difference in the relative masses of the electrons and nuclei allows us to approximately separate the wavefunction as a product of nuclear and electronic terms. The electronic wavefucntion $\phi_e(\mathbf{r};\mathbf{R})$ is solved for a given set of nuclear coordinates,

$$\hat{H}_e\phi_e(\mathbf{r};\mathbf{R}) = \left\{-\frac{1}{2}\sum_i \nabla_i^2 - \sum_{A,i}\frac{Z_A}{r_{Ai}} + \sum_{i>j}\frac{1}{r_{ij}}\right\}\phi_e(\mathbf{r};\mathbf{R}) = E_e(\mathbf{R})\phi_e(\mathbf{r};\mathbf{R}) \tag{181}$$

and the electronic energy obtained contributes a potential term to the motion of the nuclei described by the nuclear wavefunction $\phi_N(\mathbf{R})$.

$$\hat{H}_N\phi_N(\mathbf{R}) = \left\{-\sum_A \frac{1}{2M_A}\nabla_A^2 + E_e(\mathbf{R}) + \sum_{A>B}\frac{Z_A Z_B}{R_{AB}}\right\}\phi_N(\mathbf{R}) = E_{tot}\phi_N(\mathbf{R}) \tag{182}$$

As a final note, many textbooks, including Szabo and Ostlund [4], mean total energy *at fixed geometry* when they use the term "total energy" (i.e., they neglect the nuclear kinetic energy). This is just E_{el} of equation (169), which is also E_e plus the nuclear-nuclear repulsion. A somewhat more detailed treatment of the Born-Oppenheimer approximation is given elsewhere [6].

7.3 Separation of the Nuclear Hamiltonian

The nuclear Schrödinger equation can be approximately factored into translational, rotational, and vibrational parts. McQuarrie [1] explains how to do this

for a diatomic in section 10-13. The rotational part can be cast into the form of the rigid rotor model, and the vibrational part can be written as a system of harmonic oscillators. Time does not allow further comment on the nuclear Schrödinger equation, although it is central to molecular spectroscopy.

8 Solving the Electronic Eigenvalue Problem

Once we have invoked the Born-Oppenheimer approximation, we attempt to solve the electronic Schrödinger equation (171), i.e.

$$\left[-\frac{1}{2}\sum_i \nabla_i^2 - \sum_{iA}\frac{Z_A}{r_{iA}} + \sum_{i>j}\frac{1}{r_{ij}}\right]\psi_e(\mathbf{r};\mathbf{R}) = E_e\psi_e(\mathbf{r};\mathbf{R}) \tag{183}$$

But, as mentioned previously, this equation is quite difficult to solve!

8.1 The Nature of Many-Electron Wavefunctions

Let us consider the nature of the electronic wavefunctions $\psi_e(\mathbf{r};\mathbf{R})$. Since the electronic wavefunction depends only parametrically on $\mathbf{R}$, we will suppress $\mathbf{R}$ in our notation from now on. What do we require of $\psi_e(\mathbf{r})$? Recall that $\mathbf{r}$ represents the set of all electronic coordinates, i.e., $\mathbf{r} = \{\mathbf{r}_1, \mathbf{r}_2, \ldots \mathbf{r}_N\}$. So far we have left out one important item—we need to include the *spin* of each electron. We can define a new variable $\mathbf{x}$ which represents the set of all four coordinates associated with an electron: three spatial coordinates $\mathbf{r}$, and one spin coordinate ω, i.e., $\mathbf{x} = \{\mathbf{r}, \omega\}$.

Thus we write the electronic wavefunction as $\psi_e(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$. Why have we been able to avoid including spin until now? Because the non-relativistic Hamiltonian does not include spin. Nevertheless, spin must be included so that the electronic wavefunction can satisfy a very important requirement, which is the *antisymmetry principle* (see Postulate 6 in Section 4). This principle states that for a system of fermions, the wavefunction must be antisymmetric with respect to the interchange of all (space *and* spin) coordinates of one fermion with those of another. That is,

$$\psi_e(\mathbf{x}_1, \ldots, \mathbf{x}_a, \ldots, \mathbf{x}_b, \ldots, \mathbf{x}_N) = -\psi_e(\mathbf{x}_1, \ldots, \mathbf{x}_b, \ldots, \mathbf{x}_a, \ldots, \mathbf{x}_N) \tag{184}$$

The Pauli exclusion principle is a direct consequence of the antisymmetry principle.

A very important step in simplifying $\psi_e(\mathbf{x})$ is to expand it in terms of a set of one-electron functions, or "orbitals." This makes the electronic Schrödinger equation considerably easier to deal with.[3] A *spin orbital* is a function of the space and spin coordinates of a single electron, while a *spatial* orbital is a function of a single electron's spatial coordinates only. We can write a spin orbital as a product of a spatial orbital one of the two spin functions

$$\chi(\mathbf{x}) = \psi(\mathbf{r})|\alpha\rangle \tag{185}$$

or

$$\chi(\mathbf{x}) = \psi(\mathbf{r})|\beta\rangle \tag{186}$$

Note that for a given spatial orbital $\psi(\mathbf{r})$, we can form *two* spin orbitals, one with α spin, and one with β spin. The spatial orbital will be doubly occupied. It is possible (although sometimes frowned upon) to use one set of spatial orbitals for spin orbitals with α spin and another set for spin orbitals with β spin.[4]

Where do we get the one-particle spatial orbitals $\psi(\mathbf{r})$? That is beyond the scope of the current section, but we briefly itemize some of the more common possibilities:

- Orbitals centered on each atom (atomic orbitals).
- Orbitals centered on each atom but also symmetry-adapted to have the correct point-group symmetry species (symmetry orbitals).
- Molecular orbitals obtained from a Hartree-Fock procedure.

We now explain how an N-electron function $\psi_e(\mathbf{x})$ can be constructed from spin orbitals, following the arguments of Szabo and Ostlund [4] (p. 60). Assume we have a complete set of functions of a single variable $\{\chi_i(x)\}$. Then any function of a single variable can be expanded exactly as

$$\Phi(x_1) = \sum_i a_i \chi_i(x_1). \tag{187}$$

[3] It is not completely *necessary* to do this, however; for example, the Hylleras treatment of the Helium atom uses two-particle basis functions which are not further expanded in terms of single-particle functions.

[4] This is the procedure of the Unrestricted Hartree Fock (UHF) method.

How can we expand a function of *two* variables, e.g. $\Phi(x_1, x_2)$?

If we hold x_2 fixed, then

$$\Phi(x_1, x_2) = \sum_i a_i(x_2)\chi_i(x_1). \tag{188}$$

Now note that each expansion coefficient $a_i(x_2)$ is a function of a single variable, which can be expanded as

$$a_i(x_2) = \sum_j b_{ij}\chi_j(x_2). \tag{189}$$

Substituting this expression into the one for $\Phi(x_1, x_2)$, we now have

$$\Phi(x_1, x_2) = \sum_{ij} b_{ij}\chi_i(x_1)\chi_j(x_2) \tag{190}$$

a process which can obviously be extended for $\Phi(x_1, x_2, \ldots, x_N)$.

We can extend these arguments to the case of having a complete set of functions of the variable $\mathbf{x}$ (recall $\mathbf{x}$ represents x, y, and z and also ω). In that case, we obtain an analogous result,

$$\Phi(\mathbf{x}_1, \mathbf{x}_2) = \sum_{ij} b_{ij}\chi_i(\mathbf{x}_1)\chi_j(\mathbf{x}_2) \tag{191}$$

Now we must make sure that the antisymmetry principle is obeyed. For the two-particle case, the requirement

$$\Phi(\mathbf{x}_1, \mathbf{x}_2) = -\Phi(\mathbf{x}_2, \mathbf{x}_1) \tag{192}$$

implies that $b_{ij} = -b_{ji}$ and $b_{ii} = 0$, or

$$\begin{aligned} \Phi(\mathbf{x}_1, \mathbf{x}_2) &= \sum_{j>i} b_{ij}[\chi_i(\mathbf{x}_1)\chi_j(\mathbf{x}_2) - \chi_j(\mathbf{x}_1)\chi_i(\mathbf{x}_2)] \\ &= \sum_{j>i} b_{ij}|\chi_i\chi_j\rangle \end{aligned} \tag{193}$$

where we have used the symbol $|\chi_i\chi_j\rangle$ to represent a *Slater determinant*, which in the genreral case is written

$$|\chi_1\chi_2\cdots\chi_N\rangle = \frac{1}{\sqrt{N!}}\begin{vmatrix} \chi_1(\mathbf{x}_1) & \chi_2(\mathbf{x}_1) & \cdots & \chi_N(\mathbf{x}_1) \\ \chi_1(\mathbf{x}_2) & \chi_2(\mathbf{x}_2) & \cdots & \chi_N(\mathbf{x}_2) \\ \vdots & \vdots & & \vdots \\ \chi_1(\mathbf{x}_N) & \chi_2(\mathbf{x}_N) & \cdots & \chi_N(\mathbf{x}_N) \end{vmatrix} \tag{194}$$

We can extend the reasoning applied here to the case of N electrons; any N-electron wavefunction can be expressed exactly as a linear combination of all possible N-electron Slater determinants formed from a complete set of spin orbitals $\{\chi_i(\mathbf{x})\}$.

8.2 Matrix Mechanics

As we mentioned previously in section 2, Heisenberg's matrix mechanics, although little-discussed in elementary textbooks on quantum mechanics, is nevertheless formally equivalent to Schrödinger's wave equations. Let us now consider how we might solve the time-independent Schrödinger equation in matrix form.

If we want to solve $\hat{H}\psi_e(\mathbf{x}) = E_e\psi_e(\mathbf{x})$ as a matrix problem, we need to find a suitable linear vector space. Now $\psi_e(\mathbf{x})$ is an N-electron function that must be antisymmetric with respect to interchange of electronic coordinates. As we just saw in the previous section, any such N-electron function can be expressed *exactly* as a linear combination of Slater determinants, within the space spanned by the set of orbitals $\{\chi(\mathbf{x})\}$. If we denote our Slater determinant basis functions as $|\Phi_i\rangle$, then we can express the eigenvectors as

$$|\Psi_i\rangle = \sum_j^I c_{ij}|\Phi_j\rangle \tag{195}$$

for I possible N-electron basis functions (I will be infinite if we actually have a complete set of one electron functions χ). Similarly, we construct the matrix $\mathbf{H}$ in this basis by $H_{ij} = \langle\Phi_i|H|\Phi_j\rangle$.

If we solve this matrix equation, $\mathbf{H}|\Psi_n\rangle = E_n|\Psi_n\rangle$, in the space of all possible Slater determinants as just described, then the procedure is called *full configuration-interaction*, or full CI. A full CI constitues the *exact* solution to the time-independent Schrödinger equation within the given space of the spin orbitals χ. If we restrict the N-electron basis set in some way, then we will solve Schrödinger's equation *approximately*. The method is then called "configuration interaction," where we have dropped the prefix "full." For more information on configuration interaction, see the lecture notes by the present author [7] or one of the available review articles [8, 9].

Module 2 Spectroscopic techniques

Lecture 3 Basics of Spectroscopy

Spectroscopy deals with the study of interaction of electromagnetic radiation with matter. Electromagnetic radiation is a simple harmonic wave of electric and magnetic fields fluctuating orthogonal to each other (Figure 3.1A).

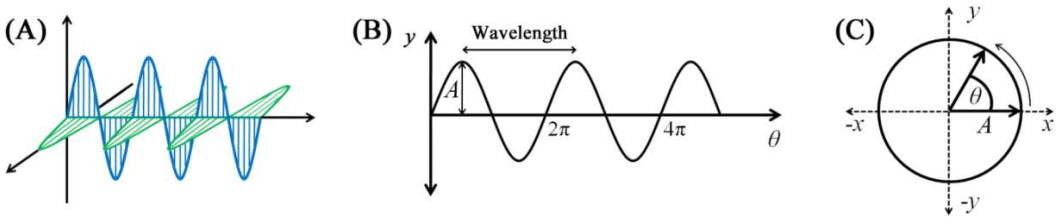


Figure 3.1: An electromagnetic wave showing orthogonal electric and magnetic components (A); a sine wave (B); and uniform circular motion representation of the sine function (C).

A simple harmonic function can be represented by a sine wave (Figure 3.1B):

$$y = A \sin \theta \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (3.1)$$

Sine wave is a periodic function and can be described in terms of the circular motion (Figure 3.1C). The value of *y* at any point is simply the projection of vector *A* on the y-axis, which is nothing but *A sinθ*. Equation (1) can therefore be written in terms of angular velocity, ω.

$$y = A \sin(\omega t) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (3.2)$$

$$y = A \sin(2\pi \nu t) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (3.3)$$

$$y = A \sin(2\pi \nu \frac{z}{c}) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (3.4)$$

where, *z* = displacement in time *t* and *c* is the velocity of the electromagnetic wave

If the wave completes ν cycles/s and the wave is travelling with a velocity *c* metres/ sec, then the wavelength of the wave must be $\frac{c}{\nu}$ metres.

$$y = A \sin(\frac{2\pi z}{\lambda}) \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (3.5)$$

Energy of electromagnetic radiation:

Energy of an electromagnetic radiation is given by

$$E = h\nu = h\frac{c}{\lambda} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (3.6)$$

where *h* is Planck's constant and has a value of 6.626×10^{-34} $m^2 \cdot kg \cdot s^{-1}$. Based on the energy, electromagnetic radiation has been divided into different regions. The region of electromagnetic spectrum human beings can see, for example, is called visible region or visible spectrum. The visible region constitutes a very small portion of the electromagnetic spectrum and corresponds to the wavelengths of ~400 – 780 nm (Figure 3.2). The energy of the visible spectrum therefore ranges from $\sim 2.5 \times 10^{-19}$ to $\sim 5 \times 10^{-19}$ Joules. It is not convenient to write such small values of energy; the energies are therefore written in terms of electronvolts (eV). One electronvolt equals 1.602×10^{-19} Joules. Therefore, the energy range of the visible spectrum is ~1.6 – 3.1 eV. Spectroscopists, however, prefer to use wavelength (λ) or frequency (ν) or wavenumber ($\bar{\nu}$) instead of energy.

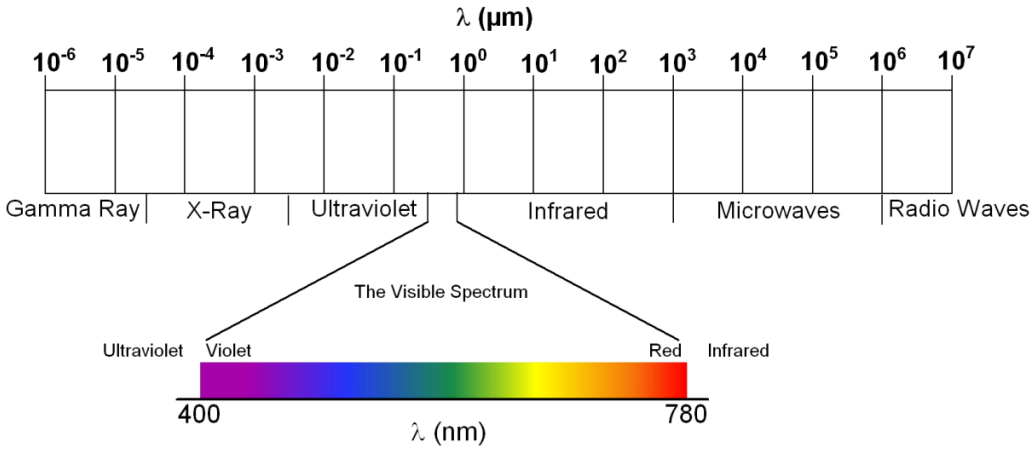


Figure 3.2 Electromagnetic spectrum

Quantization of energy:

As put forward by Max Planck while studying the problem of Blackbody radiation in early 1900s, atoms and molecules can absorb or emit the energy in discrete packets, called quanta (singular: quantum). The quantum for electromagnetic energy is called a photon which has the energy given by equation 3.6. A molecule can possess energies

in different forms such as vibrational energy, rotational energy, electronic energy, etc. Introduction to the structure of an atom in a General Chemistry course mentions about the electrons residing in different orbits/orbitals surrounding the nucleus, typically the first exposure to the discrete electronic energy levels of atoms. In much the same way, rotational and vibrational energy levels of molecules are also discrete. A molecule can jump from one energy level to another by absorbing or emitting a photon of energy that separate the two energy levels (Figure 3.3).

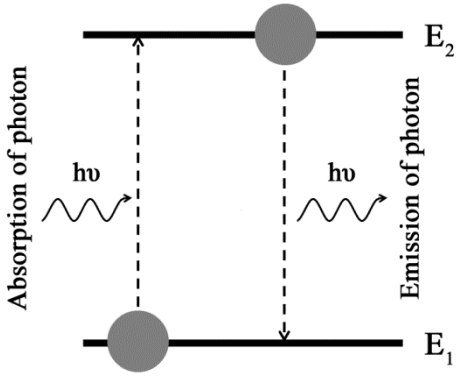


Figure 3.3 Transitions of a molecule between energy levels, E_1 and E_2 by absorbing/emitting the electromagnetic radiation

Electromagnetic spectrum and the atomic/molecular processes:

Molecules undergo processes like rotation, vibration, electronic transitions, and nuclear transitions. The energies underlying these processes correspond to different regions in the electromagnetic spectrum (Figure 3.4):

i. Radiofrequency waves: Radiofrequency region has very low energies that correspond to the energy differences in the nuclear and electron spin states. These frequencies, therefore, find applications in nuclear magnetic resonance and electron paramagnetic resonance spectroscopy.

ii. Microwaves: Microwaves have energies between those of radiofrequency waves and infrared waves and find applications in rotational spectroscopy and electron paramagnetic resonance spectroscopy.

iii. Infrared radiation: The energies associated with molecular vibrations fall in the infrared region of electromagnetic spectrum. Infrared spectroscopy is therefore also known as vibrational spectroscopy and is a very useful technique for functional group identification in organic compounds.

iv. UV/Visible region: UV and visible regions are involved in the electronic transitions in the molecules. The spectroscopic methods using UV or visible light therefore come under 'Electronic spectroscopy'.

v. X-ray radiation: X-rays are high energy electromagnetic radiation and causes transitions in the internal electrons of the molecules.

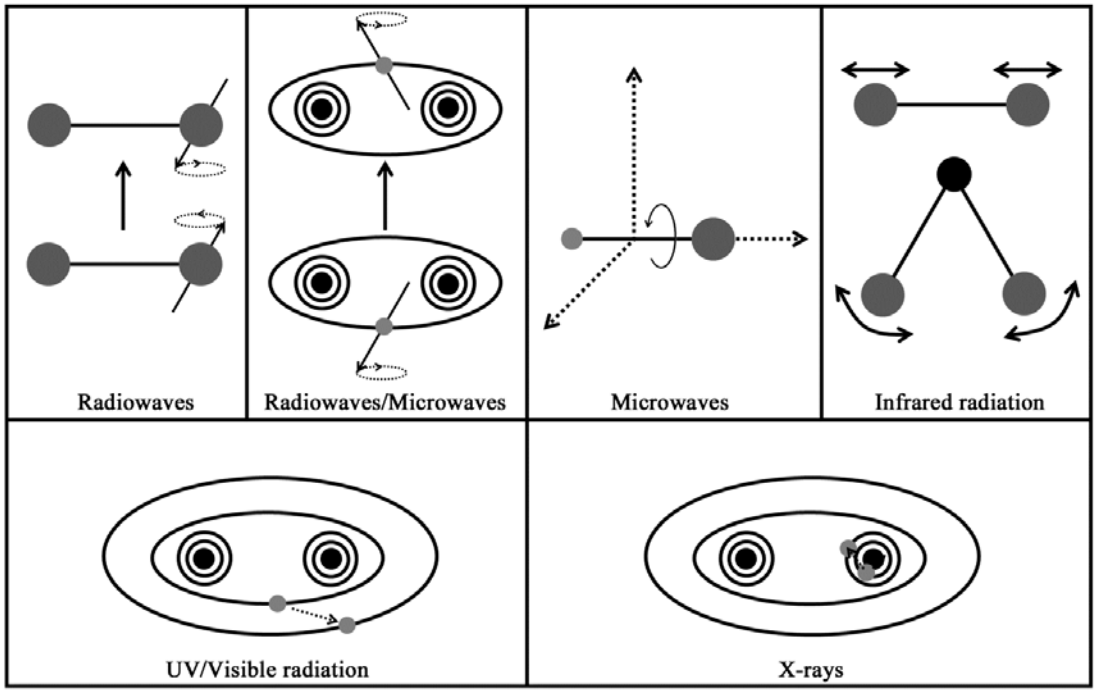


Figure 3.4 The range of atomic/molecular processes the electromagnetic radiation is involved in.

Mechanisms of interaction of electromagnetic radiation with matter:

In order to interact with the electromagnetic radiation, the molecules must have some electric or magnetic effect that could be influenced by the electric or magnetic components of the radiation.

i. In NMR spectroscopy, for example, the nuclear spins have magnetic dipoles aligned with or against a huge magnetic field. Interaction with radiofreqency of appropriate energy results in the change in these dipoles.

ii. Rotations of a molecule having a net electric dipole moment, such as water will cause changes in the directions of the dipole and therefore in the electrical properties (Figure 3.5A and B). Figure 3.5B shows the changes in the y-component of the dipole moment due to rotation of water molecule.

iii. Vibrations of molecules can result in changes in electric dipoles that could interact with the electrical component of the electromagnetic radiation (Figure 3.5C).

iv. Electronic transitions take place from one orbital to another. Owing to the differences in the geometry, size, and the spatial organization of the different orbitals, an electronic transition causes change in the dipole moment of the molecule (Figure 3.5D).

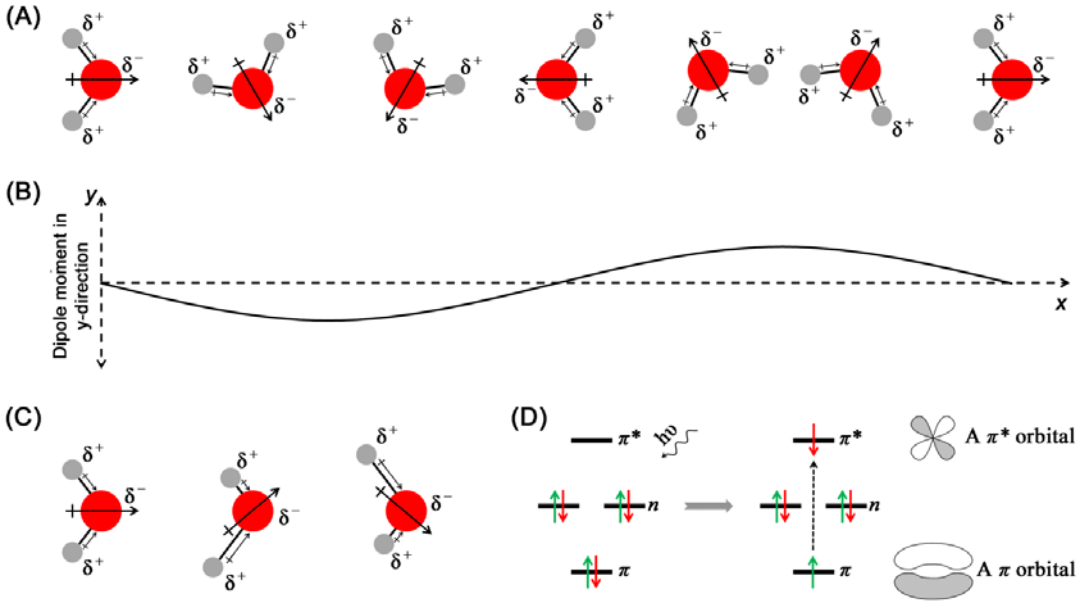


Figure 3.5 Panel A shows the rotation of a water molecule around its centre of mass (A). The change in the dipole moment as a result of rotation in plotted in panel B. Panel C shows the change in dipole moment of water due to asymmetric stretching vibrations of O—H bond. Panel D shows an electronic transition from π to π^* orbital and the geometry of the two orbitals.

The above examples suggest that a change in either electric or magnetic dipole moment in a molecule is required for the absorption or emission of the electromagnetic radiation.

THINK TANK??

Will there be a change in the dipole moment if there is symmetric stretching of O—H bond in water molecule?

Absorption peaks and line widths:

Absorption of radiation is the first step in any spectroscopic experiment. Absorption spectra are routinely recorded for the electronic, rotational, and vibrational spectroscopy. It is therefore important to see how an absorption spectrum looks like.

As we have already seen, a transition between states takes place if the energy provided by the electromagnetic radiation equals the energy gap between the two states *i.e.* $\Delta E = h\upsilon = \frac{hc}{\lambda}$. This implies that the molecule precisely absorbs the radiation of wavelength, λ and ideally a sharp absorption line should appear at this wavelength (Figure 3.6A). In practice, however, the absorption lines are not sharp but appear as fairly broad peaks (Figure 3.6B) for the following reasons:

i. Instrumental factors: The slits that allow the incident light to impinge on the sample and the emerging light to the detector have finite widths. Consider that the transition occurs at wavelength, λ_t. When the wavelength is changed to $\lambda_t + \Delta\lambda$ or $\lambda_t - \Delta\lambda$, the finite slit width allows the radiation of wavelength, λ_t to pass through the slits and a finite absorbance is observed at these wavelengths. The absorption peaks are therefore symmetrical to the line at $\lambda = \lambda_t$.

ii. Sample factors: Molecules in a liquid or gaseous sample are in motion and keep colliding with each other. Collisions influence the vibrational and rotational motions of the molecules thereby causing broadening. Two atoms/molecules coming in close proximity will perturb the electronic energies, at least those of the outermost electrons resulting in broadening of electronic spectra. Motion of molecules undergoing transition also causes shift in absorption frequencies, known as Doppler broadening.

iii. Intrinsic broadening: Intrinsic or natural broadening arises from the Heisenberg's uncertainty principle which states that the shorter the lifetime of a state, the more uncertain is its energy. Molecular transitions have finite lifetimes, therefore their energy is not exact. If Δt is the lifetime of a molecule in an excited state, the uncertainty in the energy of the states is given by:

$$\Delta E \times \Delta t \geq \frac{h}{4\pi} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (3.7)$$

$$\Delta E \times \Delta t \geq \frac{\hbar}{2} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (3.8)$$

where, $\hbar = \frac{h}{2\pi}$

Two more features worth noticing in the Figure 3.6B are the fluctuations in the baseline and the baseline itself, which is not horizontal. The small fluctuations in the baseline are referred to as noise. Noise is the manifestation of the random weak

signals generated by the instrument electronics. To identify the sample peaks clear of the noise, the intensity of the sample peaks has to be at least 3-4 times higher than the noise. A better signal-to-noise ratio is obtained by recorded more than one spectra and averaging; the noise being random gets cancelled out. Instrumental factors are responsible for the non-horizontal baseline observed in Figure 3.6B: The light sources used in the instruments emit radiations of different intensities at different wavelengths and usually the detector sensitivity is also wavelength-dependent. A reasonable horizontal baseline for the samples can easily be obtained by subtracting the spectrum obtained from the solvent the sample is dissolved in.

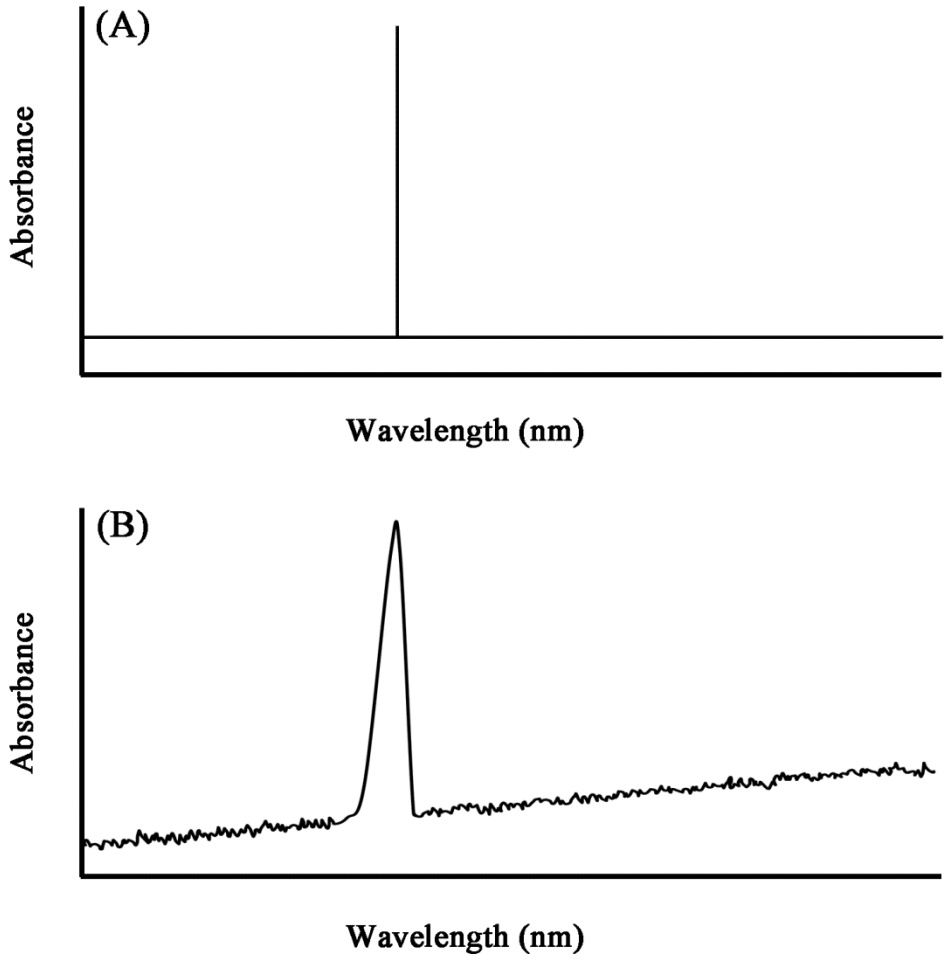


Figure 3.6 An idealized spectrum for a single wavelength transition (A) and an experimentally obtained spectrum (B)

Other features of the spectroscopy and the spectra obtained will be discussed as and when they arise in the following lectures.

Lecture4 UV/Visible Absorption Spectroscopy-I

We see a lot of colorful things around us. What exactly is the color and what make the things exhibit these colors? We know that the color we see is the visible region of the electromagnetic spectrum. We also know that matter can absorb the electromagnetic radiation of different energy (or wavelengths). The region of electromagnetic energy that is not absorbed is simply reflected back or getstransmitted through the matter. The colored compounds are colored because they absorb the visible light. The color that is perceived is called the complement color to the absorbed wavelength and is represented by a color wheel (Figure 4.1).

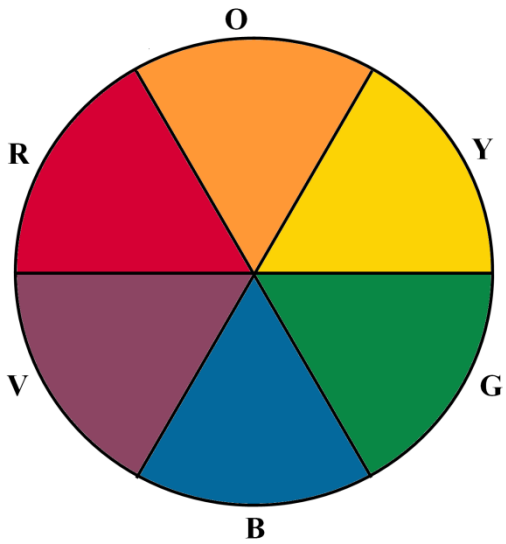


Figure 4.1 A simplified color wheel showing complementary colors. Green is interesting as it can arise from the absorption of radiation to either end of the visible spectrum.

Absorption of ultraviolet (UV) and visible radiation is one of the most routinely used analytical tools in life sciences research. The simplest application of UV/Visible radiation is to quantify the amount of a substance present in a solution.UV region of electromagnetic radiation encompasses the wavelengths ranging from ~10 nm – ~400 nm while visible region encompasses the wavelengths from ~400 nm – ~780 nm. For the sake of convenience in discussing the observations, UV region is loosely divided into near UV (wavelength region nearer to the visible region, λ ~ 250 nm – 400 nm), far UV region (wavelength region farther to the visible region, λ ~ 190 nm – 250 nm) and vacuum UV region (λ< 190 nm). The wavelength ranges defined for these regions are not strict and people use slightly different ranges to define these regions. We shall, however, stick to the wavelengths defined here. As has been discussed in the previous lecture, the absorption of UV and visible light is through the transition of an electron in the molecule from lower to a higher energy molecular orbital. The various electronic transitions observed in organic compound are shown in Figure 4.2.

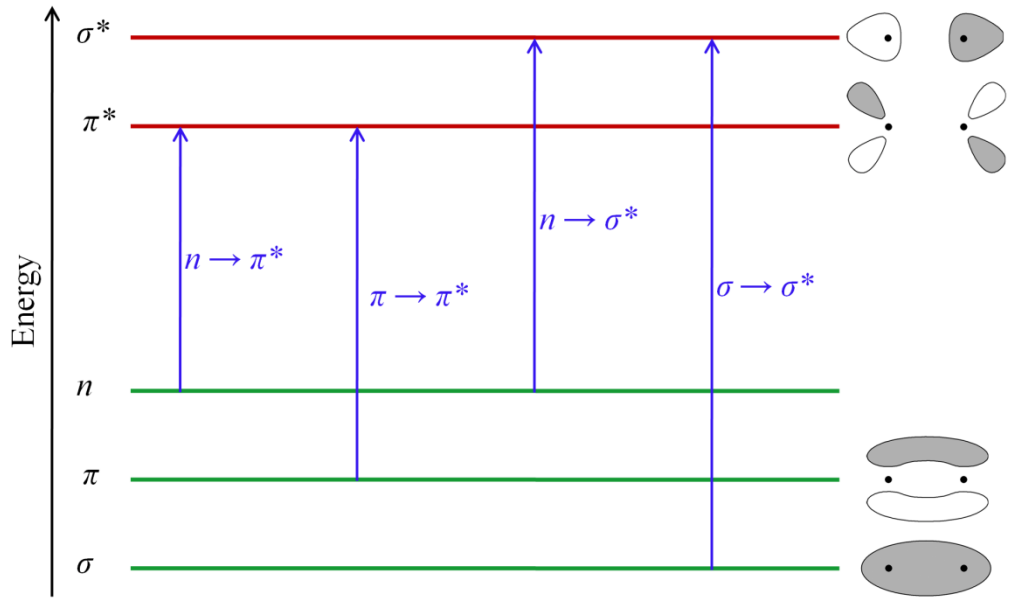


Figure 4.2 Schematic diagram showing energy levels of different orbitals and possible absorption transitions

As shown in figure 4.2, $\sigma \rightarrow \sigma^*$ transition is a high energy process and therefore lies in the vacuum UV region. Alkanes, wherein only $\sigma \rightarrow \sigma^*$ transition is possible show absorption bands ~150 nm wavelength. Alkenes haveπ and π^* orbitals and can show several transition; the lowest-energy transition, $\pi \rightarrow \pi^*$ gives an absorption band ~170-190 nm for non-conjugated alkenes (effects of conjugation on electronic transitions are discussed later). The presence of nonbonding electrons in a molecule

further expands the number of possible transitions. The entire molecule, however, is not generally involved in the absorption of the radiation in a given wavelength range. In an aliphatic ketone, for example, the absorption band around 185 nm arises due to the $\pi \rightarrow \pi^*$ transition in the carbonyl group. Atoms that comprise the molecular orbitals involved in the electronic transitions constitute the molecular moiety that is directly involved in the transition. Such a group of atoms is called a **chromophore**. A structural modification in a chromophore is generally accompanied by changes in the absorption properties.

Instrumentation:

Figure 4.3A shows a schematic diagram of a single-beam spectrophotometer. The light enters the instrument through an entrance slit, is collimated and focused on to the dispersing element, typically a diffraction grating. The light of desired wavelength is selected simply by rotating the monochromator and impinged on the sample. The intensity of the radiation transmitted through the sample is measured and converted to absorbance or transmittance (discussed later). Double beam spectrophotometers overcome certain limitations of the single beam spectrophotometers and are therefore preferred over them. A double beam spectrophotometer has two light beams, one of which passes through the sample while other passes through a reference cell (Figure 4.3B). This allows more reproducible measurements as any fluctuation in the light source or instrument electronics appears in both reference and the sample and therefore can easily be removed from the sample spectrum by subtracting the reference spectrum. Modern instruments can perform this subtraction automatically. The most commonly used detectors in the UV/Visible spectrophotometers are the photomultiplier tubes (PMT). Modern instruments also use photodiodes as the detection systems. These diodes are inexpensive and can be arranged in an array so that each diode absorbs a narrow band of the spectrum. Simultaneous recording at multiple wavelengths allows recording of the entire spectrum at once. The monochromator in these spectrophotometers is placed after the sample so that the sample is exposed to the entire spectrum of the incident radiation and the transmitted radiation is dispersed into its components.

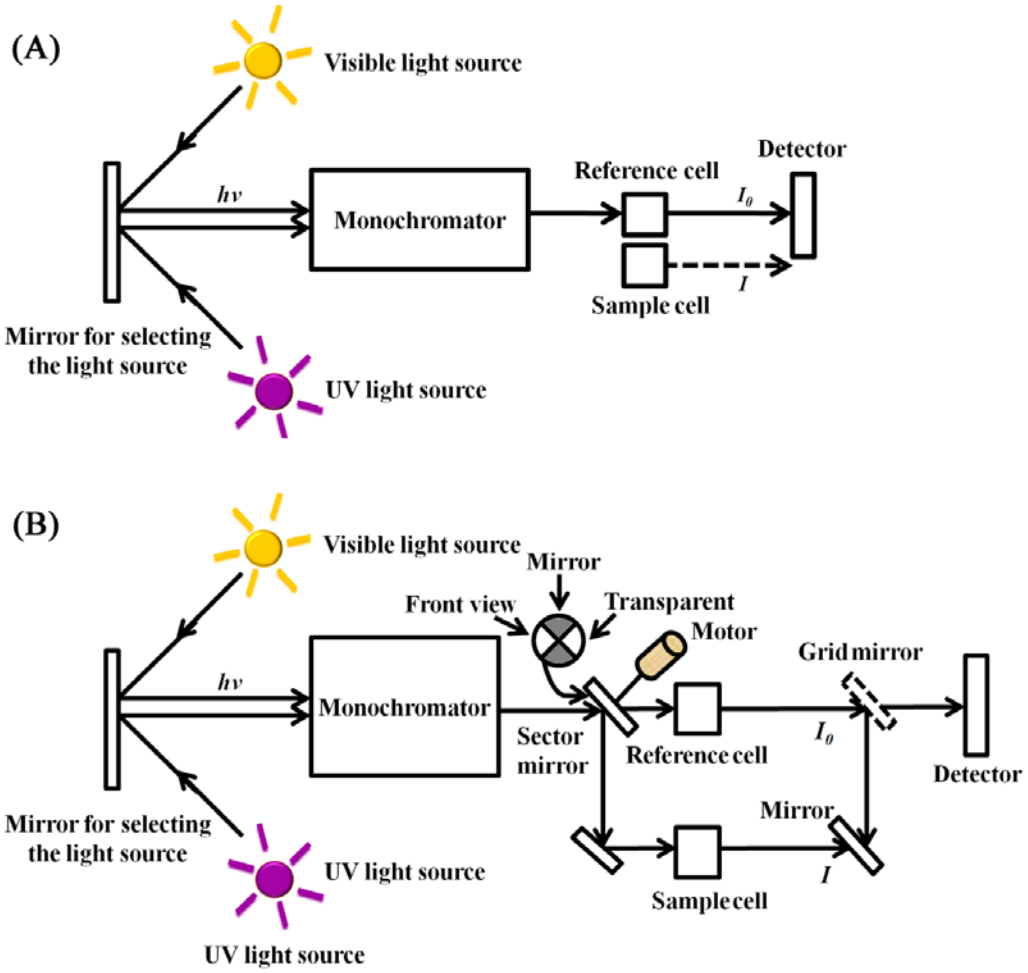


Figure 4.3 Schematic diagram showing a single beam (A) and a double beam (B) spectrophotometer

Beer-Lambert Law

It is quite intuitive that a higher concentration of the absorbing species in a sample would lead to higher absorption of the light. Furthermore, the higher thickness of the sample should result in higher absorption. Consider a cell (also called cuvette) of length, *l*, containing a solution of an absorbing molecule. The absorbing species in the sample can be represented by discs of cross-sectional area, σ. Now, let us consider a slab of infinitesimal thickness, *dx* and area, A (Figure 4.4). If an incident radiation of the resonance frequency (the frequency that causes maximum transition) having intensity I_o enters the sample cell, its intensity decreases as it penetrates the sample. Let us suppose that the intensity of the radiation before entering the infinitesimal slab is I_x.

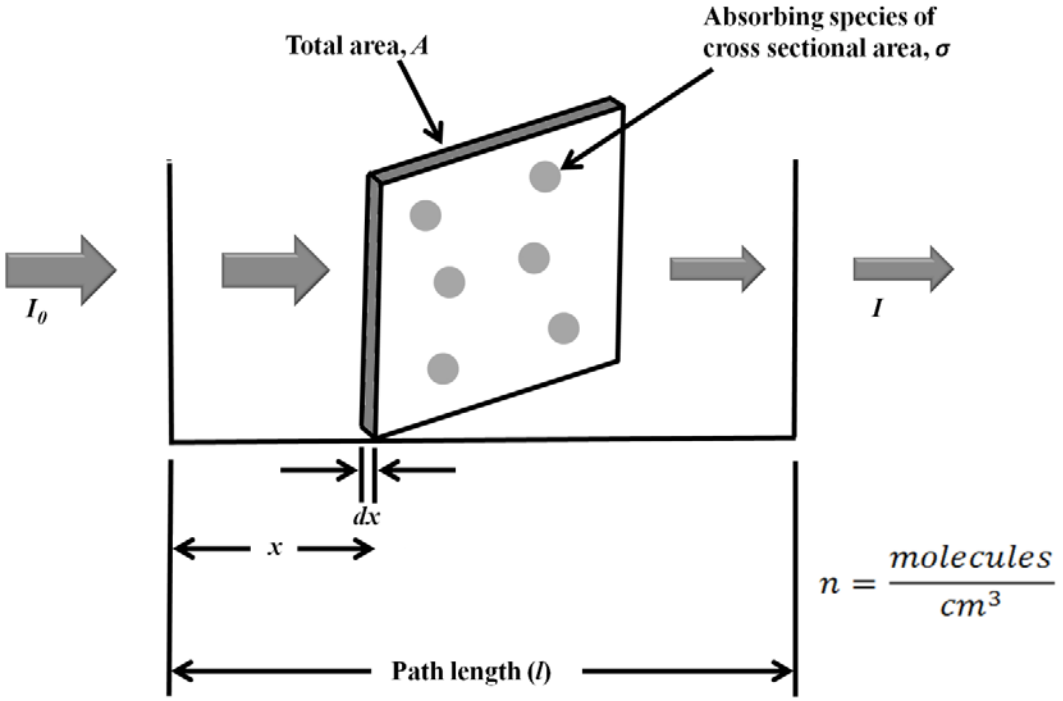


Figure 4.4A diagrammatic representation of light absorption by sample molecules in an infinitesimal thin slab within the sample

If the concentration of the absorbing molecules = $n\ \frac{molecules}{cm^3}$, the fraction of the area occupied by the molecules in the slab $= \frac{\sigma \times n \times Volume\ of\ slab}{A} = \frac{\sigma \times n \times A \times dx}{A} = \sigma \times n \times dx$

Therefore, the fraction of the photons $(\frac{dI}{I_x})$ absorbed is proportional to $\sigma \times n \times dx$. Assuming the probability of absorption if a photon strikes the molecule to be unity,

$$\frac{dI}{I_x} = -\sigma \times n \times dx \qquad \cdots\cdots\cdots\cdots\cdots\cdots (4.1)$$

The negative sign represents a decrease in intensity

Integrating equation 4.1 from $x = 0$ to $x = l$

$$\ln I|_{I_o}^{I} = -\sigma \times n \times x|_0^l \qquad \cdots\cdots\cdots\cdots\cdots\cdots (4.2)$$

$$\ln I - \ln I_o = -\sigma \times n \times l \qquad \cdots\cdots\cdots\cdots\cdots\cdots (4.3)$$

$$-\ln \frac{I}{I_o} = \sigma \times n \times l \qquad \cdots\cdots\cdots\cdots\cdots\cdots (4.4)$$

Now, the molar concentration of the molecules, *c* can be given by:

$$c\left(\frac{\text{moles}}{\text{litre}}\right) = n\left(\frac{\text{molecules}}{\text{cm}^3}\right) \times \frac{1}{6.022 \times 10^{23}}\left(\frac{\text{mole}}{\text{molecules}}\right) \times 1000\left(\frac{\text{cm}^3}{\text{litre}}\right)$$

Substituting for *n* in equation 4.4 and converting natural logarithm, *ln* into $\log_{10}$ gives:

$$-2.303 \times \log\frac{I}{I_o} = \sigma \times \{c \times 6.022 \times 10^{20}\} \times l \quad \cdots\cdots\cdots\cdots\cdots\cdots \text{(4.5)}$$

$$-\log\frac{I}{I_o} = \sigma \times c \times \left(\frac{6.022 \times 10^{20}}{2.303}\right) \times l \quad \cdots\cdots\cdots\cdots\cdots\cdots \text{(4.6)}$$

$-\log\frac{I}{I_o}$ is defined as the absorbance and $\sigma \times \left(\frac{6.022 \times 10^{20}}{2.303}\right)$ is defined as the molar absorption coefficient, denoted by the Greek alphabet, ε. Therefore, equation 4.6 can be written as:

$$\text{Absorbance, } A = \varepsilon cl \quad \cdots\cdots\cdots\cdots\cdots\cdots \text{(4.7)}$$

This equality showing linear relationship between absorbance and the concentration of the absorbing molecule (or chromophore, to be precise) is known as the Beer-Lambert law or Beer's law.

Transmittance is another way of describing the absorption of light. Transmittance (*T*) is simply the ratio of the intensity of the radiation transmitted through the sample to that of the incident radiation. Transmittance is generally represented as percentage transmittance (*%T*):

$$\%T = \frac{I}{I_o} \times 100$$

As is clear from the definition of absorbance and transmittance, both are dimensionless quantities. Absorbance and transmittance are therefore represented in arbitrary units (AU). The quantity of interest in an absorption spectrum is the molar absorption coefficient, ε which varies with wavelength (Figure 4.5). The wavelength at which highest molar absorption coefficient (ε_{max}) is observed is represented as λ_{max}. Area of cross-section of the absorbing species puts an upper limit to the molar absorption coefficient.

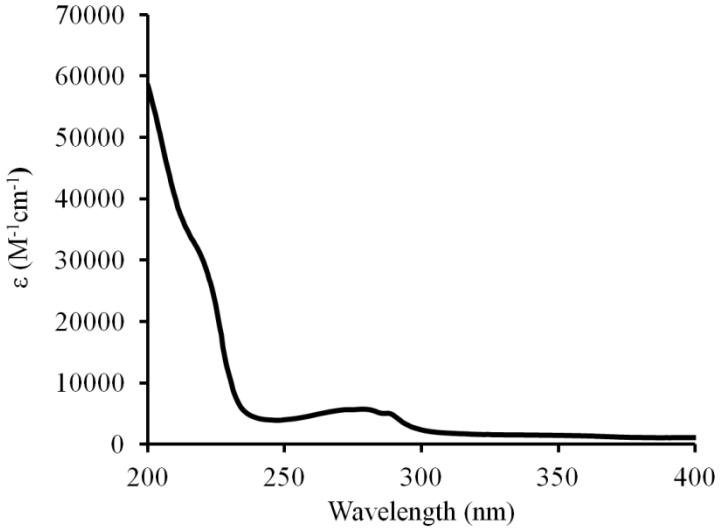


Figure 4.5 An absorption spectrum of N-acetyl-tryptophanamide

Deviations from Beer-Lambert law:

Beer-Lambert law can be used to determine the ε values of a compound by recording its absorption spectra at known concentrations. Alternatively, knowledge of ε enables the user to calculate the concentration of a compound in a given solution. It is, however, not uncommon to observe deviations from the Beer-Lambert law. Three major reasons that are responsible for the breakdown of linear relationship between absorbance and the concentration of the absorbing molecule are:

i. *High sample concentration*: The Beer-Lambert law generally holds good only for dilute solutions. At higher concentrations, the molecules come in close proximity thereby influencing their electronic properties. Although introduced as a constant at a particular wavelength for a compound, ε depends on the concentration of the compound and therefore results in deviation from linearity. At lower concentrations, however, ε can practically be assumed to be a constant.

ii. *Chemical reactions*: If a molecule undergoes a chemical reaction and the spectroscopic properties of the reacted and unreacted molecules differ, a deviation from Beer-Lambert law is observed. Change in the color of the pH indicator dyes is a classical example of this phenomenon.

iii. *Instrumental factors*: As ε is a function of wavelength, Beer-Lambert law holds good only for monochromatic light. Use of polychromatic radiation will result in deviation for linearity between absorbance and concentration.

For practical purposes, the samples giving absorbance values between 0.05 – 0.5 are considered highly reliable. At lower concentrations, the signal to noise ratio is small while at higher concentrations, absorbance values underestimate the concentration of the compound as increase in absorbance no longer matches the increase in concentration. If the absorbance values are higher, a sample can be diluted or a sample cell with smaller path length can be used; usually dilution of sample is preferred.

In the following lecture, we shall discuss the various factors that influence the absorption spectra of molecules and look at the applications of UV/Visible absorption spectroscopy for studying the biomolecules.

Lecture 5 UV/Visible Absorption Spectroscopy-II

In the previous lecture, we studied that UV/Visible radiation is absorbed by the molecules through transition of electrons in the chromophore from low energy molecular orbitals to higher energy molecular orbitals. We are interested in the transitions that lie in the far UV, near UV, and visible regions of the electromagnetic spectrum. The molecules that absorb in these regions invariably have unsaturated bonds. Plants are green due to unsaturated organic compounds, called chlorophylls. A highly unsaturated alkene, lycopene, imparts red color to the tomatoes (Figure 5.1).

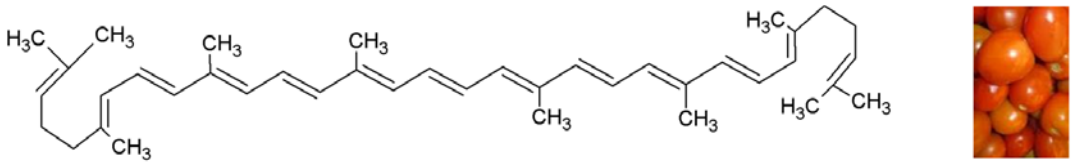


Figure 5.1 Structure of lycopene, the pigment that imparts red color to the tomatoes

As can be seen from its structure, lycopene is a highly conjugated alkene. As compared to the simple non-conjugated alkenes that typically absorb in vacuum UV region, absorption spectrum of lycopene is hugely shifted towards higher wavelengths (or lower energy). There can be factors that could shift the absorption spectra to smaller wavelengths or can increase/decrease the absorption intensity. Before understanding how conjugation causes shift in the absorption spectra, let us look at some important terms that are used to refer to the shifts in absorption spectra (Figure 5.2):

Bathochromic shift: Shift of the absorption spectrum towards longer wavelength

Hypsochromic shift: Shift of the absorption spectrum towards smaller wavelength

Hyperchromic shift: An increase in the absorption intensity

Hypochromic shift: A decrease in the absorption intensity

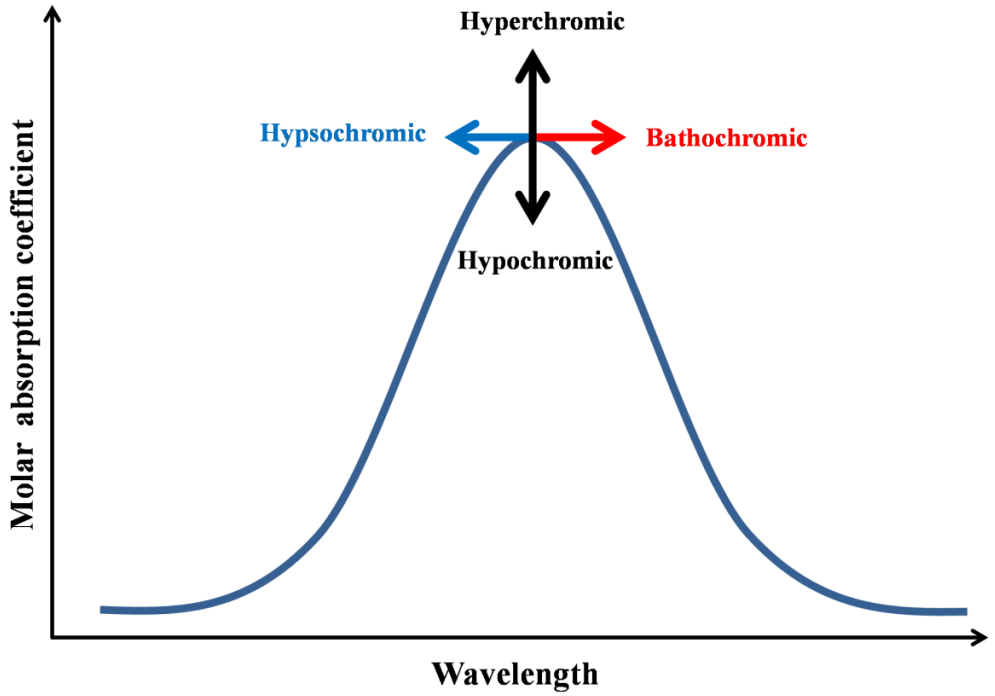


Figure 5.2 Terminology for shifts in absorption spectra

Conjugation: Conjugation brings about a bathochromic shift in the absorption bands. The higher the extent of conjugation, the more is the bathochromic shift. Such shift in absorption spectra can easily be explained using molecular orbital theory. Figure 5.3 shows the molecular orbitals drawn for ethylene; 1,3-butadiene; and 1,3,5-hexatriene on a qualitatively same energy scale for comparing their energies. As is clear from the figure, the energy differences between the highest occupied molecular orbital (HOMO) and the lowest unoccupied molecular orbital (LUMO) decreases as the conjugation increases. This provides an explanation as to why an electronic transition is possible at lower energy (higher wavelength) as the conjugation increases.

Energy levels of conjugated alkenes' molecular orbitals: The energy levels of the orbitals increase as the number of vertical nodes increase. The lowest energy π orbital has no nodes while the highest energy π^* orbital has $n–1$ nodes where n is the number of p–orbitals combined.

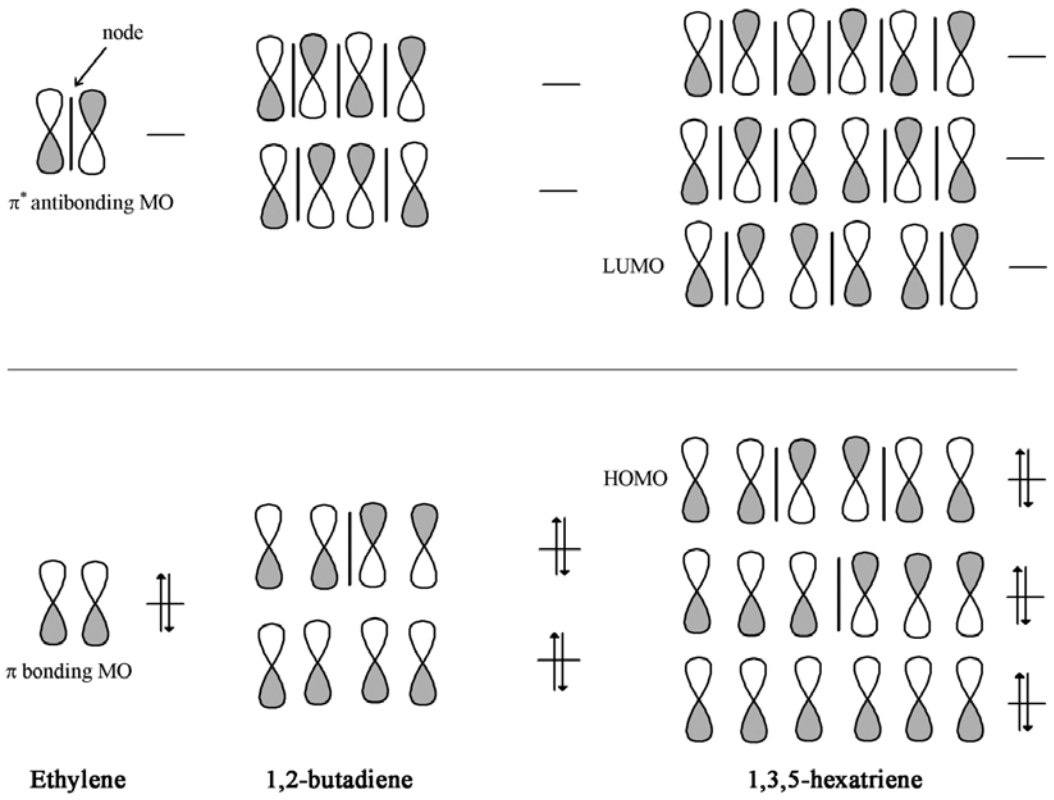


Figure 5.3 Molecular orbitals of ethylene; 1,3-butadiene; and 1,3,5-hexatriene. Notice the decrease in the energy gap of HOMO and LUMO as the conjugation increases.

Auxochrome: Auxochromes are the chemical groups that result in a bathochromic shift when attached to a chromophore. The strongest auxochromes like –OH, $–NH_2$, –OR, etc. possess nonbonding electrons. They exhibit bathochromism by extending conjugation through resonance.

The auxochrome modified chromophore is a new chromophore in real sense. The term auxochrome is therefore rarely used these days, and the entire group (basic chromophore + auxochrome) can be considered as a chromophore different from the basic chromophore. Alkyl groups also result in the bathochromic shifts in the absorption spectra of alkenes. Alkyl groups do not have non-bonded electrons, and the effect is brought about by another type of interaction called *hyperconjugation*.

Solvents: The solvents used in any spectroscopic method should ideally be transparent (non-absorbing) to the electromagnetic radiation being used. Table 5.1 shows the wavelength cutoffs (the lowest working wavelength) of some of the solvents used in UV/visible spectroscopy.

Table 5.1 Solvents commonly used in UV/visible spectroscopy

Solvent	**Wavelength cutoff**
Water	190 nm
Acetonitrile	190 nm
Cyclohexane	195 nm
Methanol	205 nm
95% ethanol	205 nm

Water, the solvent of biological systems, thankfully is transparent to the UV/visible region of interest *i.e.* the regions above $\lambda > 190$ nm. Solvents also play important role on the absorption spectra of molecules. Spectrum of a compound recorded in one solvent can look significantly different in intensity, wavelength of absorption, or both from that recorded in another. This is not something unexpected because energies of different electronic states will depend on their interaction with solvents. Polarity of solvents is an important factor in causing shifts in the absorption spectra. Conjugated dienes and aromatic hydrocarbons are little affected by the changes in solvent polarity. α,β-unsaturated carbonyl compounds are fairly sensitive to the solvent polarity. The two electronic transitions $\pi \rightarrow \pi^*$ and $n \rightarrow \pi^*$ respond differently to the changes in polarity. Polar solvents stabilize all the three molecular orbitals (*n, π,* and π^*), albeit to different extents (Figure 5.4). The non-bonding orbitals are stabilized most, followed by π^*. This results in a bathochromic shift in the $\pi \rightarrow \pi^*$ absorption band while a hypsochromic shift in $n \rightarrow \pi^*$ absorption band. Shift to different extents of the two bands will result in the different shape of the overall absorption spectrum.

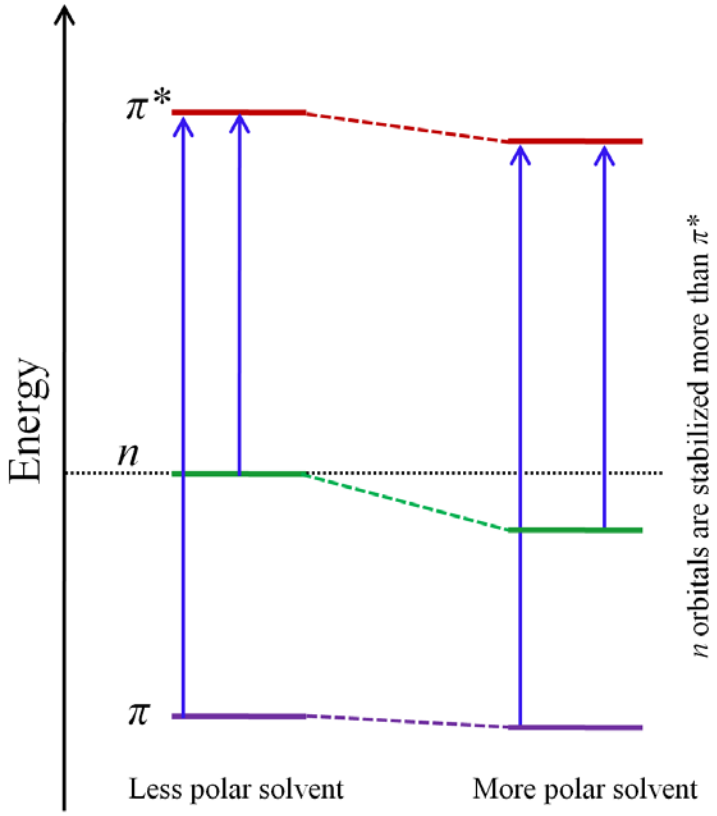


Figure 5.4 Differential stabilization of molecular orbitals in polar solvents

Biological chromophores

Amino acids and proteins: Among the 20 amino acids that constitute the proteins, tryptophan, tyrosine, and phenylalanine absorb in the near UV region. All the three amino acids show structured absorption spectra. The absorption by phenylalanine is weak with an ε_{max} of ~200 $M^{-1}cm^{-1}$ at ~250 nm. Molar absorption coefficients of ~1400 $M^{-1}cm^{-1}$ at 274 nm and ~5700 $M^{-1}cm^{-1}$ at 280 nm are observed for tyrosine and tryptophan, respectively. Disulfide linkages, formed through oxidation of cysteine resides, also contribute to the absorption of proteins in near UV region with a weak ε_{max} of ~300 $M^{-1}cm^{-1}$ around 250-270 nm. The absorption spectra of proteins are therefore largely dominated by Tyr and Trp in the near UV region. In the far UV region, peptide bond emerges as the most important chromophore in the proteins. The peptide bond displays a weak $n \rightarrow \pi^*$ transition ($\varepsilon_{max} \approx 100$ $M^{-1}cm^{-1}$) between 210-230 nm, the exact band position determined by the H-bonding interactions the peptide backbone is involved in. A strong $\pi \rightarrow \pi^*$ transition ($\varepsilon_{max} \approx 7000$ $M^{-1}cm^{-1}$) is observed around 190 nm. Side chains of Asp, Glu, Asn, Gln, Arg, His also contribute

to the absorbance in the far UV region. Figure 5.5 shows an absorption spectrum of a peptide

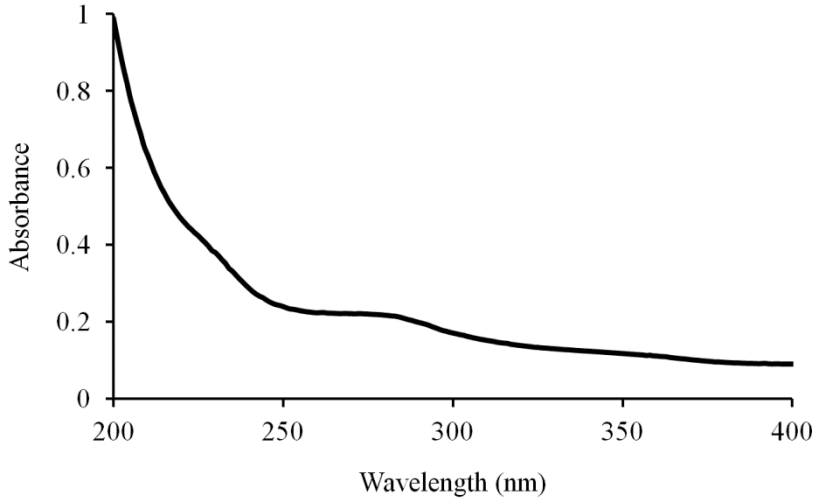


Figure 5.5 Absorption spectrum of a peptide. The absorption band ~280 nm is due to aromatic residues. Absorption band in the far UV region arises due to peptide bond electronic transitions.

Nucleic acids: Nucleic acids absorb very strongly in the far and near UV region of the electromagnetic spectrum. The absorption is largely due to the nitrogenous bases. The transitions in the nucleic acid bases are quite complex and many $\pi \rightarrow \pi^*$ and $n \rightarrow \pi^*$ transitions are expected to contribute to their absorption spectra. A 260 nm wavelength radiation is routinely used to estimate the concentration of nucleic acids. Though the molar absorption coefficients vary for the nucleotides at 260 nm, the average ε_{max} can be taken as ~10^4 $M^{-1}cm^{-1}$. It is important to mention that nucleotides show hyperchromicity when exposed to aqueous environment. The absorbance of the free nucleotides is higher than that of single stranded nucleic acid which is higher than that of the double stranded nucleic acid (assuming equal amount of the nucleotides present in all three).

Other chromophores: Nucleotides like NADH, NADPH, FMN, and FAD; porphyrins such as heme, chlorophylls and other plant pigments; retinal (light sensing molecule); vitamins; and a variety of unsaturated compounds constitute chromophores in the UV and visible region.

Having studied the principles of the UV/visible absorption spectroscopy and various factors that influence the electronic transitions, we can now have a look at its applications, especially the applications for analyzing the biological samples.

Applications:

i. *Determination of molar absorption coefficient*: From Beer-Lambert law, $A = \varepsilon cl$. It is therefore straightforward to calculate the molar absorption coefficient of a compound if the concentration of compound is accurately determined.

ii. *Quantification of compounds*: This is perhaps the most common application of a UV/visible spectrophotometer in a bioanalytical laboratory. If the molar absorption coefficient at a wavelength is known for the compound, the concentration can easily be estimated using Beer-Lambert law. The compounds can still be quantified if their molar absorption coefficients are not known. Estimation of total protein concentration in a given solution is an important example of this. As the given solution is a mixture of many different proteins, the ε is not available. There are, however, dyes that specifically bind to the proteins producing colored complex. The color produced will be proportional to the amount of the protein present in the solution. Performing the experiment under identical conditions using known concentrations of a protein gives a standard graph between absorbance of the dye and the amount of protein. This standard graph is then used to estimate the concentration of the given protein sample.

iii. *Quality control*: A given organic compound such as a drug can be studied for its purity. Comparison of spectrum with the standard drug will detect the impurities, if any. UV/Visible absorption is often used to detect the nucleic acid contamination in the protein preparations. Aromatic amino acids as well as the nucleotides show absorption band in the near UV region and there is a considerable overlap in the absorption spectra of aromatic amino acids and the nucleotides. A nucleic acid contamination in a protein, however, can be determined by measuring

$\frac{A_{260}}{A_{280}}$ ratio is not useful in detecting protein contaminations in DNA preparations. This is because of the large difference in molar absorption coefficients of these molecules. To cause an appreciable change in the $\frac{A_{260}}{A_{280}}$ ratio, there should a large amount of protein present.

absorbances at 260 and 280 nm. A typical nucleic acid containing all four bases shows an absorption band centered ~260 nm while a protein having aromatic amino acids shows absorption band centered ~280 nm. It is possible to determine the purity of protein preparations by recording absorbances at both 260 and 280 nm. A ratio of the absorbance at 260 nm to that at 280 nm *i.e.* $\frac{A_{260}}{A_{280}}$ is a measure of the purity.

iv. *Chemical kinetics*: UV/visible spectroscopy can be used to monitor the rate of chemical reactions if one of the reactants or products absorbs in a region where no other reactant or product absorbs significantly.

v. *Detectors in liquid chromatography instruments*: UV/visible detectors are perhaps the most common detectors present in liquid chromatography systems. Modern instruments use photodiode array detectors that can detect the molecules absorbing in different spectral regions (Figure 5.6).

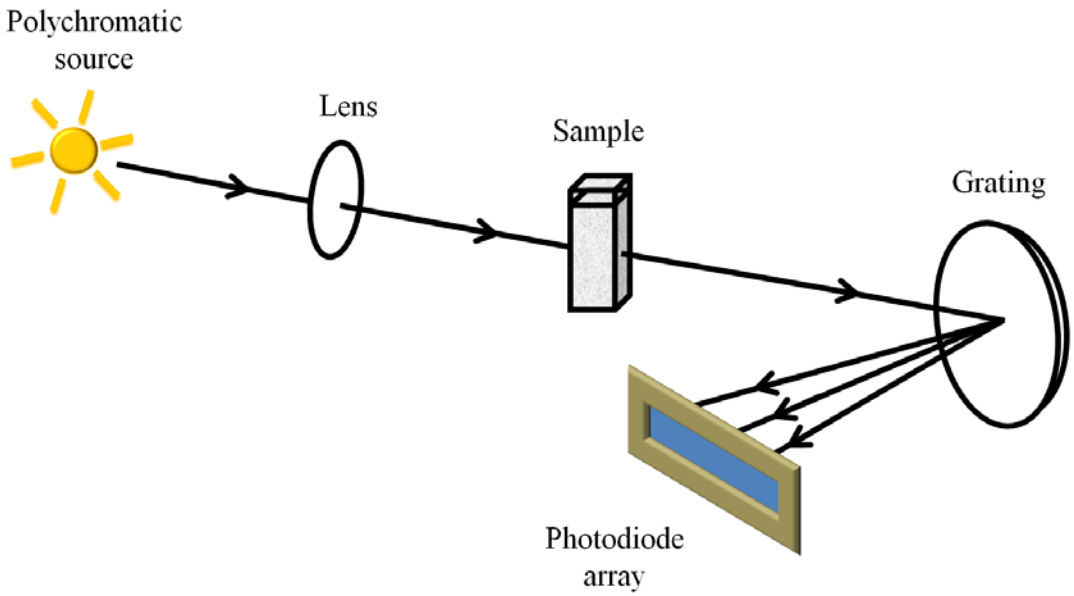


Figure 5.6 Diagram of a photodiode array detector

vi. *Determination of melting temperature of DNA*: A double stranded DNA molecule can be denatured into the single strands by heating it. Melting temperature, T_m is the temperature at which 50% of the DNA gets denatured into single strands. Denaturation of DNA is accompanied by hyperchromic shift in the absorption spectra in the near UV region. A melting curve (plot between temperature and absorbance at 260 nm) is plotted and T_m is determined (Figure 5.7).

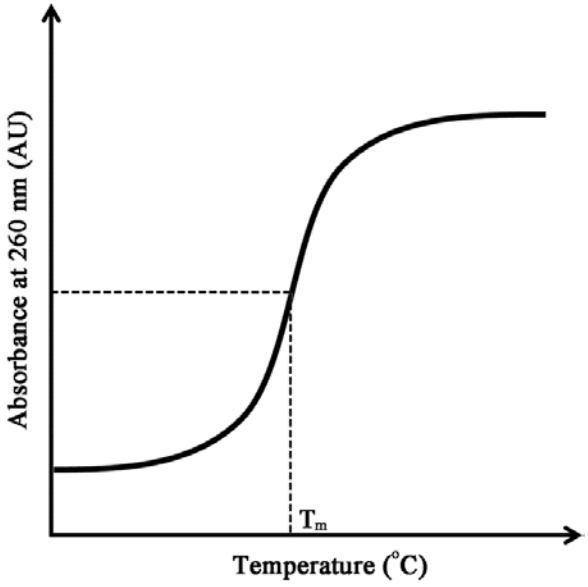


Figure 5.7 Thermal denaturation of a DNA sample; a plot of absorbance at 260 nm against the temperature allows determination of the melting temperature (T_m).

vii. *Microbial growth kinetics*: A UV/visible spectrophotometer is routinely used to monitor the growth of microorganisms. *The underlying principle behind this, however, is not absorbance but scattering*. As the number of microbial cells increase in a culture, they cause more scattering in light. The detector therefore receives less amount of radiation, recording this as absorbance. To distinguish this from actual absorbance, the observed value is referred to as the optical density.

QUIZ

Q1: The molar absorption coefficient of tyrosine in water is 1280 $M^{-1}cm^{-1}$ at 280 nm. Calculate the concentration of a tyrosine solution in water if the absorbance of the solution is 0.34 in a 1 cm path length cell.

Ans: Given:

λ = 280 nm $\quad \varepsilon_{280\ nm}$ = 1280 $M^{-1}cm^{-1}$ $\quad l$ = 1 cm $\quad A$ = 0.34

From Beer-Lambert Law:

$$A = \varepsilon c l$$

$$c = \frac{A}{\varepsilon l} = \frac{0.32}{1280 M^{-1}cm^{-1} \times 1\ cm} = 0.00025\ M = 250\ \mu M$$

Q2: Calculate the concentration of a tryptophan solution that gives an absorbance of 0.25 at 280 nm in a 1 mm path length cell (Given $\varepsilon_{280\ nm}$ = 5690 $M^{-1}cm^{-1}$).

Ans: The concentration of the given sample can be estimated using Beer-Lambert law:

$$A = \varepsilon c l$$

$$c = \frac{A}{\varepsilon l}$$

$$c = \frac{0.25}{5690\ M^{-1}cm^{-1} \times 1\ mm}$$

$$c = \frac{0.25}{5690\ M^{-1}cm^{-1} \times 0.1\ mm}$$

$$c = 4.39 \times 10^{-5}\ M = 43.9\ \mu M$$

Q3: Concentration of a pure compound in solution can easily be determined by taking absorbance at any wavelength in a given spectral region if ε at these wavelengths is known. Why then absorbance is generally recorded at λ_{max}?

Ans: This is done for the following reasons:

a) At λ_{max}, the ε value is maximum, therefore reliable absorbance *i.e.* A between 0.05 – 0.5 can be obtained at lower concentrations of the compound.

b) At λ_{max}, the slope of the absorption spectrum, $\frac{dA}{d\lambda}$ or $\frac{d\varepsilon}{d\lambda}$, is zero. This ensures that for a given bandwidth of the incident radiation, the ε is relatively constant in this region as compared to the regions of non-zero slopes. If ε is not constant, the linearity of the Beer Lambert law is compromised.

Lecture 6 Fluorescence Spectroscopy-I

This lecture is a very concise review of the phenomenon of fluorescence and the associated processes. Let us move a step forward from the absorption of the UV/visible radiation. What happens to the electrons that absorb UV/visible light and occupy the high energy molecular orbitals? In a UV/visible absorption experiment, the samples continue absorbing light. This means that the higher energy molecular orbitals never get saturated. This further implies that after excitation, the molecules somehow get rid of the excess energy and return back to the ground state. The electrons can return back to the ground state in different ways such as releasing the excess energy through collisions or through emitting a photon. In fluorescence, the molecules return back to the ground state by emitting a photon. The molecules that show fluorescence are usually referred to as the *fluorophores*. Various electronic and molecular processes that occur following excitation are usually represented on a Jablonski diagram as shown in Figure 6.1.

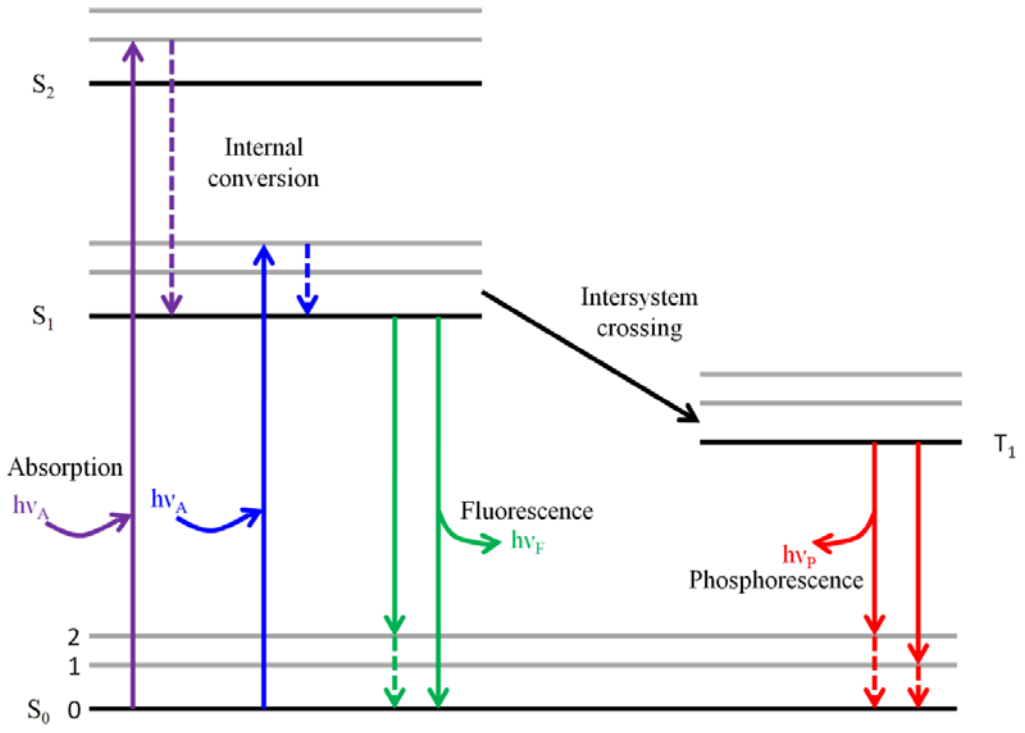


Figure 6.1 Jablonski diagram showing various processes following absorption of light by the fluorophore

S_0, S_1, and S_2 represent the singlet electronic states while the numbers 0, 1, 2 represent the vibrational energy levels associated with the electronic states. T_1 depicts the first triplet electronic state. Let us go through the processes shown in Figure 6.1:

Absorbance: S_0 state with 0^{th} vibrational level is the state of lowest energy and therefore, the highest populated state. Absorption of a photon of resonant frequency usually results in the population of S_1 or S_2 electronic states; but usually a higher vibrational state. Transition of electrons from low energy molecular orbital to a high energy molecular orbital through absorption of light is a femtosecond (10^{-15} s) phenomenon. The electronic transition, therefore, is too quick to allow any significant displacement of the nuclei during transition.

Internal conversion: Apart from few exceptions, the excited fluorophores rapidly relax to the lowest vibrational state of S_1 through non-radiative processes. Non-radiative electronic transition from higher energy singlet states to S_1 is termed as *internal conversion* while relaxation of a fluorophore from a higher vibrational level of S_1 to the lowest vibration state is termed as *vibrational relaxation*. The terms 'internal conversion' and 'vibrational relaxation', however, are often interchangeably used. The timescale of internal conversion/vibrational relaxation is of the order of 10^{-12} seconds.

Fluorescence: Fluorescence lifetimes are of the order of 10^{-8} seconds, implying that the internal conversion is mostly complete before fluorescence is observed. Therefore, fluorescence emission is the outcome of fluorophore returning back to the S_0 state through $S_1 \rightarrow S_0$ transition emitting a photon. This also explains why emission spectra are usually independent of the excitation wavelength, also known as Kasha's rule (However, there are exceptions wherein fluorescence is observed from $S_2 \rightarrow S_1$ transition). The $S_1 \rightarrow S_0$ transition, like $S_0 \rightarrow S_1$ transition, typically results in the population of higher energy vibrational states. The molecules then return back to the lowest vibrational state through vibrational relaxation.

Intersystem crossing: Intersystem crossing referes to an isoenergetic non-radiative transition between electronic states of different multiplicities. It is possible that a molecule in a vibrational state of S_1 can move to the isoenergetic vibrational state of T_1. The molecule then relaxes back to the lowest vibrational state of the triplet state.

Phosphorescence: The molecule in the triplet state, T_1, can return back to the S_0 state emitting a photon. This process is known as phosphorescence and has time scales of several orders of magnitudes higher than that of fluorescence (10^{-3} – 10 s).

Characteristics of fluorescence:

Figure 6.2 shows absorption and fluorescence emission spectrum of a hypothetical fluorophore. The important characteristics of the fluorescence emission can be briefly summarized as follows:

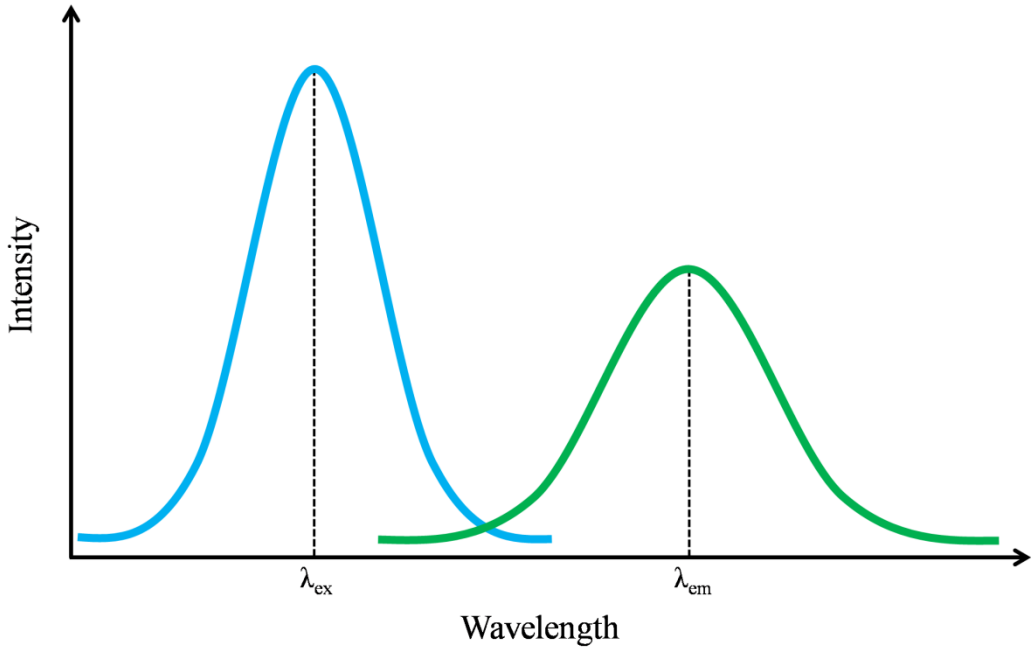


Figure 6.2 Absorption and fluorescence emission spectrum of a hypothetical fluorophore

Stokes shift: A fluorescence emission spectrum is always shifted towards longer wavelengths with respect to the absorption spectrum. This shift is known as *Stokes shift* and is expected as excited molecules lose energy through processes like internal conversion and vibrational relaxation. The emitted radiation is therefore expected to be of lower energy *i.e.* higher wavelength.

Kasha's rule: As fluorescence emission is observed from $S_1 \rightarrow S_0$ transtions (except a few exceptions), fluorescence absorption spectrum is independent of the excitation wavelength.

Franck-Condon principle: The Franck-Condon principle states that the positions of the nuclei do not change during electronic transitions. The transitions are said to be vertical. This implies that if the probability of $0^{th} \rightarrow 2^{nd}$ vibrational transition during $S_0 \rightarrow S_1$ transition is highest, the $2^{nd} \rightarrow 0^{th}$ transition will be most probable in the reciprocal transition (Figure 6.3).

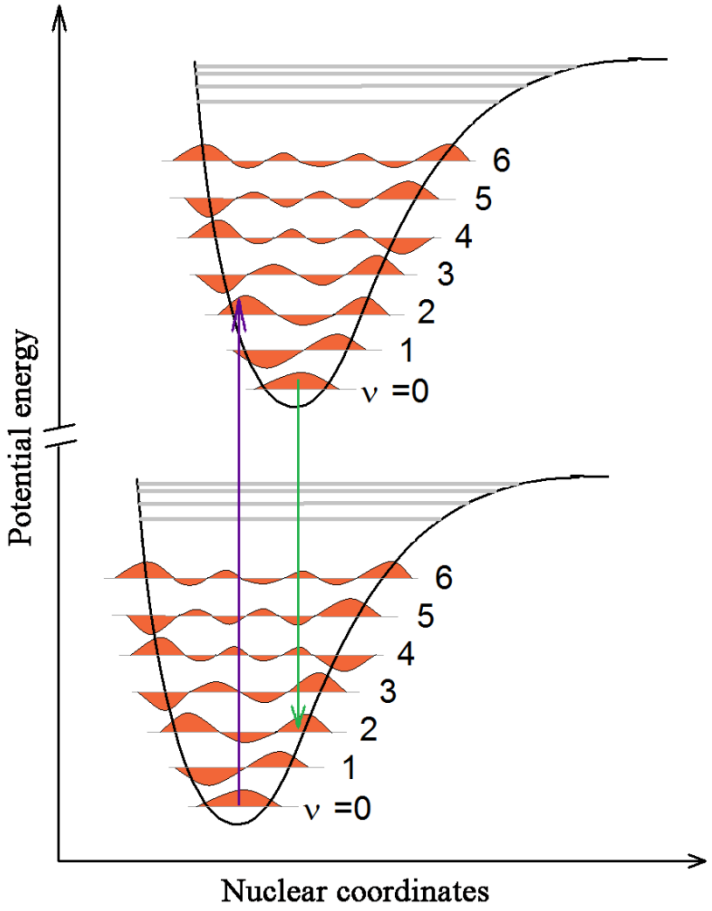


Figure 6.3 Potential energy diagrams showing the Franck-Condon principle

This results in an emission spectrum that is a mirror image of the $S_0 \rightarrow S_1$ transition in terms of the shape. There are several exceptions to the mirror image rule that arise largely due to the excited state reactions of the molecule.

Quantum yield: As has been mentioned earlier, an excited molecule can come back to the ground state through non-radiative pathways.

Fluorescence quantum,

$$Q = \frac{number\ of\ photons\ emitted}{number\ of\ photons\ absorbed} = \frac{\Gamma}{\Gamma + k_{nr}}$$

where,

Γ is the rate of radiative process *i.e.* fluorescence

k_{nr} is the rate of all the non-radiative processes bringing molecule to the S_0 state

Fluorescence lifetime: Lifetime of a fluorophore is defined as the average time it spends in the excited state before returning to the S_0 state. It is therefore the reciprocal of the rate of processes de-exciting the molecule.

$$\text{Fluorescence lifetime, } \tau = \frac{1}{\Gamma + k_{nr}}$$

Fluorescence quenching, resonance energy transfer and anisotropy

Fluorescence spectroscopy comprises of experiments exploiting various different phenomena related to it. Discussion of all these experiments is beyond the scope of this course, but we shall have a quick look at a few important phenomena related to fluorescence.

Fluorescence quenching: A decrease in fluorescence intensity is referred to as quenching. A molecule that quenches the fluorescence of a fluorophore is called a quencher. A quencher can be either a collisional quencher or a static quencher. A collisional quencher brings about decrease in fluorescence intensity by de-exciting the excited fluorophore through collisions. Addition of another non-radiative process to the system leads to lower quantum yield. A static quencher forms a non-fluorescent complex with the fluorophore. It effectively leads to a decrease in the concentration of the fluorophore thereby decreasing the fluorescence emission intensity.

Resonance energy transfer: Resonance energy transfer (RET), also known as fluorescence resonance energy transfer (FRET) is an excited state phenomeneon wherein energy is transferred from a donor molecule (*D*) to an acceptor molecule (*A*). The prerequisite for the energy transfer is that there should be an overlap between the emission spectrum of the *D* and the absorption spectrum of the *A* (Figure 6.4).

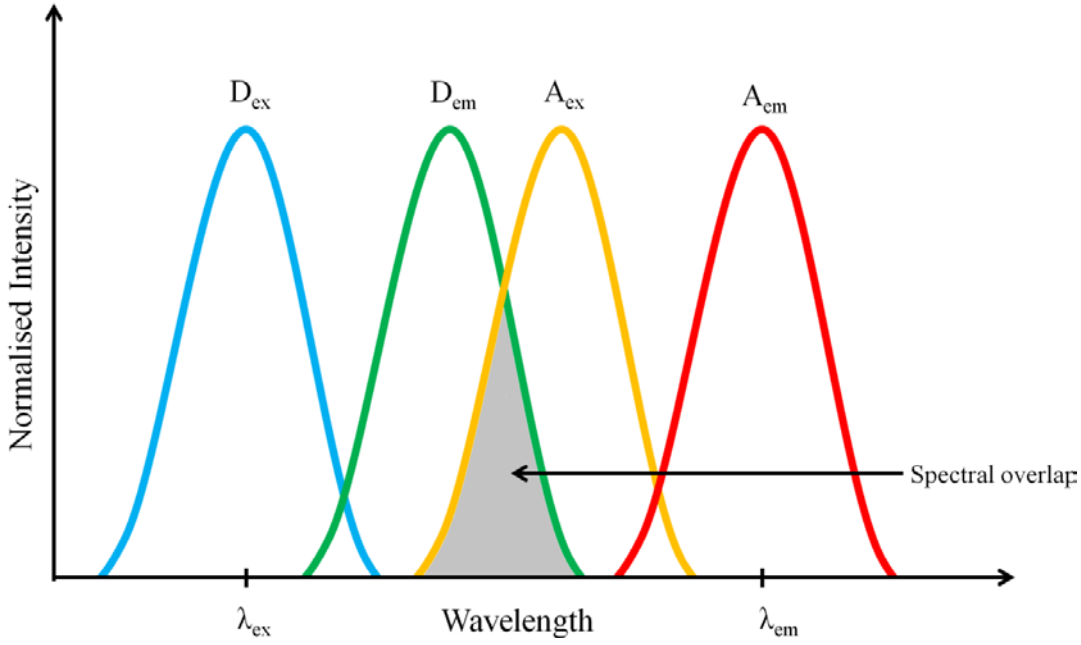


Figure 6.4 Diagrammatic representation of spectral overlap between donor's emission and acceptor's absorption spectrum.

The efficiency of energy transfer depends upon

i. the distance between *D* and *A*
ii. the relative orientation of the transition dipoles of *D* and *A*
iii. the extent of the overlap between *D*'s emission spectrum and *A*'s absorption spectrum

$$\text{Efficiency of energy transfer } E = \frac{R_0^6}{R_0^6 + r^6}$$

where,

r is the distance between *D* and *A*.

R_0 (also called the Förster distance) is the distance (*r*) between *D* and *A* at which the efficiency of energy transfer is 50%, and is characteristic of a *D*-*A* FRET pair.

Resonance energy transfer can be used to determine the distances between *D* and *A*, and is therefore also termed as molecular ruler.

Fluorescence anisotropy: The radiation emitted by a sample following excitation with polarized light can be polarized. Polarization is measured in terms of anisotropy. Zero anisotropy implies isotropic/non-polarized radiation while non-zero anisotropy implies some degree of polarization. Figure 6.5 shows how fluorescence anisotropic measurements are made.

Transition dipole moment: The transition dipole moment represents the transient dipole moment generated from the charge displacement during a transition. The transition dipole moments are defined vector quantities for the transitions of a particular molecule.

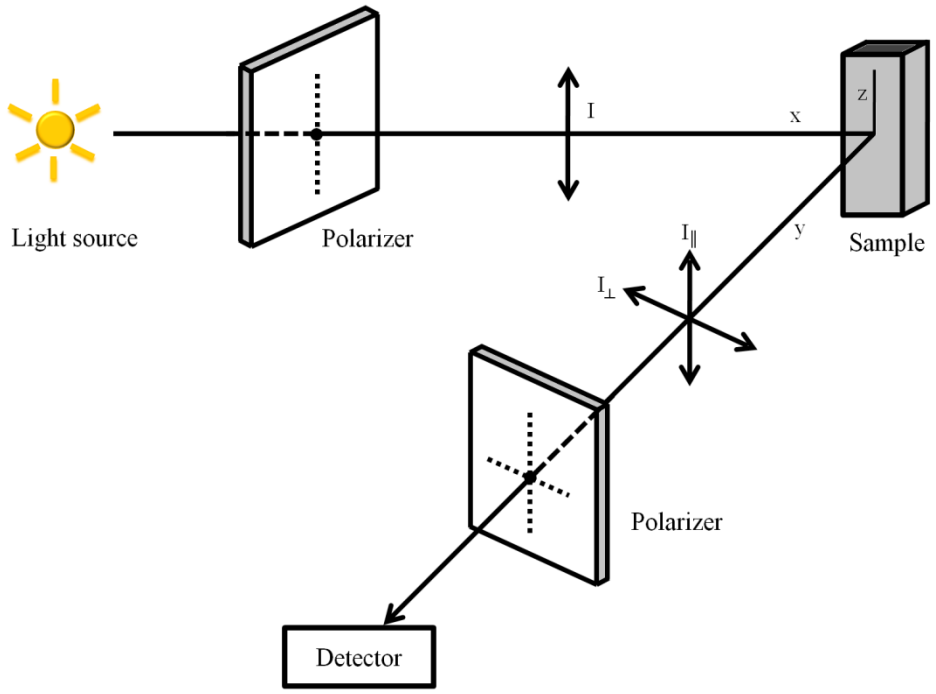


Figure 6.5 A schematic diagram showing the measurement of fluorescence anisotropy

The sample is excited with the linearly polarized light and emission is recorded at 90°. A polarizer is placed before the detector that allows intensity measurement of the light polarized parallel ($I_{\parallel}$) and perpendicular ($I_{\perp}$) to the direction of excitation radiation. The anisotropy (r) is given by

$$r = \frac{I_{\perp} - I_{\parallel}}{I_{\perp} + 2I_{\parallel}}$$

Molecular tumbling before emission changes the orientation of the transition dipole moment, resulting in the loss of polarization (Figure 6.6). As rotational diffusion of the molecules depends on their sizes, fluorescence anisotropy can be used to measure the diffusion coefficient and therefore the sizes of the molecules.

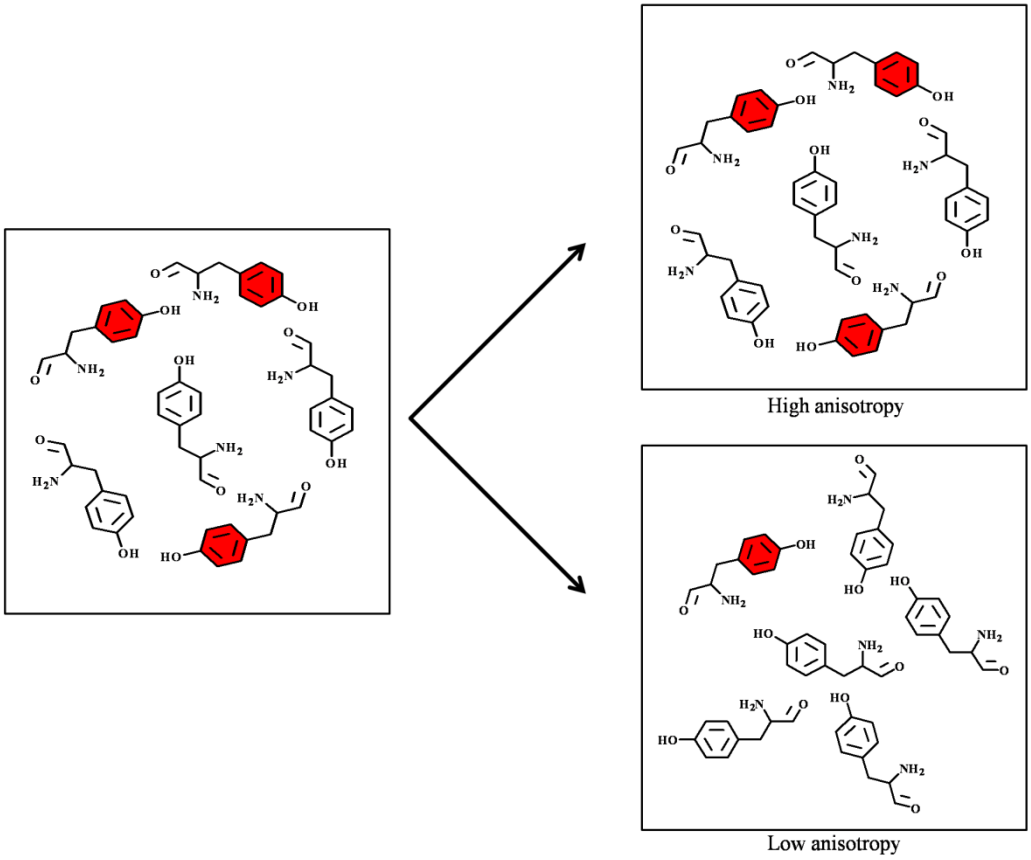


Figure 6.6 Depolarization of radiation as a result of molecular tumbling

We shall, in the next lecture, discuss the biological fluorophores and the applications of fluorescence in understanding the biomolecules.

Lecture 7 Fluorescence Spectroscopy-II

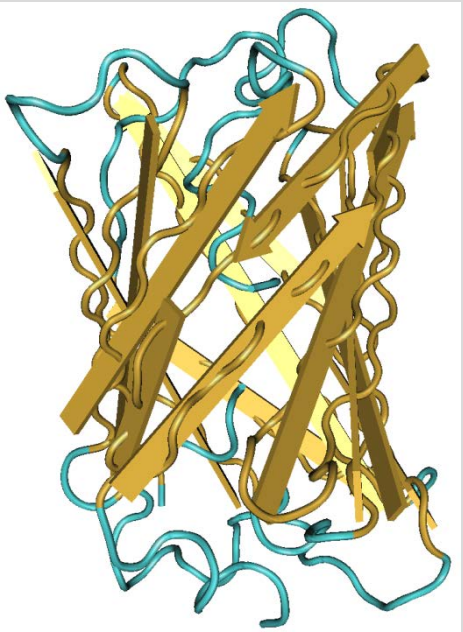
Biological fluorophores

Amino acids: Aromatic amino acids tryptophan (Trp), tyrosine (Tyr), and phenylalanine (Phe) are perhaps the most important intrinsic biological fluorophores. Proteins harboring these amino acids become intrinsically fluorescent.

Proteins: Proteins are fluorescent due to the presence of aromatic amino acids that fluoresce in the near UV region. Certain proteins, however, do fluoresce in the visible region. Green fluorescent protein (GFP), for example, fluoresces in the green region of the electromagnetic spectrum. The discovery of green fluorescent protein has revolutionized the area of cell biology research. It is therefore important to see what green fluorescent protein is and why it fluoresces in the visible region (See Box 1).

Box 7.1: Green Fluorescent Protein (GFP)

Green fluorescent protein, abbreviated as GFP was discovered by Shimomura and coworkers in 1962. The protein was isolated from the jellyfish, *Aequorea victoria*, that glows in the dark. GFP is a 238 amino acid long protein that folds into an 11-stranded β-barrel structure wherein an α-helix passes through the barrel.



PDB ID: 1EMA

The fluorophore of the GFP, p-hydroxybenylideneimidazolinone is formed by the residues 65-67 (Ser-Tyr-Gly) and is present in the α-helix passing through the barrel.

The excitation spectrum of GFP exhibits a strong absorption band at 395 nm and a weak band at 475 nm. Emission is observed at ~504 nm *i.e.* in the green region. GFP is an excellent fluorophore with a molar absorption coefficient of ~30000 $M^{-1}cm^{-1}$ at 395 nm and fluorescence quantum yield of 0.79. GFP has been engineered through extensive mutations to remove the undesirable properties that could affect its use as a potential fluorophore. For example, a Ser65 → Thr65 mutant has improved quantum yield and its major excitation band shifted to 490 nm. GFP has the tendency to form oligomers, seriously questioning its use as a fluorescent probe. The aggregation tendency has also been removed through extensive mutations. GFP can easily be tagged to a protein by expressing the fusing gene (GFP gene fused with the gene expressing the desired protein). The GFP then acts as a reporter for all the processes the linked protein is involved in. Several color variants of GFP have been generated through modifications in the residues that constitute the fluorophore. Development of the GFP variants with varying excitation and emission characteristics has made it possible to label the proteins differentially. This is a huge breakthrough and allows easy monitoring of the biological processes using fluorescence microscopy as discussed in lectures 15 and 20.

Nucleotides: Nicotinamide adenine dinucleotide in its reduced form, NADH and the flavin adenine dinucleotide in its oxidized form, FAD are fluorescent in the visible region of the electromagnetic spectrum. It is not necessary for all the biomolecules to have an intrinsic fluorophore to perform fluorescence experiments. Fluorescent groups can be covalently incorporated into the molecules making them fluorescent with desirable fluorophore. Such externally incorporated fluorophores are called extrinsic fluorophores.

Applications of fluorescence

Protein folding: High sensitivity of tryptophan fluorescence to the polarity of solvent makes it an interesting intrinsic fluorescent probe for studying protein folding. In the proteins having Phe and Tyr, Trp can be selectively excited at 295 nm. In water and other aqueous solutions, tryptophan fluoresces with an emission maximum, λ_{max} around 350 nm. A tryptophan present in the hydrophobic environment usually displays a blue shift in the emission spectrum and an increase in quantum yield. Due to the hydrophobic nature of the indole side chain, tryptophans are usually buried inside the core of the proteins. The folding can therefore be studied by monitoring the Trp fluorescence as protein folds burying the water-exposed Trp residues inside the protein.

Peptide-lipid interactions: Interaction of the peptides having Trp residues with lipid bilayers can easily be studied using fluorescence spectroscopy. Interaction of the peptide with lipids brings the tryptophan in relatively hydrophobic environment causing a blue shift in emission spectrum (Figure 7.1).

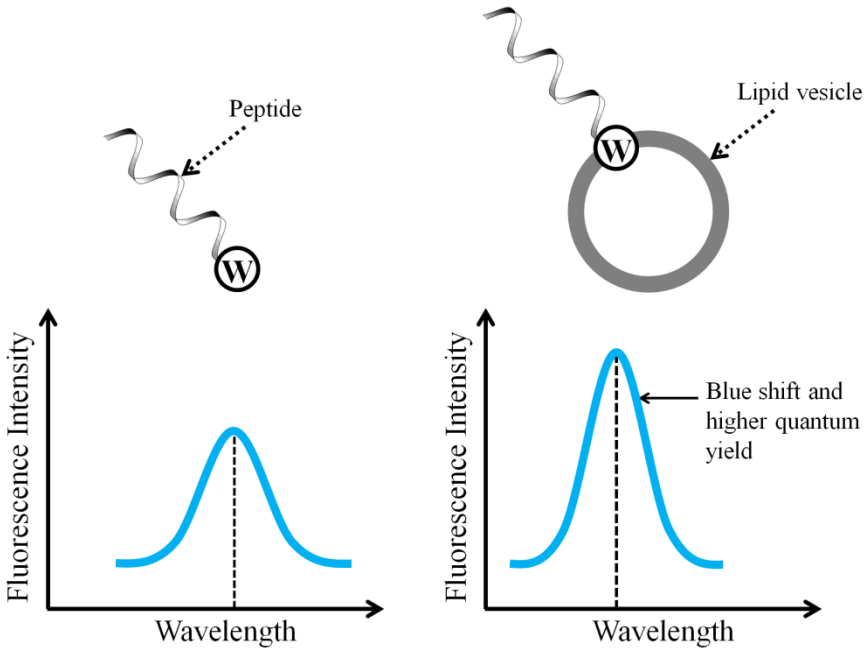


Figure 7.1 Spectral changes in tryptophan fluorescence upon binding to lipid bilayers

Binding studies: Binding of small fluorescent molecules to the biomacromolecules can be studied using fluorescence anisotropy. Binding of the fluorophore to a macromolecule will reduce its tumbling (increase its rotational correlation time) thereby resulting in higher fluorescence anisotropy.

FRET:

i. The distance between two sites in a biomacromolecule such as a protein can be calculated by labeling these sites with suitable donor-acceptor FRET pair. FRET can also be used to study the intermolecular interaction if the interacting molecules comprise of the fluorophores making a FRET pair.

ii. Interactions of peptides and other molecules with lipid bilayers comprising fluorophore labeled lipids. If the interacting molecule makes a FRET pair with the fluorescent lipid, the distance between them can be calculated providing information about the insertion of the molecule in the lipid bilayer.

iii. FRET has been utilized to study the kinetics of enzymatic reactions. For example, a DNA molecule, tagged with the fluorescence donor at one end and an acceptor at the other end can be used as a substrate to study the restriction endonuclease activity and cleavage reaction kinetics (Figure 7.2). A similar assay can be used to study the proteases using peptides as the substrates.

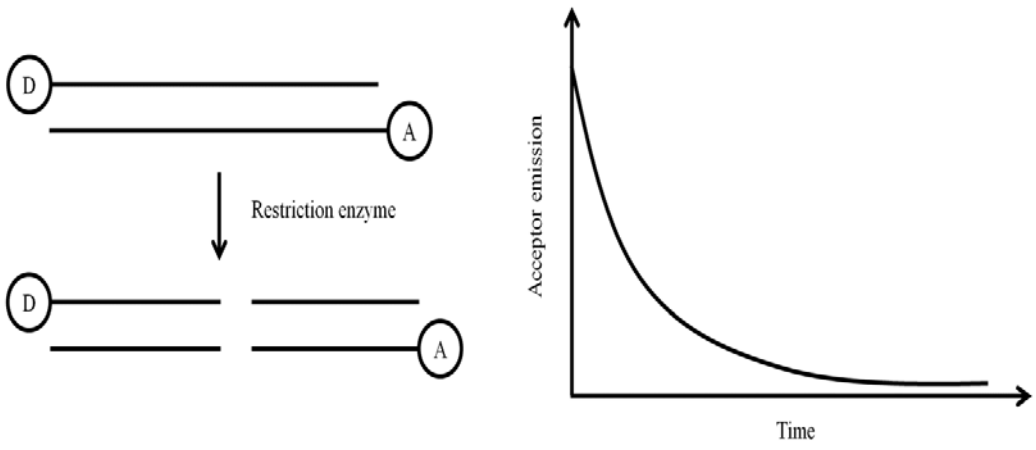


Figure 7.2 Decrease in fluorescence intensity of the acceptor following cleavage of DNA molecules.

Fluorescence quenching:

i. Interaction of a fluorophore with another molecule(s) may provide it protection against a collisional quencher. For example, interaction of a Trp containing peptide with lipid bilayers can be studied using iodide (I^-) as the collisional quencher. The peptide sample in the presence of lipid vesicle is titrated with the potassium iodide (KI) and fluorescence spectra recorded at each quencher concentration. The collisional fluorescence quenching is described by a plot of 'the ratio of quantum yield in the absence of quencher to that in the presence of quencher' against 'the quencher concentration'. Such a plot is known as the Stern-Volmer plot. The Stern-Volmer equation is given by

$$\frac{F_0}{F} = 1 + k_q\tau_0[Q] = 1 + K_{sv}[Q] \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (7.1)$$

where,

F_0 = Fluorescence intensity in the absence of quencher

F = Fluorescence intensity in the presence of quencher

k_q = Bimolecular quenching constant

τ_0 = Fluorescence lifetime in the absence of quencher

$[Q]$ = Quencher concentration

K_{sv} = Stern-Volmer constant

A normalized accessibility factor (NAF) is defined as the ratio of 'the K_{sv} in the presence of the binding partner of the fluorophore' to 'that without the binding partner'.

ii. The fluorescence intensity of a sample increases with an increase in the fluorophore concentration. Beyond certain concentration, however, the fluorescence intensity decreases due to self collisional quenching. This property is often used to study the membranolytic activities of a compound. A fluorescent dye at self-quenching concentrations is trapped inside a lipid vesicle. A membranolytic compound results in the release of the fluorescent dye causing increase in fluorescence emission intensity (Figure 7.3).

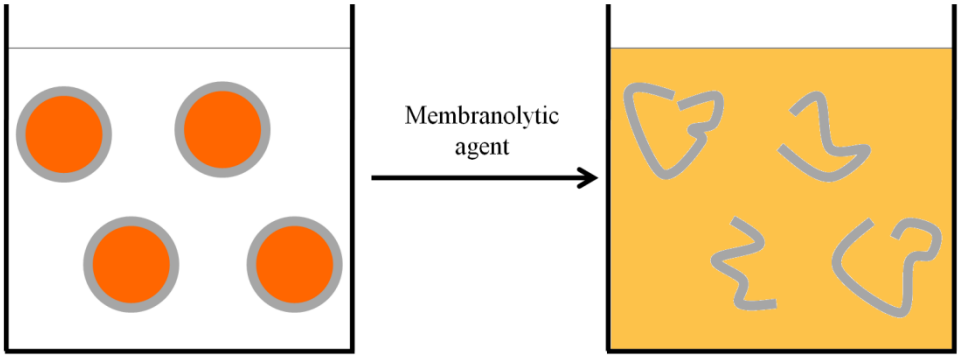


Figure 7.3 Membranolytic activity of a compound monitored through dye release assay. Release of dye from the lipid vesicle diminishes the self-quenching resulting in enhanced fluorescence emission.

iii. Fusion of lipid vesicles can also be studied using the same approach. Vesicles that contain self-quenching concentrations of the fluorescent dye are titrated with the vesicles without fluorophores. A fusion will result in the dilution of fluorophores; the consequent decrease in self-quenching is exhibited as an increase in the fluorescence intensity (Figure 7.4).

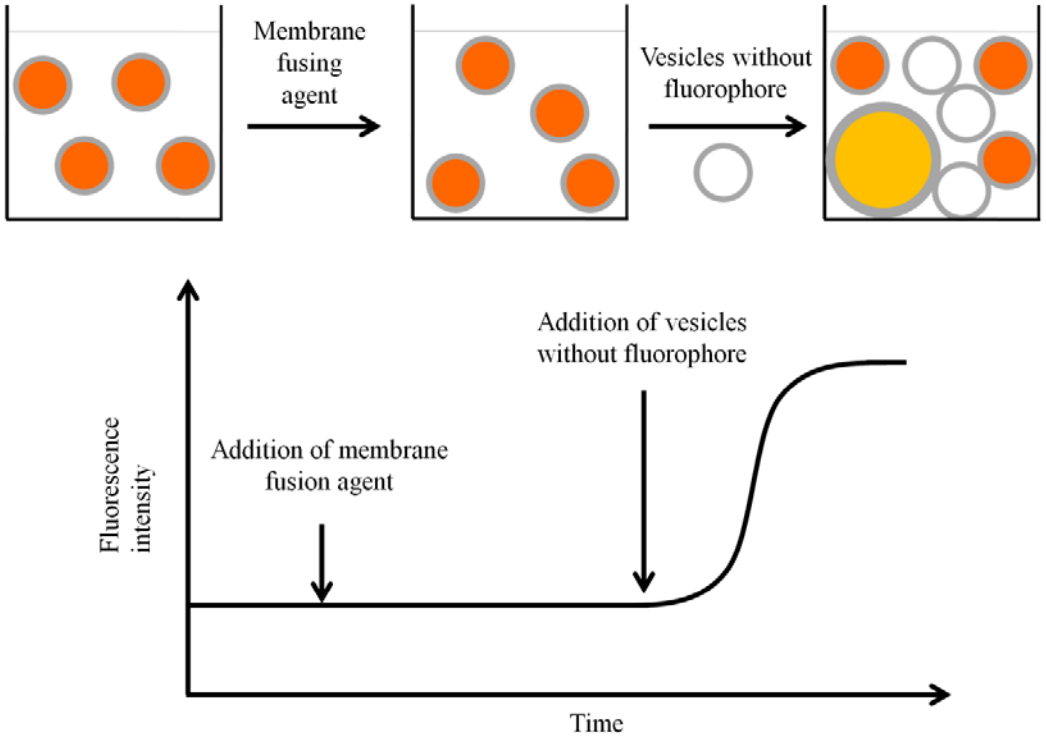
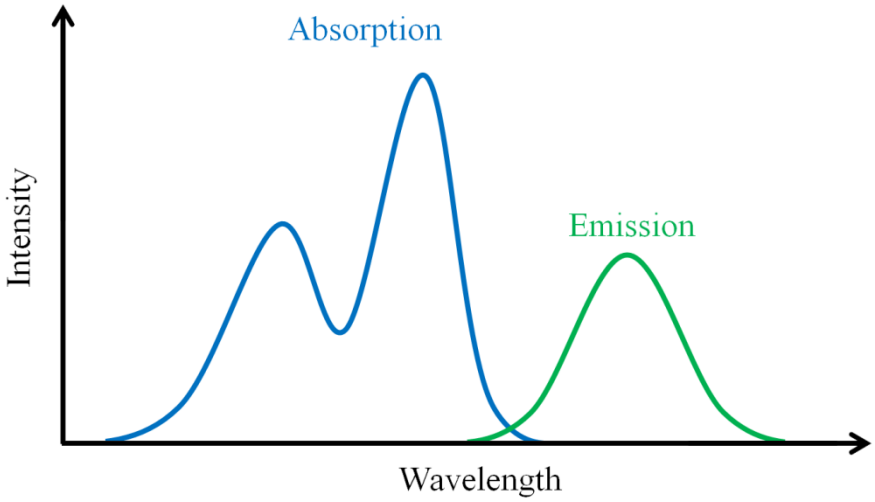


Figure 7.4 Fusion of fluorescent dye-containing lipid vesicles with vesicles without dye results in dilution of dye. The dilution results in lesser self-quenching thereby increasing the fluorescence intensity.

QUIZ

Q1: Shown below are the absorption and emission spectra of a fluorophore. The fluorescence emission for this fluorophore is not the mirror image of the absorption spectrum. How do you explain this?



Ans: The low wavelength absorption band is likely to be arising from $S_0 \rightarrow S_2$ transition. As fluorophore relaxes back to S1 state prior to emission, the fluorescence band is the mirror image of the band arising from $S_0 \rightarrow S_1$ transition, not the entire absorption spectrum.

Q2: If the efficiency of energy transfer between a donor and acceptor is 80%. Calculate the distance between them if the Förster distance between them is 40 nm?

Ans: The efficiency of energy transfer, *E* is given by:

$$E = \frac{R_0^6}{R_0^6 + r^6}$$

Given: E = 80% = 0.8, R_0 = 40 nm

Rearranging the expression for the efficiency of energy transfer

$$E = \frac{1}{1+\left(\frac{r}{R_0}\right)^6}$$

$$0.8 = \frac{1}{1+\left(\frac{r}{40}\right)^6}$$

$$1 + \left(\frac{r}{40}\right)^6 = \frac{1}{0.8} = 1.25$$

$$\left(\frac{r}{40}\right)^6 = 1.25 - 1 = 0.25$$

$$\frac{r}{40} = (0.25)^{\frac{1}{6}} = 0.7937$$

$$r = 0.7937 \times 40 = 31.748 \text{ nm} \approx 31.75 \text{ nm}$$

Lecture 8 Circular Dichroism Spectroscopy-I

Introduction

Before going ahead to see what circular dichroism (abbreviated as CD) means, let us have a quick revisit on the polarized light. Light, as we have discussed in lecture 3 is electromagnetic radiation where electric field and the magnetic field are always perpendicular to each other. From now on, we shall mention only electric field; it is implicit that at all points in time and space, the magnetic field vector is perpendicular to both the electric field vector and the direction of the propagation of light. Unpolarized light is comprised of several electromagnetic waves with their electric field vectors (and therefore magnetic field vectors also) pointing in all possible directions, but perpendicular to the direction of light propagation. If the vectors in all, but one, directions are cut off, the resulting radiation is a plane polarized light as the electric field vector is confined to one plane (Figure 8.1). Looking towards the light source will exhibit electric field fluctuations in one line; the plane polarized light is therefore also referred to as the linearly polarized light.

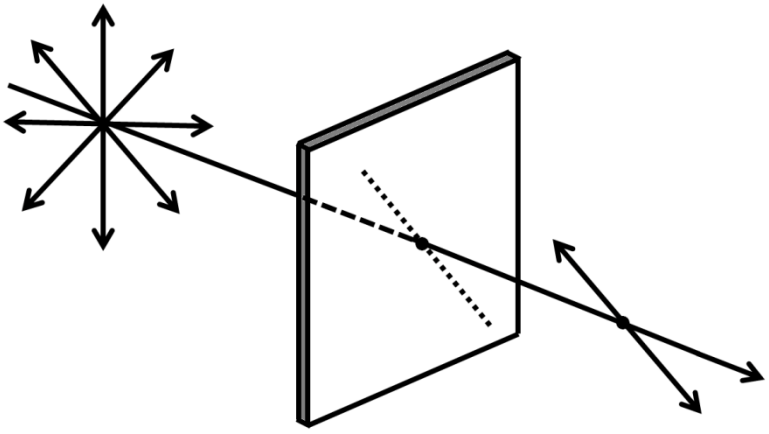


Figure 8.1 Plane polarized light produced by a linear polarizer

Superposition of polarized waves

Two electromagnetic waves can be superposed through vector addition of their electric field vectors. The properties of the resultant waves depend on the wavelength, polarization, and the phase of the superposing waves. In-phase superposition of two waves of same wavelength that are linearly polarized in two perpendicular planes results in a linearly polarized light with its electric field vector oscillating in a plane that is inclined at an angle of 45° to the polarization planes of both the waves (Figure 8.2).

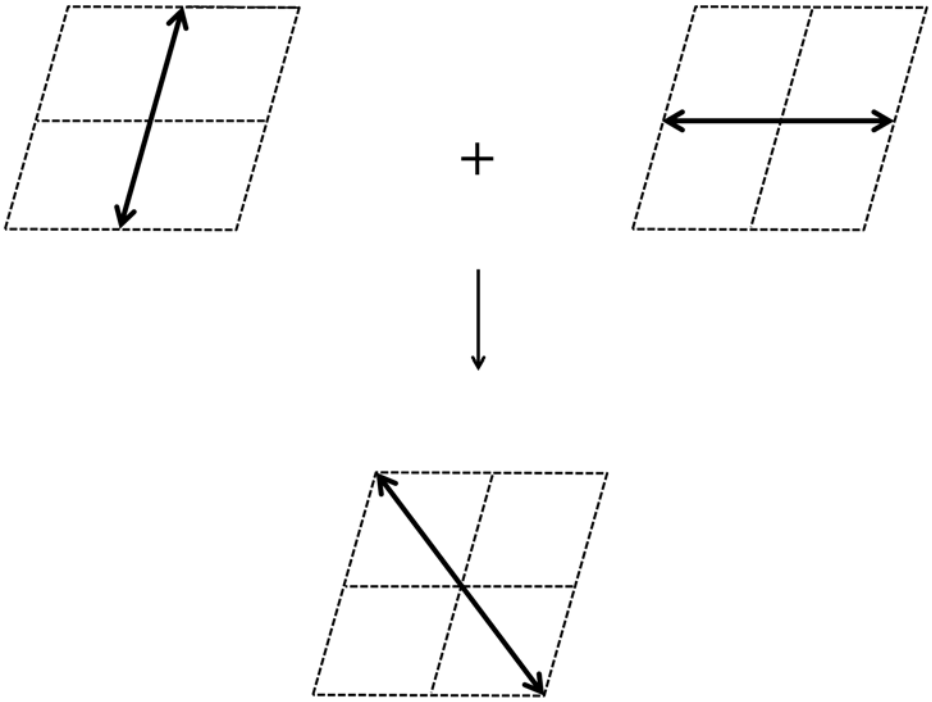


Figure 8.2 Superposition of linearly polarized waves

Let us see what happens when the two plane polarized waves, polarized in two perpendicular planes meet each other out of phase. Suppose the two waves have a phase difference of 90°. As the two waves have same wavelength, a 90° phase difference implies that when one of the wave is at maximum amplitude, the amplitude of the other one is minimum and vice versa. If the amplitudes of the two waves are equal, their superposition with a 90° phase difference results in a wave wherein electric field vector traverses a circular path (Figure 8.3). The electric field of the resultant wave is never zero but a vector of constant length. When looked at the travelling wave from the direction of propagation, the electric field appears to be

rotating in a circle. The resulting light is therefore termed as circularly polarized light (Figure 8.3).

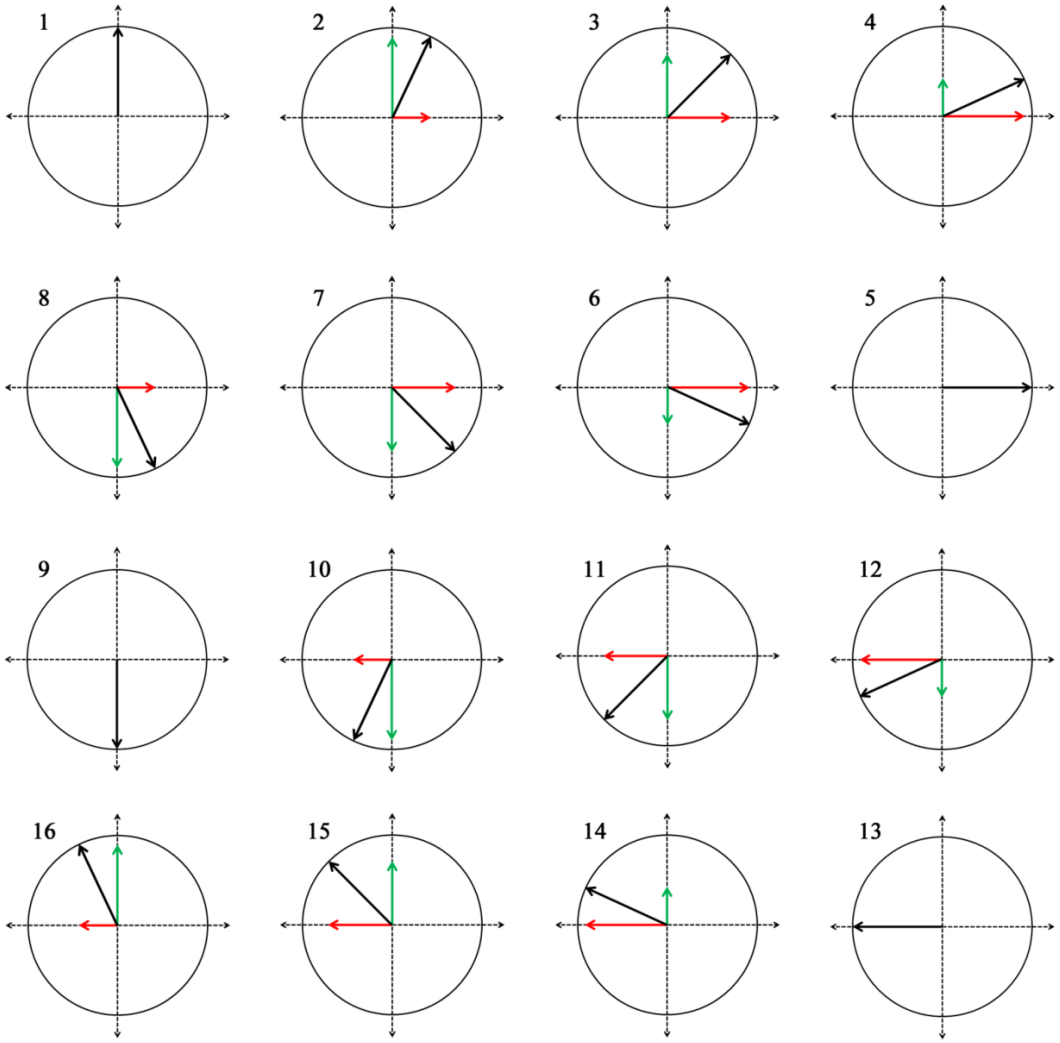


Figure 8.3 Superposition of waves linearly polarized in mutually perpendicular plain and that meet together 90° out of phase.

The direction of rotation depends on phase difference; a -90° phase difference would result in a circularly polarized light where the electric field rotates in opposite direction. When looked towards the light source, the electric field vector of a right circularly polarized wave appears to rotate counterclockwise in space while that of a left circularly polarized wave rotates clockwise. What happens when the right circularly polarized light (RCPL) and the left circularly polarized light (LCPL) superpose? The resultant wave is a linearly polarized wave (Figure 8.4). A linearly polarized light can therefore be considered as being composed of a right circularly polarized light and a left circularly polarized light.

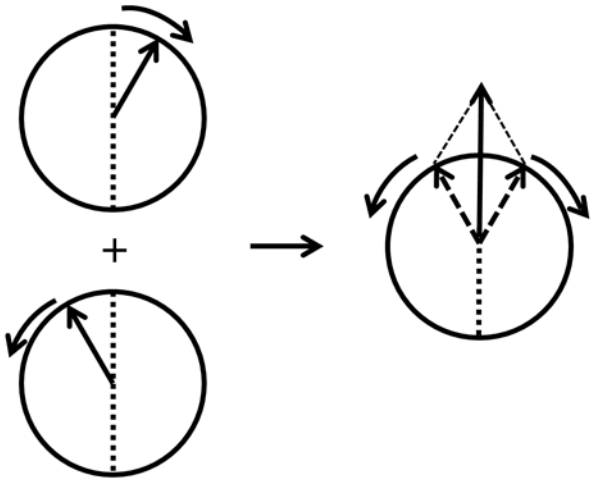


Figure 8.4 Superposition of left and right circularly polarized light resulting in plane polarized light.

Circular Dichroism

Circular dichroism, abbreviated as CD, is a chiroptical spectroscopic method. A chiral molecule or an achiral molecule in asymmetric environment interacts differently with the LCPL and the RCPL. The literal meaning of dichroism is 'two colors'. In chiroptical spectroscopy, dichroism means differential absorption of the lights with different polarizations. Circular dichroism, therefore, refers to the differential absorption of the left and right circularly polarized light and is defined as:

$$CD = \Delta A = A_l - A_r \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(6.1)$$

where, A_l and A_r are the absorbances for the left and right circularly polarized lights, respectively.

We can therefore say that the molar absorption coefficients for the two lights are different and can write the equation 6.1 can be written as:

$$CD = (\varepsilon_l - \varepsilon_r)cl \quad \cdots\cdots(6.2)$$

$$CD = \Delta\varepsilon cl \quad \cdots\cdots(6.3)$$

The preferential absorption of LCPL over RCPL (or vice versa) results in elliptical polarized light (Figure 8.5).

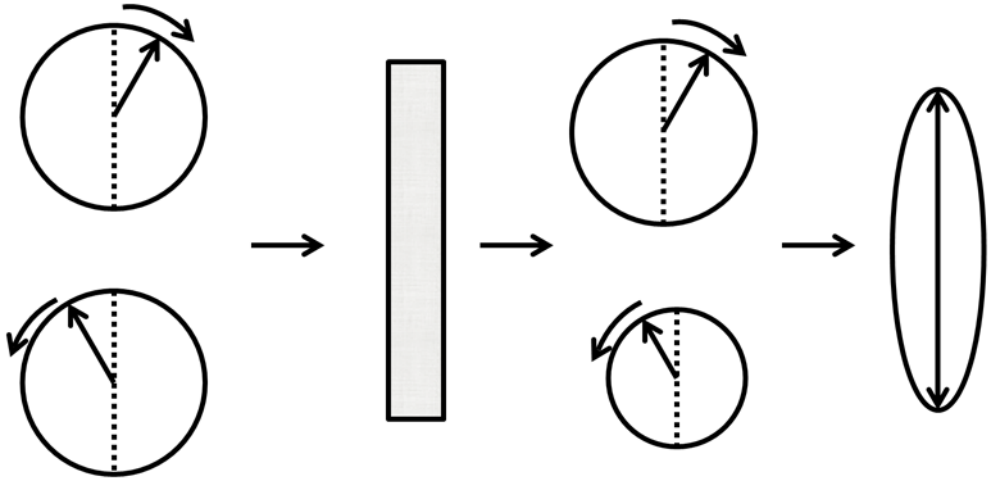


Figure 8.5 Differential absorption of the left and right circularly polarized light resulting in elliptically polarized light. Notice that if one component is completely absorbed, the resultant wave will be circularly polarized.

CD is historically represented in terms of ellipticity (θ) which is the tangent of ratio of minor to major axis of the ellipse. The relationship between CD and θ is give by:

$$\theta \text{ (radians)} = \frac{2.303}{4} \times CD \quad \cdots\cdots(6.4)$$

$$\theta \text{ (degrees)} = \frac{2.303}{4} \times CD \times \frac{180}{\pi} \quad \cdots\cdots(6.5)$$

$$\theta \text{ (degrees)} \approx 33.0 \times CD \quad \cdots\cdots(6.6)$$

A plot between ΔA or $\Delta\varepsilon$ or θ against the wavelength of light represents a CD spectrum. In this lecture, we shall be discussing only electronic CD. That means that we shall be looking at the electromagnetic region that causes electronic transition, which of course is UV/Visible region.

Circular birefringence

If a sample reduces the velocity of the LCPL and RCPL to different extents, the sample is said to be circularly birefringent and the phenomenon circular birefringence. Let us see what happens when the linearly polarized light (having two components, LCPL and RCPL) traverses a circular birefringent medium: the velocities of the two components are reduced to different extents *i.e.* they have different wavelengths in the

sample. After emerging from the samples, the wavelength is restored but two components can be out of phase. This results in the rotation of the polarization axis. If the material is not circularly dichroic, the plane of the linearly polarized light is rotated (Figure 8.6A). If the material is both circularly dichroic and birefringent, the plane polarized light will become elliptically polarized light with the major axis of the ellipse tilted with respect to the polarization axis of the incident polarized light (Figure 8.6B).

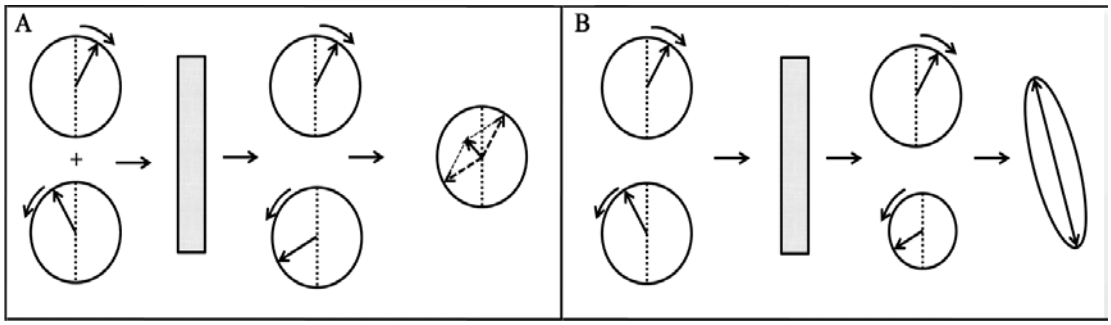


Figure 8.6 A linearly polarized light passing through a circular birefringent but not circular dichroic material (A) and through a material that is both circular birefringent and circular dichroic (B). Circular dichroism results in elliptically polarized light while circular birefringence causes change in the polarization axis.

Instrumentation

As CD is simply the difference in the absorbance of the LCPL and RCPL lights, a CD spectrometer, also known as a CD spectropolarimeter, is basically an absorption spectrophotometer (Figure 8.7). The instrument has a light source, usually a Xenon lamp. The polychromatic light from the source is converted to monochromatic radiation which is further converted to linearly polarized light by a polarizer. The linearly polarized light passes through a photoelastic modulator that alternately converts the linearly polarized light into LCPL and RCPL. The LCPL and the RCPL, therefore pass through the sample alternately and their absorbance gets recorded. Absorbance is recorded at various wavelengths to obtain a CD spectrum.

Photoelastic modulator: A photoelastic material is the one that exhibits birefringenece under mechanical stress. The photoelastic modulator in a CD instrument comprises of a quartz crystal fused to a piezoelectric material. Oscillations in the piezoelectric material drive the quartz crystal to oscillate at the same frequency. The crystal optical axis is at 45° to the linearly polarized light. The crystal retards one component of the light more than the other when compressed. When expanded the velocity of the two components gets reversed. A PEM, therefore gives alternating LCPL and RCPL.

Single wavelength CD values are also important in studying the fast reactions such as protein folding/unfolding (discussed in the next lecture).

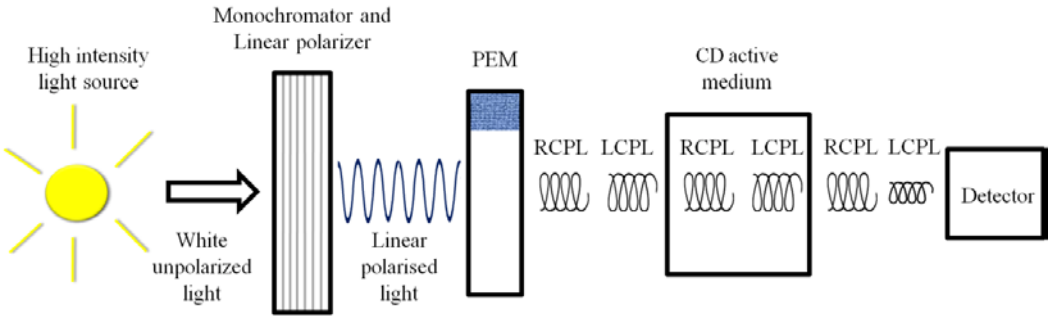


Figure 8.7 Schematic diagram of a CD spectropolarimeter.

Lecture 9 Circular Dichroism Spectroscopy-II

CD of biomolecules

Most biomolecules are chiral and the biomacromolecules are composed of chiral components. Folding of biomacromolecules into higher order structures further imparts them the asymmetry. CD has not been used as much to study other biomolecules probably, as it has been used to study proteins.

CD of proteins

Proteins are usually composed of 20 amino acids, 19 of which (except glycine) are chiral. This chirality also reflects in the higher order structures that the polypeptides adopt; α-helix, for example, is a right handed helix. If a polypeptide adopting α-helical structure is synthesized using D-amino acids, it folds into the left-handed α-helix under identical conditions. The other structural features of a polypeptide backbone include β-sheets, that are comprised of extended polypeptide chains; β-turns, that usually, but not essentially, link the β-strands in an antiparallel β-sheet; and unordered conformation. CD spectra of the proteins contain information about the asymmetric features of the polypeptide backbone. Furthermore, it can provide information about the orientation of the side chains. CD, therefore, is capable of providing information about the structure of proteins which in turn helps understanding their function. The chromophore that provides information about the conformation of the peptide backbone is the peptide bond (Figure 9.1); the spectra are therefore recorded in the far UV region, the region where peptide bond absorbs.

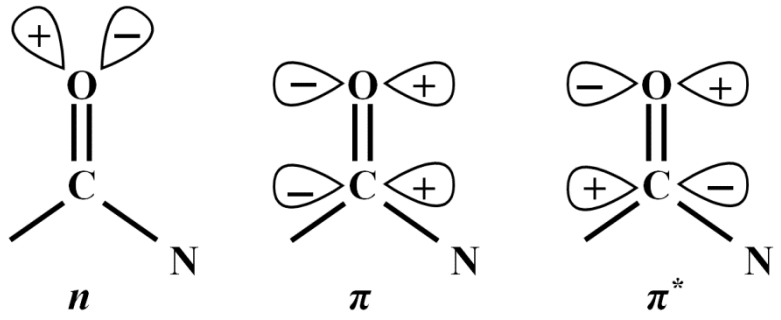


Figure 9.1 The peptide bond showing molecular orbitals involved in electronic transitions

Let us have a look at the CD spectra characteristic of the different structural components of the proteins (Figure 9.2).

- *α-helix:* The right handed α-helix displays two negative absorption bands centered around 222 nm ($n \rightarrow \pi^*$ transition) and 208 nm (a part of the $\pi \rightarrow \pi^*$ transition) and a strong positive band around 192 nm (a part of the $\pi \rightarrow \pi^*$ transition).
- *β-sheet:* β-sheets are characterized by the presence of a negative band centered around 216-218 nm ($n \rightarrow \pi^*$ transition) and a positive band of comparable intensity at around 195 nm ($\pi \rightarrow \pi^*$ transition).
- *β-turn:* A β-turn comprises of a four residue protein motif that causes the polypeptide backbone to take an approximately 180° turn. The CD spectrum for a β-turn is not well defined. A typical β-turn, however, shows a weak negative band around 225 nm ($n \rightarrow \pi^*$ transition), a strong positive band between 200 – 205 nm ($\pi \rightarrow \pi^*$ transition), and a strong negative band ($\pi \rightarrow \pi^*$ transition) between 180 – 190 nm.
- *Random coil:* Random coil or unordered conformation shows a weak positive band around 218 nm ($n \rightarrow \pi^*$ transition) and a strong negative band ($\pi \rightarrow \pi^*$ transition) below 200 nm.

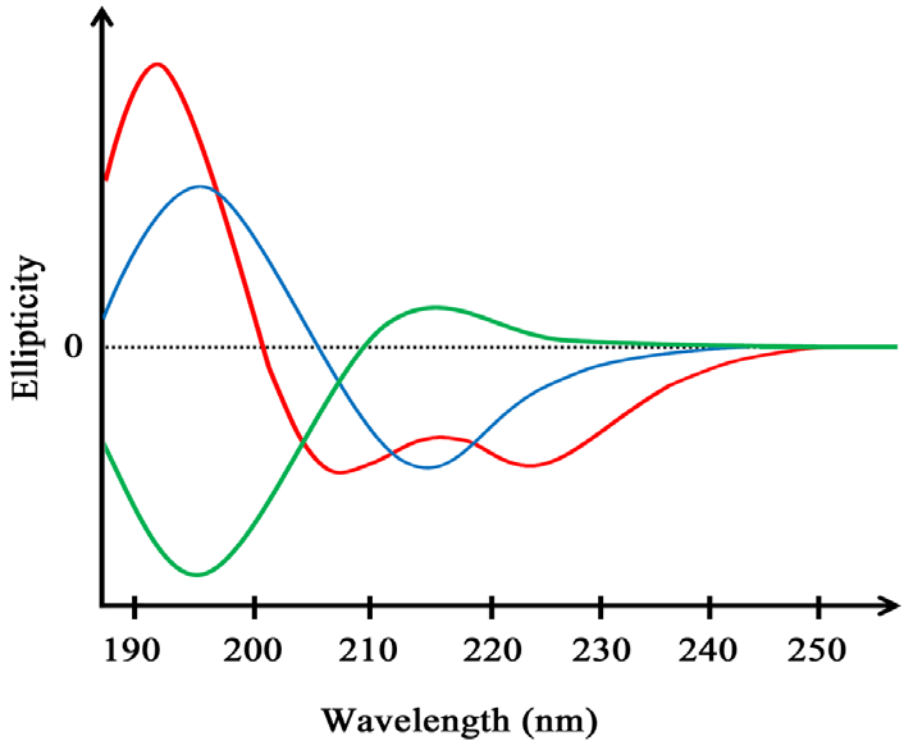


Figure 9.2 Far UV circular dichroism spectra of α-helix (red), β-sheet (blue), and unordered conformation (green)

The CD spectrum of a protein can be written as a linear combination of the spectra of all the structural components:

$$CD\ (protein) = a\ CD(\alpha\text{-helix}) + b\ CD(\beta\text{-sheet}) + c\ CD\ (Random\ coil)$$

As the CD spectra of different structural components are quite distinct, it is possible to estimate the fraction of different structural components in a protein from its CD spectrum. As discussed in lecture 5, proteins also have chromophores that absorb in the near UV region. These include the aromatic amino acids and disulfide linkages. The CD of aromatic amino acids is highly dependent on their environment and therefore near UV CD of proteins can provide the information about the environments these residues reside in as well as their orientations in the structure. As it provides information about the tertiary region, near UV CD is also referred to as tertiary CD in the context of the proteins.

CD of nucleic acids

As mentioned in lecture 5, nitrogenous bases constitute the chromophores of nucleic acids in the near and far UV region. The CD of the stacked bases is larger in magnitude as compared to that of the isolated bases. As the double helical nucleic acids have stacked bases, what we measure essentially is the CD that arises due to coupling of the chromophores. As the stacking geometries are different for different forms of nucleic acids such as B-DNA, Z-DNA, and A DNA; CD can help in determining which DNA form is present in a given sample.

Applications in biomolecular analysis

i. Determination of protein/peptide structure: As has already been discussed earlier, far UV CD spectroscopy provides information about the secondary structural elements in a protein. A mixture of structures can be deconvoluted to obtain the fraction of different structural elements. Furthermore, near UV CD provides information about the tertiary structure of the protein.

ii. Comparison of structures: Mutants of proteins are often required for understanding the functions of the proteins. It, however, needs to be ascertained that the mutation does not cause any significant change in the overall structure of the protein. CD spectroscopy happens to be a fast and extremely reliable tool to compare the conformations of the wild type proteins with their mutants.

iii. Stability of proteins: Stability of the proteins to denaturants or heat can be studied using CD spectroscopy. In such studies CD is usually monitored at a single wavelength, typically around 220 nm. Plotting the change in ellipticity against increasing denaturant concentration/temperature provides the denaturation curve. Figure 9.3 shows the denaturation curves for three related proteins. The denaturation curves suggest that the protein indicated with the blue trace is most stable while the one indicated with red trace the least.

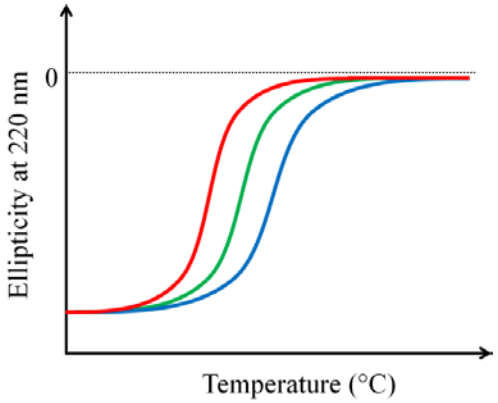


Figure 9.3 Comparison of thermostability of three related proteins. The blue trace represents the most stable protein.

iv. Binding of ligands to proteins: Binding of a ligand to a protein usually does not affect the secondary structural elements significantly. However, such a binding can cause changes in the local tertiary structure. Binding of ligands accompanying such conformational changes can be studied using tertiary CD if the binding region happens to have one or more aromatic residues. Short peptides, on the other hand, can undergo large scale structural changes sometime involving completely switching from one secondary structure to another. Such changes can easily be observed using far UV CD.

v. DNA structure: CD in the 200 – 300 nm region can be used to identify which structural isoform of DNA is present in the given sample. The left-handed helical DNA form, the Z-DNA was indeed identified using CD spectroscopy. The typical CD signatures of the B, Z, and A form of DNA are:

B-DNA: In its most common form *i.e.* B-DNA with ~10.4 bases per turn, a positive band ~275 nm, a crossover ~258 nm, and a negative band at ~240 nm are observed.

Z-DNA: A negative band ~290 nm and a positive band ~260 nm; a crossover between 180-185 nm.

A-DNA: A positive band ~260 nm, a negative band ~210 nm.

vi. Protein folding/unfolding: CD is used for studying the folding and unfolding of proteins. For monitoring the fast reactions such as protein folding, a single wavelength CD is recorded in a stopped flow experiment wherein the protein solution is mixed with a denaturant and CD is recorded as a function of time. Modern instruments take ~1 millisecond time between mixing and recording data allowing the understanding of the folding/unfolding events that occur on milliseconds to seconds timescale. A diagrammatic unfolding experiment is shown in figure 9.4

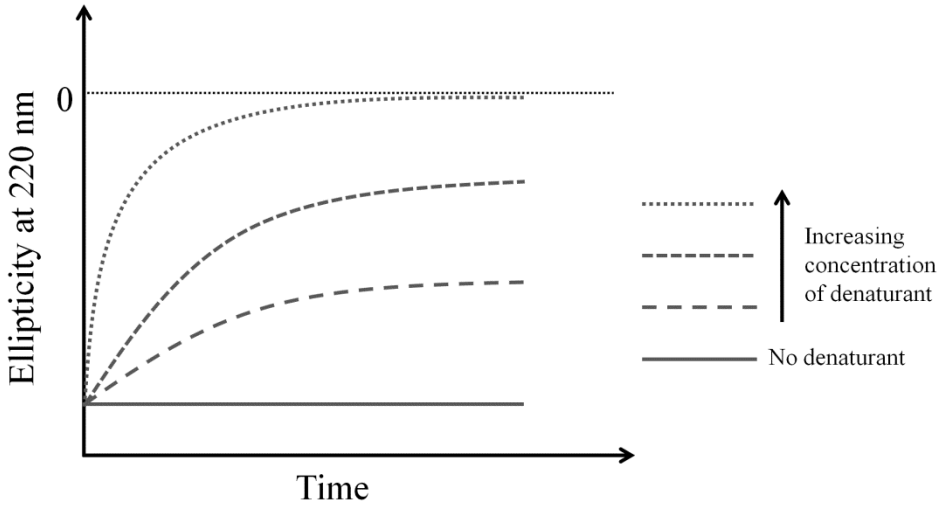


Figure 9.4 A diagram showing the kinetics of unfolding of a hypothetical protein. The protein is unfolded with different concentrations of a denaturant. Protein and denaturant are mixed in a stopped flow apparatus (mixing time typically ~1 ms) and changes in ellipticity are monitored over time.

vii. Molecular self-assembly: Self-assembly into structural and functional superstructures is integral to biomolecules and therefore to living systems. Inspired by the naturally occurring superstructures, short peptides have attracted considerable attention as the monomers for designing superstructures with novel properties and applications in biomedicine. Circular dichroism has been central in elucidating the conformations of the peptides in superstructures as well as the interactions that drive this assembly.

Circular dichroism, therefore, is a powerful tool in studying the conformations of biomolecules as well as the processes these molecules are involved in.

Lecture 10 Infrared Spectroscopy

Introduction

Infrared (IR) region of the electromagnetic spectrum lies between visible and microwave regions and therefore spans the wavelengths from 0.78 – 250 μm. The energies associated with molecular vibrations are smaller than those associated with electronic transitions and fall in the IR region. IR spectroscopy, therefore, is used to probe the vibrations in molecules and is also known as vibrational spectroscopy. Infrared region is usually divided into three regions: near infrared, mid-infrared, and far infrared (Figure 10.1). IR spectroscopists use wavenumbers ($\bar{\nu}$) to represent the IR spectra and we shall be following the same convention. Mid-IR region (λ = 2.5 -25 μm; $\bar{\nu}$ = 4000 – 400 cm^{-1}) is the region of interest for studying molecular vibrations.

Conventions for IR radiation

Wavelength: The wavelength of IR region ranges from ~780 nm – 250000 nm. Writing such big number is avoided by expressing the wavelengths in micrometers (0.78 – 250 μm).

Wavenumber ($\bar{\nu}$): Wavenumber means the number of wavelengths per unit distance. Therefore, 100 cm^{-1} implies there are 100 wavelengths per cm. $\bar{\nu}$ in cm^{-1} is given by:

$$\bar{\nu}\ (cm^{-1}) = \frac{1}{\lambda\ (\mu m)} \times 10^4$$

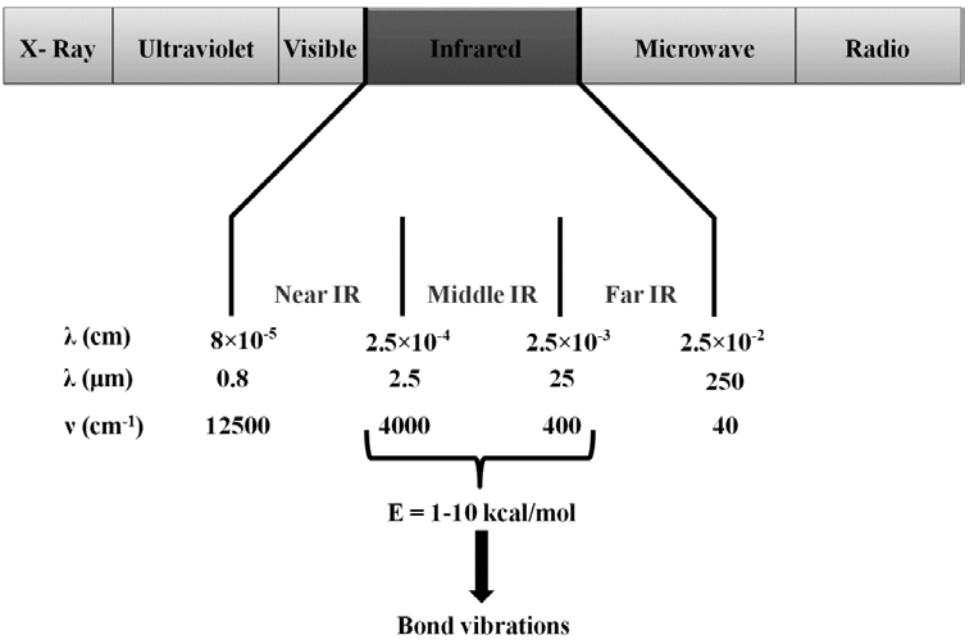


Figure 10.1 Infrared region of the electromagnetic spectrum

Degrees of freedom and molecular vibrations

At non-zero temperatures, *i.e.* temperatures above 0 K, all the atoms in a molecule are in motion. The molecule itself also is in translational and rotation motion. In a three dimensional space, an atom in isolation has 3 degrees of freedom, corresponding to the motion along the three independent coordinate axes. *A molecule composed of N atoms has a total of 3N degrees of freedom* (Figure 10.2).

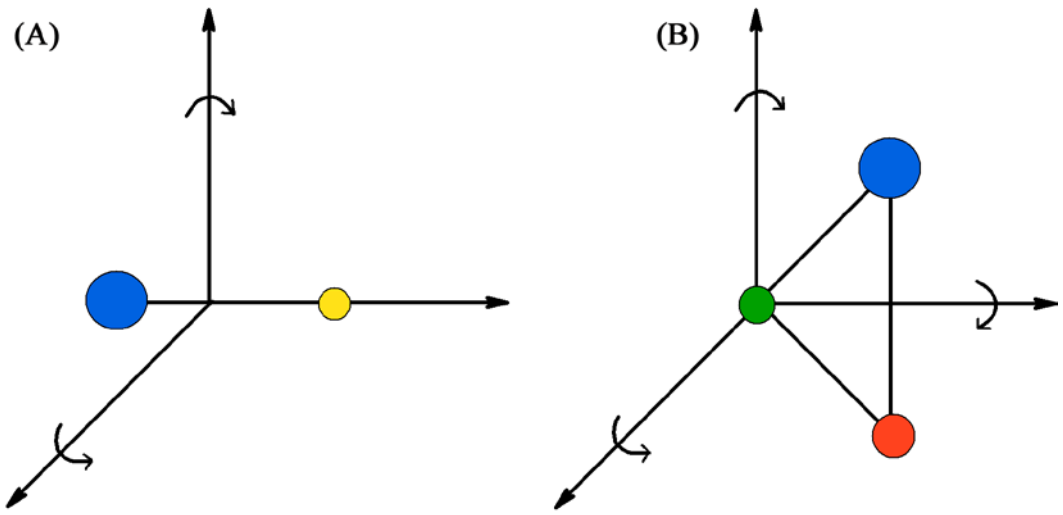


Figure 10.2 Degrees of rotational freedom for a diatomic (A) and a triatomic (B) molecule

For a non-linear molecule, three of these 3*N* degrees of freedom correspond to translational motion, three correspond to rotational motion while rest 3N-6 are the vibrational degrees of freedom. For a linear molecule, there are only two rotational degrees of freedom that correspond to the rotation about the two orthogonal axes perpendicular to the bond (Figure 10.2). A linear molecule, therefore, has 3*N*-5 vibrational degrees of freedom. Let us have a look at the degrees of freedom of a diatomic molecule. A diatomic molecule has a total of 3 × 2 = 6 degrees of freedom. *Three* of these *six* degrees of freedom correspond to translational motion of the molecule; *two* of them define rotational degrees of freedom; while *one* corresponds to the vibration of the atoms along the bond. The 3N-6 vibrational degrees of freedom (3N-5 for linear molecules) represent the true/fundamental modes of vibration of a molecule. The different types of vibrations are shown in Figure 10.3.

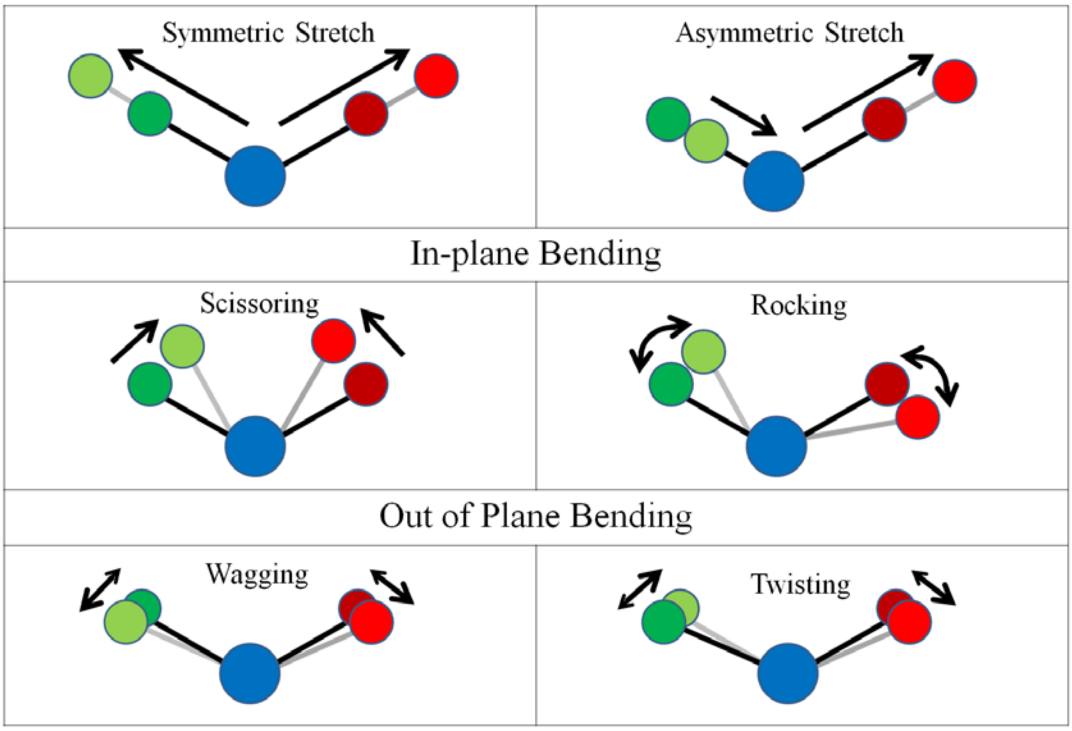


Figure 10.3 Stretching and bending vibrations in molecules

Hooke's law and frequency of vibration

We have seen that the bonds are not static but vibrating in different ways. A vibrating bond can therefore be considered a spring with its ends tethered to two atoms (Figure 10.4).

Figure 10.4 Spring analogy of a bond vibration

If the masses of the atoms are m_1 and m_2, the frequency of stretching vibration of the diatomic molecule can be given by the Hooke's law:

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{\mu}} \qquad \cdots\cdots (10.1)$$

where, ν is the frequency of vibration, k is the spring constant, and μ is the reduced mass *i.e.* $\frac{m_1 m_2}{m_1 + m_2}$

Dividing equation 10.1 by λ gives:

$$\frac{\nu}{\lambda} = \frac{1}{2\pi\lambda}\sqrt{\frac{k}{\mu}} \qquad \cdots\cdots (10.2)$$

$$\frac{1}{\lambda} = \frac{1}{2\pi(\lambda\nu)}\sqrt{\frac{k}{\mu}} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (10.3)$$

$$\bar{\nu} = \frac{1}{2\pi c}\sqrt{\frac{k}{\mu}} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (10.4)$$

Anharmonic oscillator

Real molecules are anharmonic oscillators. Unlike harmonic oscillator wherein energy levels are equally spaced; energy levels in an anharmonic oscillator are more closely spaced at higher interatomic distances. A treatment for anharmonicity is beyond the scope of our discussion.

The spring constant, *k* is the measure of the bond strength. The stronger the bond, the higher the *k*, and consequently the higher is the frequency of vibration. This treatment implies that the diatomic molecule is a simple harmonic oscillator. The energy of a quantum harmonic oscillator is given by:

$$E = \left(n + \frac{1}{2}\right)h\nu \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (10.5)$$

where, n = 0, 1, 2, …… and h is the Plancks's constant

Absorption of infrared radiation

A molecular vibration is IR active i.e. it absorbs IR radiation if the vibration results in a change in the dipole moment. A diatomic molecule, that has one mode of vibration, may not absorb an IR radiation if the vibration does not accompany a change in the dipole moment. This is true for all the homonuclear diatomic molecules such as H_2, N_2, O_2, etc. Vibration of carbon monoxide (C=O), on the other hand, causes a change in dipole moment and is therefore IR active. Vibration of a bond involving two atoms that have large electronegativity difference is usually IR active.

An IR active vibration of a particular frequency absorbs the IR radiation of same frequency. Let us calculate the position of absorption band for carbonyl stretching vibration (frequency = 5.1×10^{13} vibrations/second) in acetone.

$$\bar{\nu} = \frac{1}{\lambda} = \frac{\nu}{c}\ \text{cm}^{-1}$$

$$\bar{\nu} = \frac{5.1 \times 10^{13}\ sec^{-1}}{3 \times 10^{10}\ cm/sec} = 1700\ \text{cm}^{-1}$$

Instrumentation

Two types of infrared spectrometers are commercially available: dispersive and Fourier Transform infrared (FTIR) spectrometers.

Dispersive spectrometer: A dispersive spectrometer is very similar in design to a UV/visible spectrophotometer. It has a radiation source, a grating monochromator, and a detector. The IR radiation generated by the source is dispersed into different frequencies by a monochromator. The selected frequencies go through sample and reference cells and the transmitted light is measured by the detector. The infrared sources are usually inert solids that are electrically heated to radiate infrared radiation. The detectors usually are either thermal sensors such as thermocouples and thermistors or the semiconductor materials that conduct following absorption of IR radiation (absorption of photon causes transition of electrons from the valence band to the conduction band).

Fourier Transform Spectrometer: A Fourier transform spectrometer uses an interferometer in place of the monochromator. An interference of polychromatic radiation is generated using an interferometer, usually a Michelson interferometer (Please see Box 10.1). Absorption of any particular wavelength will bring a change in the interferogram which gets detected. An interferogram is a time domain signal and is converted to frequency domain signal though Fourier Transformation.

Dispersive infrared spectrometers are still in use but FTIR spectrometers are slowly taking over. FTIR spectrometers have several advantages over the dispersive ones:

i. Better speed: FTIR spectrometers detect absorption of all the frequencies simultaneously; consequently, they are much faster than the dispersive spectrometers that scan the entire frequency range stepwise.

ii. Better sensitivity: Their speed of data acquisition makes FTIR spectrometers more sensitive. A large number of spectra can be recorded in small time thereby giving an improved signal to noise $\left(\frac{S}{N}\right)$ ratio.

$$\left(\frac{S}{N}\right) \propto \sqrt{n} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (10.6)$$

where, *n* is the number of independent measurements

iii. More radiation energy: Dispersive spectrometers use slits that result in loss of radiation energy. FTIR spectrometers lack the slits as filtering of radiation is not required; this provides higher radiation energy for recording the absorbance.

iv. Simple design: As dispersion of the radiation and filtering are not required in the FTIR spectrometer, the movable mirror is the only moving part in the spectrometer.

v. Wavelength accuracy: The FTIR spectrometers usually have a He-Ne laser emitting light of 632.8 nm. This serves as an internal calibration for the wavelength and provides an accuracy of 0.01 cm^{-1} or better.

Attenuated Total Reflectance – Fourier Transform Infrared Spectrometer (ATR-FTIR)

An ATR-FTIR works on the principle of total internal reflection and the evanescent field (Figure 10.5). The refractive index of the ATR crystal (usually Zinc selenide, diamond, Germanium) is significantly higher than that of the samples that are to be studied. An IR beam gets refracted at the interface of the ATR material and the sample (Figure 10.5A).

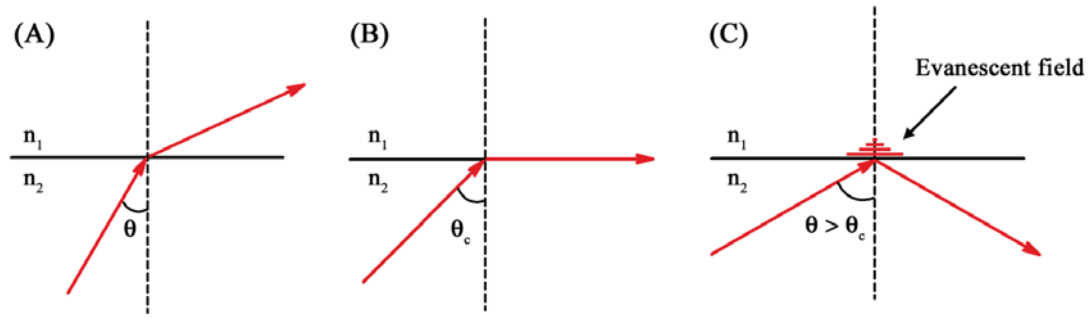


Figure 10.5 Phenomenon of total internal reflection. Notice the exponentially decaying evanescent field in the medium of lower refractive index (n_1)

If the angle of incidence, θ is more than the critical angle, θ_c (Figure 10.5B and C), the beam gets totally internally reflected. Before getting totally reflected, however, an exponentially decaying field penetrates into the medium of low refractive index. This field is called an evanescent field and can interact with the molecules that are in the close proximity of the ATR crystal. If some part of the evanescent field is absorbed by the molecules, the reflected beam will be attenuated (become less intense) by this factor. The reflected beam will therefore be of lesser intensity implying absorption of radiation. The commercially available ATR instruments are FTIR spectrometers. ATR-FTIR allows studying the samples like thin films, powders, pastes by directly placing the sample on the ATR crystal.

Functional group region and fingerprint region

The most common application of IR spectroscopy is perhaps to identify the functional groups. This is possible because different functional groups vibrate at different frequencies allowing their identification. The frequency of vibration, however, depends on additional factors such as delocalization of electrons, H-bonding, and substitutions at the nearby groups. The wavenumbers for some of the bonds are shown in Table 10.1.

Table 10.1 Typical vibrational frequencies of functional groups

Bond	**Molecule**	**Wavenumber (cm^{-1})**
C–O	Alcohols, ethers, esters, carboxylic acids, etc.	1300 – 1000
C=O	Aldehydes, ketones, esters, carboxylic acids	1750 – 1680
C=O	Amides	1680 – 1630
N–H (Stretching)	Amines and amides	3500 – 3100
–N–H (Bending)	Amines and amides	1640 – 1550
O–H	Alcohols	3650 – 3200
C–N	Amines	1350 – 1000
S–H	Mercaptans	2550

The absorption bands in the 4000 – 1500 cm^{-1} region help in the identification of functional groups; this region therefore is also termed the functional group region of the IR spectrum (Figure 10.6). The lower energy portion of the mid-IR region (1500 – 400 cm^{-1}) usually contains a very complicated set of peaks arising due to complex vibrations involving several atoms. This region is unique to a particular compound and therefore is known as the fingerprint region of the IR spectrum. Though it is

difficult to assign the vibrational modes to these peaks, these are useful to identify a compound if the spectrum of the compound is already known.

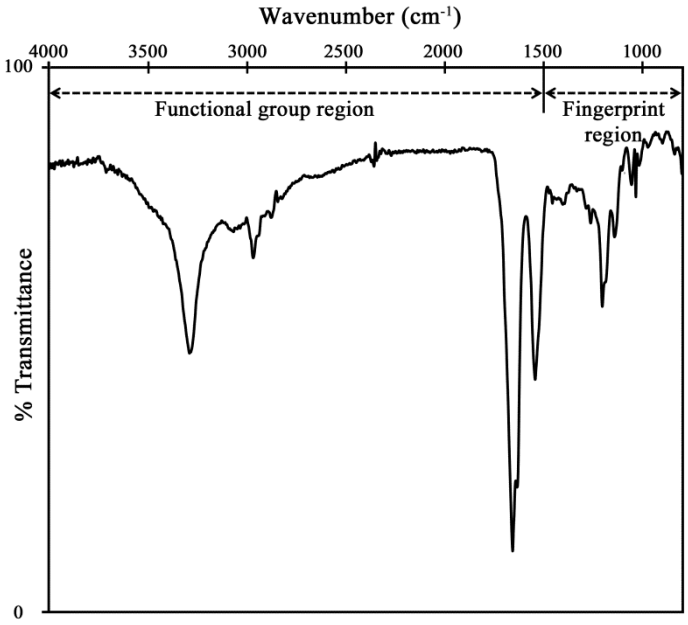


Figure 10.6. A typical IR spectrum showing functional group and fingerprint regions

Applications

i. *Identification of functional groups:* As has already been discussed, IR spectroscopy allows identification of functional groups. Carbonyl (C=O) is an interesting functional group worth discussing. Carbonyl is a double bond (high spring constant, *k*) with very high polarity. Stretching vibration of carbonyl group causes large changes in the dipole moment consequently resulting in a very intense absorption band. Furthermore, the frequency of carbonyl stretching does not differ significantly for aldehydes, ketone, carboxylic acids, and esters (Table 10.1). The large intensity and relatively unchanged frequency of carbonyl stretching allows easy identification of the carbonyl compounds (It is important to note that carbonyl stretching frequency can be much lower for amides and much higher for anhydrides and acid chlorides).

ii. *Identification of compounds:* The fingerprint region of the IR spectrum is unique to each compound. It is possible to identify a compound from its IR spectrum if the spectrum for the compound is already known and available for comparison. This is particularly useful in pharmaceutical research and development. A patented drug, if suspected to be synthesized by another

pharmaceutical company, can easily be identified by comparing the IR spectra in the fingerprint region.

iii. *Presence of impurities:* Comparison of the IR spectra of the given compound with the spectra of pure compound helps in the assessment of its purity. It is important to ascertain the purity of the active molecule and the excipients used in preparing drug formulations.

iv. *Structural transitions in lipids:* Structural lipids are those that are organized in bilayers in biological membranes. Glycerophospholipids constitute the major class of the structural lipids (Figure 10.7). The lipids have several structural phases such as a gel phase with all-*trans* conformation and a liquid crystalline phase where *gauche* conformations are also present. Methylene ($–CH_2–$) stretching vibrations give the most intense absorption band in lipids as expected for a molecule having long hydrocarbon chains.

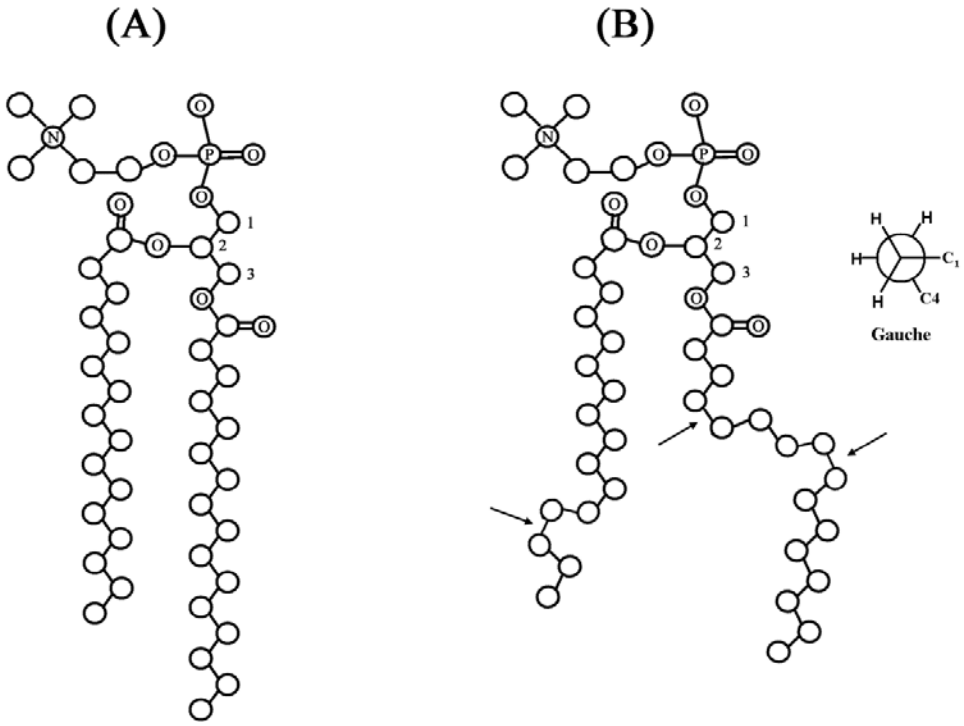


Figure 10.7. Structure of a glycerophospholipid: all-*trans* conformation (A); lipid with *trans* and *gauche* conformations (B), *gauche* conformations are indicated with arrows

Both $–CH_2–$ stretching and bending vibrations are sensitive to the conformations of the lipids and therefore provide information about the transition of lipids between different phases. Vibration modes of the head group and the interfacial region also provide useful information about local acyl chain conformation. Carbonyl stretching vibration (1750 – 1700 cm^{-1}) in the ester bond is sensitive to the conformation of the local acyl chain conformation.

v. *Protein and peptide structure:* Infrared spectroscopy is routinely used to study the structures of proteins and peptides. Like CD spectroscopy, the region of interest in determining the conformation of the polypeptide backbone is the peptide bond. The peptide group results in nine distinct bands, labeled as amide A, B, and I-VII. Amide I is the most useful band in studying the polypeptide backbone conformation. Amide I band (1700 – 1600 cm^{-1}) arises largely due to the carbonyl stretching with small components of C–N stretching and N–H bending. The frequency of carbonyl stretching vibration is sensitive to the H-bonding, and therefore to the conformation of polypeptide backbone. The frequencies of absorption of different secondary structural elements are shown in Table 10.2

Table 10.2 Vibrational frequencies of the secondary structural elements of proteins in H_2O

Structure	**Wavenumber (cm^{-1})**
α-helix	1657 – 1648
β-sheet	1641 – 1623
Unordered	1657 – 1642
Antiparallel β-sheet	1695 – 1675

There is considerable overlap of the bands arising from α-helical and the unordered conformations. It is therefore generally difficult to assign the bands appearing in this region. Recording an IR spectrum in D_2O decreases this overlap to some extent. Dissolution a protein in D_2O results in the exchange of solvent exposed amide protons by deuterium. Hydrogens of the unordered amides are more easily exchanged as compared to those involved in the secondary structures. Despite this, it is not easy to unambiguously assign the bands arising in the 1657- 1648 cm^{-1} region. Circular dichroism and IR spectroscopy therefore complement each other wherein α-helices are easily detected by CD and β-sheets by IR. Like CD, an IR spectrum of a protein can also be deconvoluted to determine the fractions of different secondary structural elements as shown in Figure 10.8.

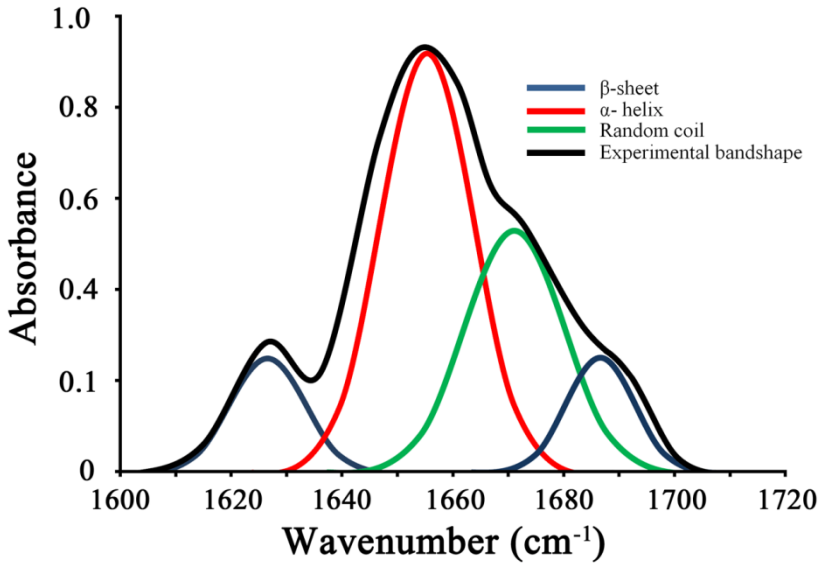
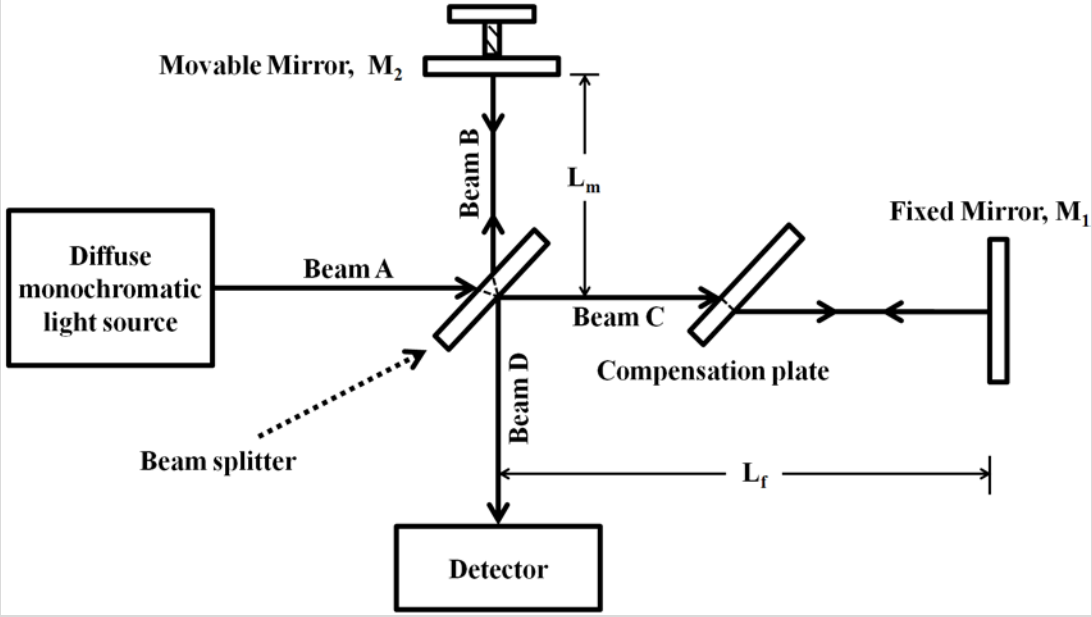


Figure 10.8. Deconvolution of Amide I band of a protein to identify the fractions of different structural elements

Box 10.1: Michelson interferometer

A Michelson interferometer has a radiation source, a collimator, a beam-splitter, a movable mirror, a fixed mirror, a compensator, and a detector.



The radiation coming from the source is collimated and focused on the beam-splitter. 50% of the radiation gets transmitted while 50% gets reflected. The mirrors reflect back the radiation towards the beam splitter that again allows 50% transmission and 50% reflection. This allows the beams, B and C to interfere and give the beam D. As the beam, B travels through the beam-splitter twice while beam C does not travel through it even once, a compensation plate of same material (un-mirrored) and thickness as the beam-splitter is, is placed between the beam-splitter and the fixed mirror. This allows the beams, B and C to travel the equal distance. The motion of the movable mirror, M_2 causes the two beams to travel different distances thereby generating interference. Let us see what happens when a monochromatic radiation is used in the Michelson interferometer. If the beams, B and C travel the equal distance, they are in phase and will interfere constructively. If however, the M_2 moves, say towards the beam-splitter by a distance of $\frac{\lambda}{4}$, the beam B travels a distance of $\frac{\lambda}{2}$ less than that travelled by beam C. This allows a phase difference of 180° resulting in destructive interference. A continuous motion of the mirror M_2, therefore, will generate a sinusoidal signal through interference. The detector therefore detects a time domain signal. If a sample placed before the

detector absorbs this radiation, the intensity of the light goes down. If a polychromatic light is used, the interference pattern generated carries all the wavelengths present in the polychromatic light. Absorption of any wavelength will result in a change in the interfering pattern. The interfering pattern, also known as interferogram is then Fourier transformed to obtain the frequency domain data.

QUIZ

Q1: If the stretching frequency of a hydrogen molecule is 1.2 × 10^{14} vibrations/sec. Calculate the wavenumber where hydrogen molecule absorption band will be observed in an IR spectrum.

Ans: The frequency of hydrogen stretching can be represented in terms of wavenumbers as follows:

$$\bar{\nu} = \frac{1}{\lambda} = \frac{\nu}{c} \text{ cm}^{-1}$$

$$\bar{\nu} = \frac{1.2 \times 10^{14}\ sec^{-1}}{3 \times 10^{10}\ cm/sec} = 4000 \text{ cm}^{-1}$$

Hydrogen, however, is a homodiatomic molecule; the stretching vibration does not cause any change in the dipole moment. Therefore, hydrogen will not absorption the IR of 4000 cm^{-1} and consequently will not appear in an IR spectrum.

Lecture 11 Mass Spectrometry-I

Mass spectrometry (abbreviated as MS) has slowly emerged as a very powerful tool in analyzing the organic molecules including biomolecules. A mass spectrometer separates the molecules based on their mass and charge. The underlying principle is conceptually very simple: a moving charged particle can be deflected by applying electric and magnetic fields. The deflection caused by the electric and magnetic fields depends on the mass and the charge of the particle. Let us see what happens to a charged particle in an electric field (Figure 11.1).

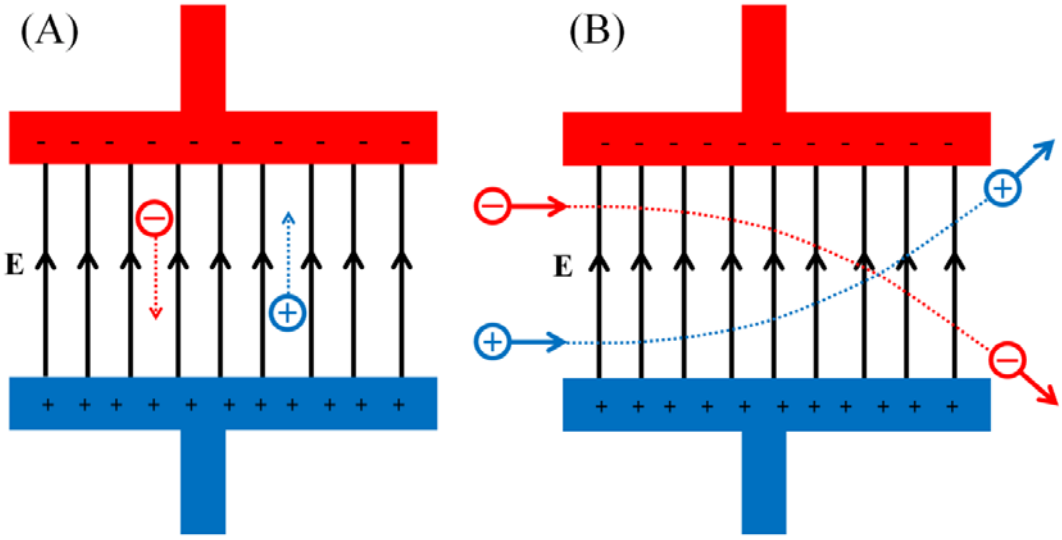


Figure 11.1 Force experienced in an electric field by stationary charged particles (A) and charged particles with uniform motion (B). Panel B represents the situation when the charged particles are in uniform motion with their initial velocity vectors perpendicular to the electric field vector.

The force experienced, F by a particle with charge, q in an electric field, E is given by:

$$F = qE \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (11.1)$$

The force causes the particle to accelerate in the electric field which is given by

$$ma = qE \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (11.2)$$

$$a = \frac{qE}{m} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (11.3)$$

where, m is the mass of the particle and a is the acceleration under the electrostatic force.

Equation 11.3 shows that the acceleration of the particle depends on the mass to charge ratio, $\frac{m}{q}$. A lighter particle is accelerated more than a heavier particle carrying the same charge. Similarly, a particle with higher charge is accelerated more as compared to the particle of same mass but having lesser charge.

In a magnetic field, a moving charged particle experiences a force, *F* that is given by the Lorentz force law:

$$F = q\ (v \times B) \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (11.4)$$

where, v is the velocity of the moving charged particle and B is the magnetic field strength.

The direction of the force can be determined using the right hand rule; If the fingers represent the magnetic field (*B*) and the thumb represents the velocity (v), then the direction of the force is given by the direction of the palm (Figure 11.2A). As the force is always perpendicular to the velocity, the deflected particle moves in a circular path (Figure 11.2B).

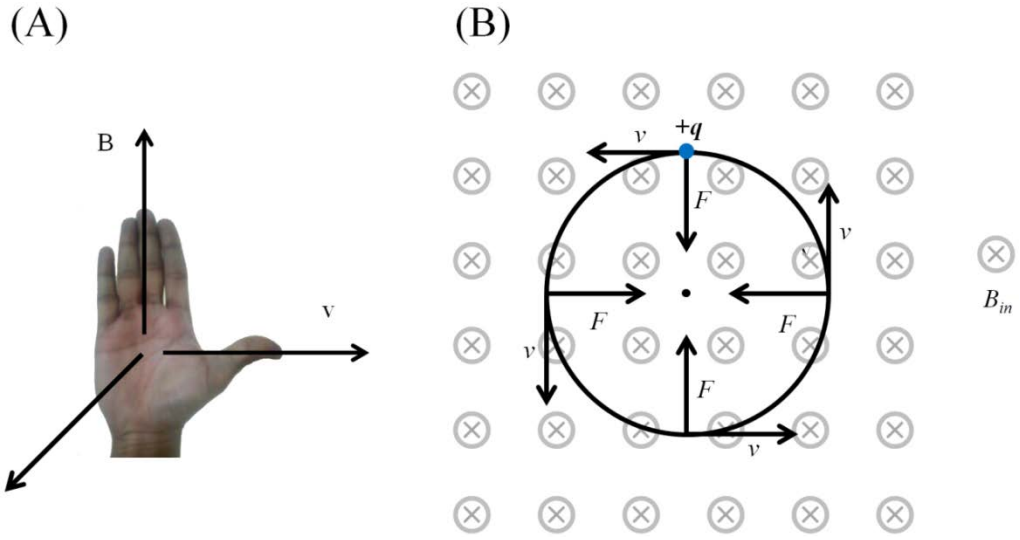


Figure 11.2 The right hand rule showing the direction of the Lorentz force in a magnetic field (A) and the circular path taken by the particle (B).

The Lorentz force therefore provides the particle a centripetal force. Therefore, equation 11.4 can be written as

$$\frac{mv^2}{r} = q\ (v \times B) \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (11.5)$$

where, *r* is the radius of the circular path

Rearranging in terms of *r*:

$$r = \frac{m\,v^2}{qvB\,sin\theta} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (11.6)$$

$$r = \frac{mv}{qB\,sin\theta} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (11.7)$$

where, θ is the angle between the velocity vector, *v* and the magnetic field vector, *B*.

In a mass spectrometer, *v* and *B* are generally orthogonal to each other; in that case:

$$r = \frac{mv}{qB} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (11.8)$$

Equation 11.8 shows that the deflection caused by a magnetic field in a moving charged particle is proportional to the mass to charge ratio. For the two particles having same charge but different masses, the one with lesser momentum deflects more ($r \propto mv$ and smaller r means larger deflection). In mass spectroscopy, the charge is usually represented as *z* and we shall be sticking to the same convention. A mass spectrum is a two dimensional plot between ion abundance and $\frac{m}{z}$ ratio (Figure 11.3)

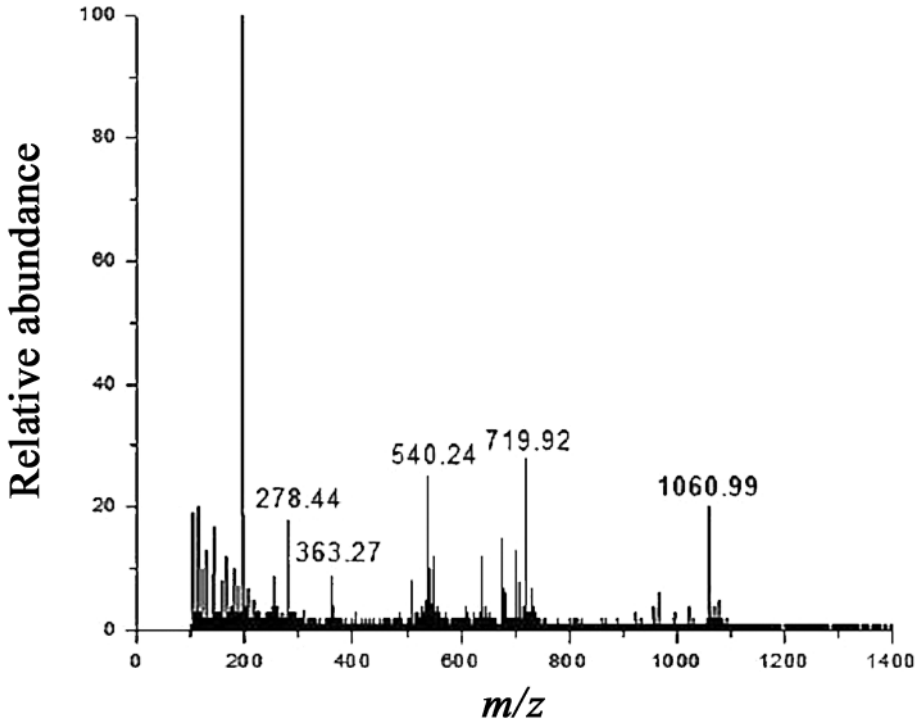


Figure 11.3 A typical mass spectrum

Let us see the design of a typical mass spectrometer (Figure 11.4). The basic requirement for an analyte molecule to be studied using mass spectrometry is that it has to be charged. A large number of molecules, however, may not be charged. The first step in an MS experiment is therefore to ionize the molecules. The spectrometer therefore has an *ionization source*. The ions generated are then separated by one or more *mass analyzers* which are then detected by a *detector*.

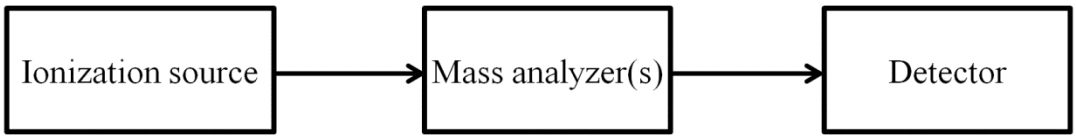


Figure 11.4 The components of a mass spectrometer

Ionization/ionization source

The first step in an MS experiment is to obtain the ions in gas phase. The mass spectrometers, therefore have an ionization chamber (also called ionization source) where the samples are introduced to achieve ionization. Ions are generated through one of the several methods that have their own merits and limitations. Some of the ionization methods are:

Ion mode

Mass spectrometric analyses are usually performed in the positive ion mode *i.e.* only cationic species are detected. It is, however, possible to study the molecules in negative ion mode as well, where anions are detected. Unless mentioned otherwise, it is usually assumed that the analysis is done in positive ion mode.

Electron Ionization (EI)

In electron ionization method (Figure 11.5), a heated filament is used to emit the electrons. The electrons are accelerated through the ionization chamber under the influence of a strong electric field. The sample in gas phase is introduced into the ionization chamber. A high energy electron can knock off an electron from an analyte molecule, M giving a molecular radical cation.

$$M + e^- \rightarrow M^{\bullet +} + 2e^- \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (11.9)$$

$M^{\bullet +}$ is referred to as the molecular ion. Loss of electron is a miniscule loss of mass; therefore mass of $M^{\bullet +}$ equals the mass of the molecule.

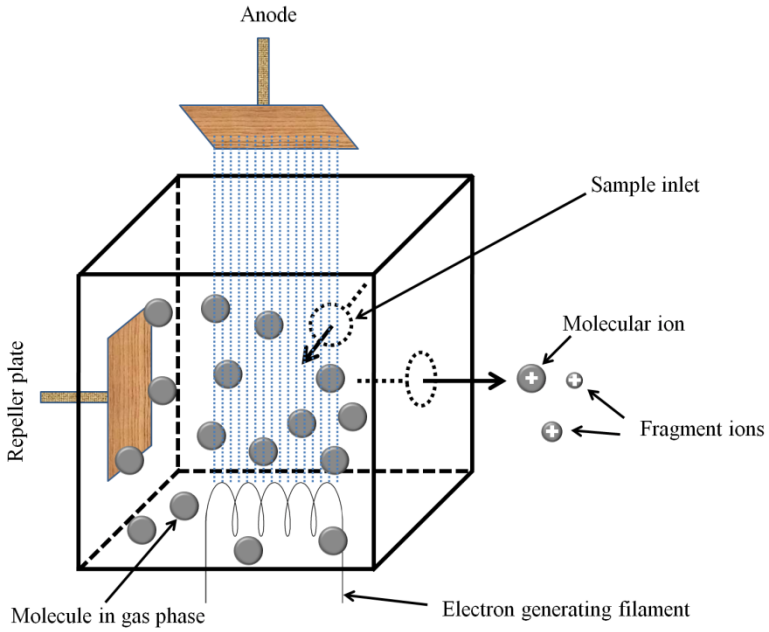


Figure 11.5 Design of an electron ionization source

The kinetic energy of the electrons is usually 70 eV in the electron ionization method. Typically 10-20 eV energy is transferred to the molecules. Around 10 eV energy is sufficient to cause ionization of most organic molecules; the radical cation is therefore left with an excess energy. Electron ionization, therefore often causes extensive fragmentation of the radical cation. Detection of these fragments can provide useful structural information about the molecule but can complicate the data for larger molecules. In some cases, molecular ion may not even be detected at all. The fragmentation is usually hemolytic, resulting in an even-electron cation and a neutral radical (Equation 11.10). Fragmentation into a neutral molecule and a smaller radical cation, however, is not uncommon (Equation 11.11).

$$M^{\bullet+} \rightarrow m_1^{\bullet} + m_2^{+} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (11.10)$$

$$M^{\bullet+} \rightarrow m_1^{\bullet+} + m_2 \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (11.11)$$

Electron ionization method is limited to samples in the gas phase. Gaseous and highly volatile samples can be directly introduced into the ionization chamber. Liquid and solid samples can be heated to obtain molecules in gaseous phase but it depends on the thermostability of the samples.

Chemical Ionization (CI)

The ions are produced through the collision of the sample molecules with the primary ions produced by a gas (called a reagent gas) in the ionization chamber. The reagent gas is ionized through electron ionization. The radical cations generated will undergo fragmentations and reactions. The most common reaction generating ions is a proton transfer from a gas cation (GH^+) to the molecule. Methane, isobutane, and ammonia are the most common reagent gases. Let us take methane as an example to understand how chemical ionization occurs:

Electron ionization: $CH_4 + e^- \rightarrow CH_4^{\bullet +} + 2e^-$

Fragmentation: $CH_4^{\bullet +} \rightarrow CH_3^+ + H^\bullet$

$CH_4^{\bullet +} \rightarrow CH_2^{\bullet +} + H_2$

The radical cations and the carbocations can react with the reagent gas to give various protonated species: Reactions: $CH_4^{\bullet +} + CH_4 \rightarrow CH_5^+ + CH_3^\bullet$

$CH_3^+ + CH_4 \rightarrow C_2H_5^+ + H_2$

The analyte molecules acquire the protons from any one of these cations:

$$M + CH_5^+ \rightarrow MH^+ + CH_4$$

The ion, MH^+ is called the quasimolecular or pseudomolecular ion as its mass is one unit more than the molecular mass. Saturated hydrocarbons usually ionize through hydride abstraction by the reagent gas cation:

$$RH + CH_5^+ \rightarrow R^+ + CH_4 + H_2$$

The chemical ionization method imparts little excess energy to the molecular ions thereby resulting in lesser fragmentation as compared to the electron ionization method. Furthermore, the degree of fragmentation depends on the reagent gas used for chemical ionization. The fragmentation caused by isobutane and ammonia is considerably less than that caused by methane. Like electron ionization, chemical ionization is also suitable only for gaseous samples limiting it to the gases and volatile liquids.

Fast Atom Bombardment (FAB)

Fast atom bombardment is a soft ionization technique *i.e.* it causes little fragmentation of the molecular ions generated. In fast atom bombardment ionization methods, the sample is dissolved in a non-volatile liquid and the ions are extracted by bombarding the sample with a beam of high energy atoms (~5 keV), usually argon (sometimes xenon). The commonly used liquid matrices include glycerol, thioglycerol, and *m*-nitrobenzyl alcohol. Fast moving Argon atoms are generated as shown in the Figure 11.6. The Argon radical cations ($Ar^{\bullet+}$), generated through electron ionization, are accelerated and focused as a sharp beam. The high energy $Ar^{\bullet+}$ ions are allowed to collide with the Ar atoms resulting in the neutralization of some of the $Ar^{\bullet+}$ ions in the beam. The residual $Ar^{\bullet+}$ in the beam are extracted out by applying an electric field, thereby resulting in a beam of fast moving atoms. The atoms collide with the sample dissolved in the liquid matrix extracting the ions into the gas phase.

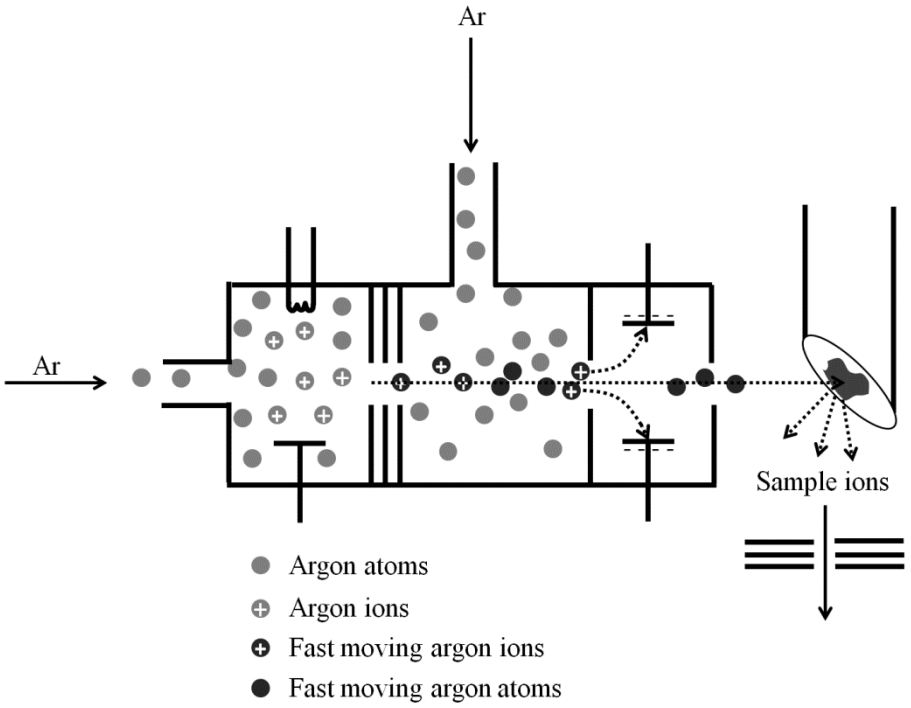


Figure 11.6 Diagram showing the generation of fast moving atoms and sample ionization in FAB ionization source

FAB causes little or no ionization but desorbs the ions already existing in the solution into the gas phase. FAB causes desorption of the ions present on the surface of the matrix; the compounds having higher surface activity are therefore detected better. FAB is particularly good for polar molecules with large molecular weights and molecules up to 10,000 Da can be detected. It is therefore possible to detect biomolecules like peptides, oligonucleotides, and oligosaccharides using FAB ionization.

Liquid secondary ion mass spectrometry (LSIMS)

LSIMS is similar to FAB, with only difference that an ion is used for bombarding the samples. Usually argon, xenon, or caesium ions are used for the LSIMS.

Laser desorption

Laser desorption or laser ablation is the ionization method wherein an intense laser beam is focused on a solid sample resulting in ablation of mass from the surface. The laser pulse causes both desorption and ionization of the molecules. The ions generated are short-lived and therefore detected simultaneously. Laser desorption, however, causes fragmentation for large molecules (Molecular mass >500 Da) therefore restricting its use to small molecules.

Owing to the difficulty in generating gas phase ions of biomolecules like proteins and nucleic acids, the application of mass spectrometry, till late 1980s, was largely restricted to the small organic molecules and biomolecules (amino acids, peptides, oligonucleotides, etc.). Therefore, mass spectrometry was not of much use for biochemists. Advent of two ionization methods around 1987-88 revolutionized the area of biomolecular mass spectrometry. Both these methods are routinely used for identifying and characterizing the biomolecules.

Matrix-assisted laser desorption ionization (MALDI)

MALDI is basically a laser desorption ionization method wherein ionization is assisted by small organic molecules, called matrix. The matrices used in positive ion mode MALDI mass spectrometry are organic acids that have strong absorption for the wavelength of the laser used. Some of the commonly used MALDI matrices are listed in Table 11.1.

Table 11.1 Some of the MALDI matrices commonly used for biomolecules	
Matrix	**Analyte**
α-Cyano-4-hydroxycinnamic acid (CHCA or HCCA)	Proteins, peptides, lipids, oligonucleotides
2,5-Dihydroxybenzoic acid (DHB)	Proteins, peptides, oligonucleotides, oligosaccharides
3,5-Dimethoxy-4-hydroxycinnamic acid (Sinapinic acid)	Proteins, peptides, lipids
3-Hydroxypicolinic acid (HPA)	Oligonucleotides
Trihydroxyacetophenone (THAP)	Oligonucleotides, oligosaccharides

The sample for MALDI is usually prepared in one of the following ways:

i. mixing the analyte solution with the matrix solution → deposition of the mixture on a metallic plate → complete drying of the sample
ii. Deposition of the matrix on the metallic plate → drying of matrix → addition of analyte solution → drying of analyte solution

Desorption and ionization is achieved by applying the laser pulse on the dried sample. Although the exact mechanism behind MALDI is not completely understood, it is believed that the absorption of light by the matrix molecules causes sublimation of matrix crystals carrying along with them the analyte molecules into the gas phase (Figure 11.7).

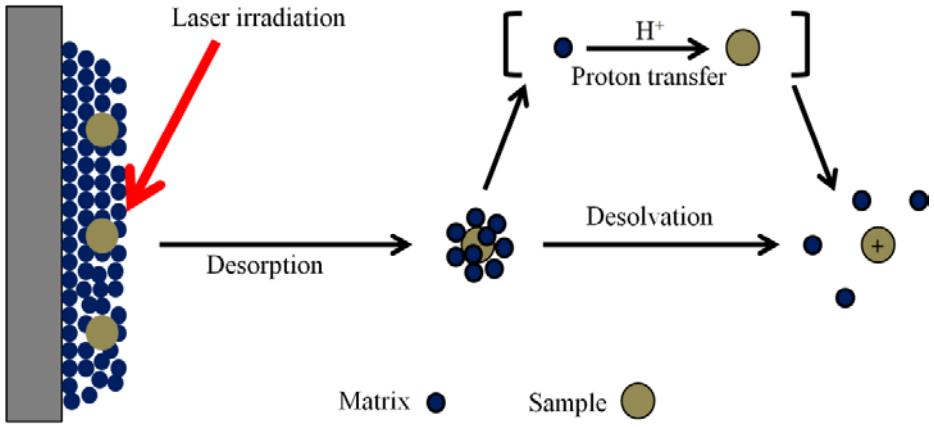


Figure 11.7 Diagram showing ionization in a MALDI source

MALDI can very efficiently generate the gas phase ions from a variety of non-volatile and thermolabile molecules such as proteins, carbohydrates, nucleic acids, and synthetic polymers. MALDI can desorb and ionize the molecules as large as 300 kDa. MALDI usually results in the molecular species having only one charge. In positive ion mode, a quasimolecular ion is formed by protonation of the molecule (MH^+). The samples or the matrices can have trace amounts of alkali metal ions, often resulting in the quasimolecular species, MNa^+ or sometimes MK^+ (Figure 11.8). Multiply charged species are also observed sometimes.

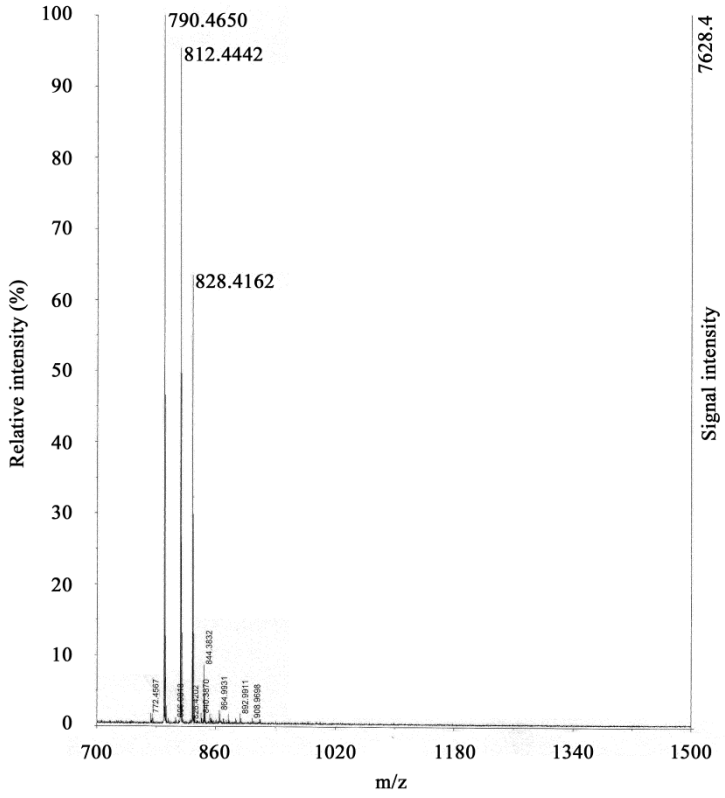


Figure 11.8 MALDI-TOF mass spectrum of a peptide with theoretical mass of 790.448 Da. The peaks at 812.4442 and 828.4162 represent [peptide–Na^+] and [peptide–K^+], respectively.

Time-of-flight (TOF) mass analyzers are well suited for the pulsed ionization techniques such as MALDI (In MALDI, laser pulses produce the ions in pulses). We shall be discussing the different mass analyzers in the next lecture.

Electrospray ionization (ESI)

In electrospray ionization, the analyte solution enters the ionization chamber, maintained at atmospheric pressure, through a fine capillary. The typical flow rates are ~1-20 µl/min. A potential difference of ~3-6 kV is applied between the capillary and the counter-electrode that is ~0.3 – 2 cm away. Under this electric field, the sample droplets appearing at the capillary end accumulate large amount of charge. If the potential of the capillary is above a threshold voltage, the drop will be dispersed into a very fine spray. A coaxial sheath is present around the capillary through which dry nitrogen is supplied for better nebulization and restricting the dispersion of the spray in space. Evaporation of the solvent causes the droplets to diminish in size and their charge density to increase (Figure 11.9). The high charge density on these droplets can further result in the production of smaller droplets. The droplets keep losing the solvent ultimately resulting in the desorption of molecular ions from the surface. As the ions are generated from the surface, a surface active molecule will be detected better. For large molecules such as proteins, the molecules do not desorb from the droplets but become ionized through complete evaporation of the solvent.

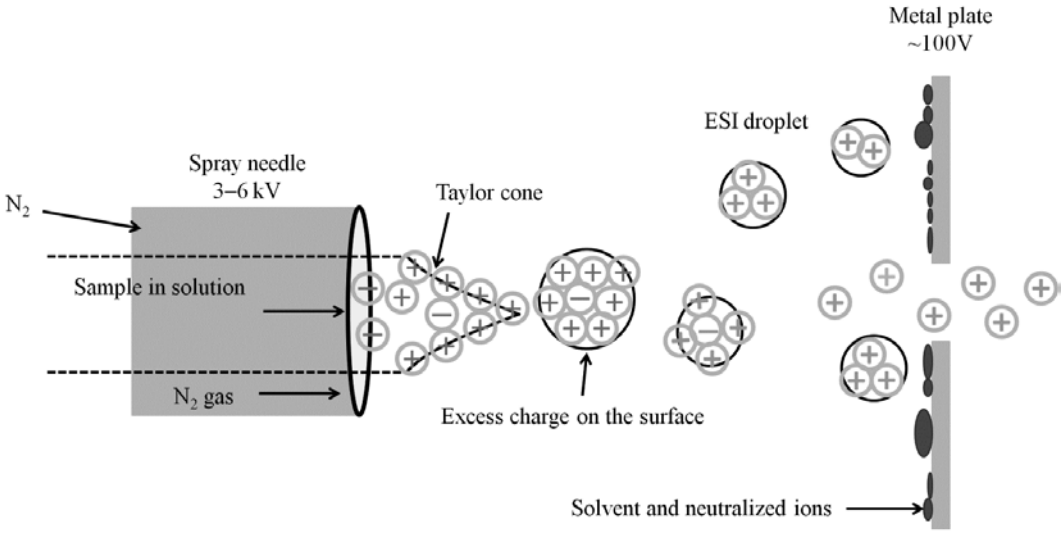


Figure 11.9 Diagrammatic representation of ionization in an ESI ion source

ESI typically results in multiply-charged molecular species. Biological macromolecules usually result in the mass spectra wherein the consecutive peaks differ by one charge unit. This information allows one to calculate the mass of the biomolecule.

Lecture 12 Mass Spectrometry-II

Following ionization, the gas phase ions are accelerated towards the mass analyzers. A great variety of mass analyzers are utilized in the mass spectrometers. All of these analyzers separate the molecules using static or dynamic electric fields and magnetic fields, alone or in combination. After mass analysis, the ions are detected and the mass spectra are generated.

Imagine what would happen if the ions collide with the air molecules in the MS tube. This can lead to the loss of charge to the molecules in the air, deviation in the ion trajectory which might lead to the collision with the MS tube, reactions with the air molecules, etc. A mass spectrometer, therefore operates under very high vacuum to ensure that the ions reach the detector without colliding with the air molecules. The mean free path of a particle is the average distance a particle travels before colliding with other particles and is inversely proportional to the pressure of the gas and the size of the colliding molecules. The mean free path, according to the kinetic theory of gases is given by:

$$\lambda = \frac{kT}{\sqrt{2}\,\pi d^2 p} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \quad (12.1)$$

where, λ is the mean free path, k is the Boltzmann constant (1.38×10^{-23} JK^{-1}), T is the temperature, d is the sum of the radii of the ion and the colliding molecule, and p is the pressure.

Let us try calculating the mean free path for a small ion in the air. We need to make a few assumptions: Air is largely nitrogen (~78%) and oxygen (~21%). The Van der Walls radii for nitrogen and oxygen are 155 pm and 152 pm, respectively. As the Van der Walls radii of oxygen and nitrogen are very close, we can assume air to be composed of a particle with ~150 pm. Let us calculate the mean free path for a small molecular ion, CH_4^+ (molecular radius ~380 pm). The mean free path at room temperature ($T \approx 298$ K) and atmospheric pressure (1.01×10^5 Pa) can be given by:

$$\lambda = \frac{(1.38 \times 10^{-23}) \times 298}{\sqrt{2}\,\pi \times (530 \times 10^{-12})^2 \times (1.01 \times 10^5)} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (12.2)$$

$$\lambda = \frac{4.11 \times 10^{-21}}{1.26 \times 10^{-13}} \qquad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (12.3)$$

$$\lambda = 3.26 \times 10^{-8} \text{ m} = 32.6 \text{ nm} \quad \cdots\cdots(12.4)$$

In a mass spectrometer, the ions have to travel large distances (usually >1 m). It is therefore absolutely essential to apply large vacuum for increasing the mean free path by several orders of magnitude. Let us now have look at some of the important mass analyzers:

Magnetic sector

Figure 12.1 shows a diagram of the magnetic sector analyzer mass spectrometer. The ions (say, cations) generated in the ionization chamber are accelerated under a strong electric field. The accelerated ions are allowed to pass through a narrow slit resulting in a sharply focused ion beam. The ions in the beam can be deflected by applying a magnetic field perpendicular to the velocity of the ions.

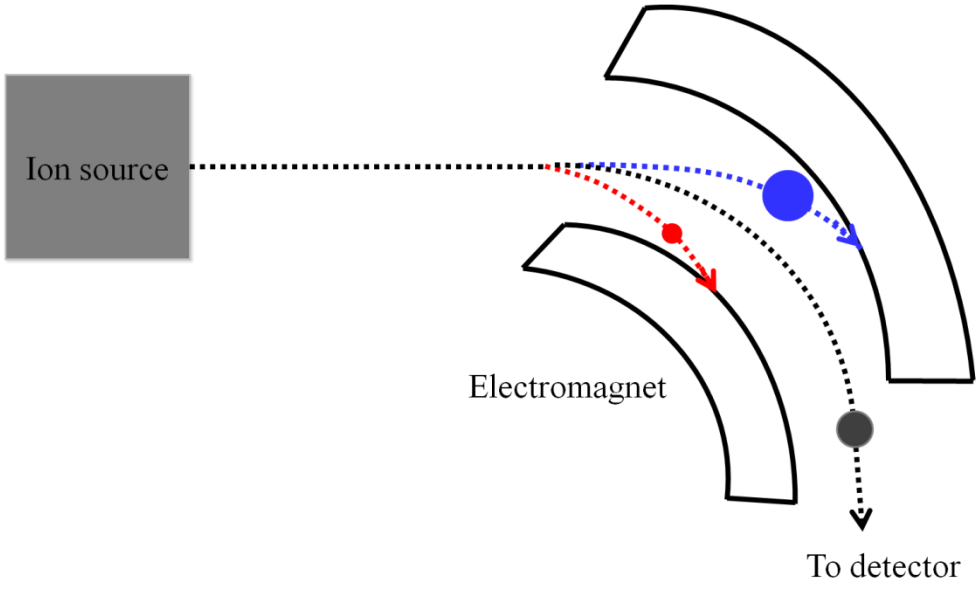


Figure 12.1 Diagram of a magnetic sector mass analyzer.

Too small (blue dotted line, Figure 12.1) or too much deflection (red dotted line, Figure 12.1), which is determined by the charge and the momentum of the ions (Equation 11.8), causes the ions to collide with the MS tube. Therefore, the ions within a small $\frac{m}{z}$ range will be allowed to go to the detector. It is possible to sequentially allow all the ionic species to reach the detector by gradually varying the magnetic field. Assume that the magnetic field strength is zero initially. All the ions will move straight (Equation 11.8) and collide with the curved MS tube losing their charge. If the magnetic field is slowly increased from zero to the maximum value, the

ions with lowest momentum and highest charge will appear first while the ions with highest momentum and lowest charge will appear last.

Time of flight (TOF)

In a time of flight mass analyzer, the time taken by the ions to reach the detector is measured. The ions are generated in bundles *e.g.* by MALDI. The ions are then accelerated towards the flight tube. The flight tube does not have any electric field and the ions drift in the flight tube according to their velocities (Figure 12.2).

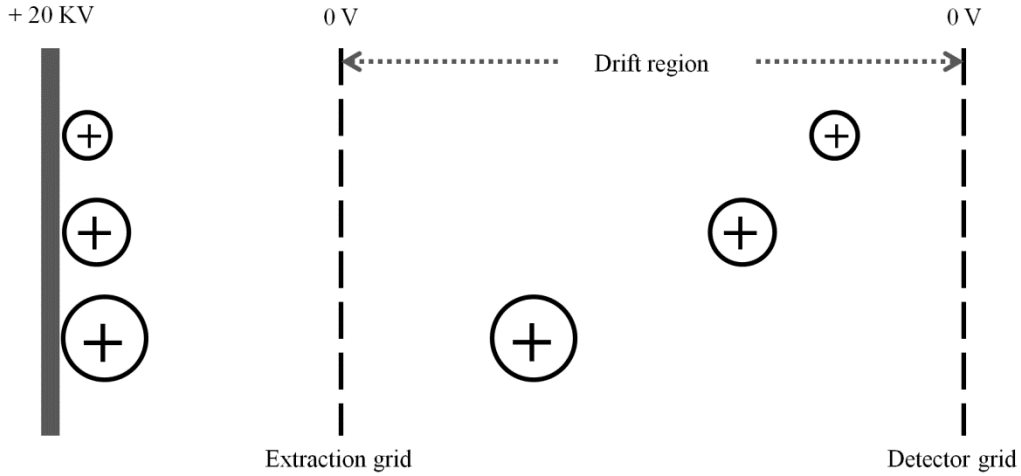


Figure 12.2 Separation of the ions in a TOF tube

If a particle with charge, q (= ez) and mass, m is accelerated towards the flight tube by electrostatic potential, V, the kinetic energy (KE) of the particle can be given by:

$$KE = qV \quad \cdots\cdots (12.5)$$

$$\frac{1}{2}mv^2 = zeV \quad \cdots\cdots (12.6)$$

where, v is the velocity of the particle.

If, L is the length of the tube, the time taken by an ion with velocity v is given by:

$$t = \frac{L}{v} \quad \cdots\cdots (12.7)$$

Substituting the value of v from equation 12.6 into equation 12.7 gives:

$$t = \sqrt{\frac{mL^2}{2zev}} \quad \cdots\cdots (12.8)$$

Equation 12.8 shows that the time taken by the particle to reach the detector is directly proportional to $\frac{m}{z}$ ratio. The $\frac{m}{z}$ ratio of the particle can therefore be calculated from the time of flight of the particle:

Rearranging equation 12.8:

$$\frac{m}{z} = \frac{2evt^2}{L^2} \quad \text{.....} \quad (12.9)$$

A serious problem with the linear TOF analyzers is their poor resolution. This happens due to difference in the flight times among the ions having same $\frac{m}{z}$ ratio. The factors responsible for the poor resolution include length of the ionization laser pulse, space distribution of the ions formed, and spread in the initial kinetic energies of the ions. In MALDI-TOF, for example, these factors severely affect the resolution. Two techniques have considerably improved the resolution in MALDI-TOF:

i. Delayed extraction: In continuous extraction, the ions generated are continuously extracted towards the TOF tube. An ion with higher initial kinetic energy reaches the detector earlier than the ion with smaller initial kinetic energy. Delayed extraction improves this situation substantially. Following ionization, the ions are allowed to move in the field free region according to their kinetic energies during a short delay. An extraction pulse is then applied; the pulse gives more energy to the ions that are nearer to the ionization source as compared to those that have moved away. The ions with small initial kinetic energies, therefore gain more energy and catch up the ions moving ahead (Figure 12.3).

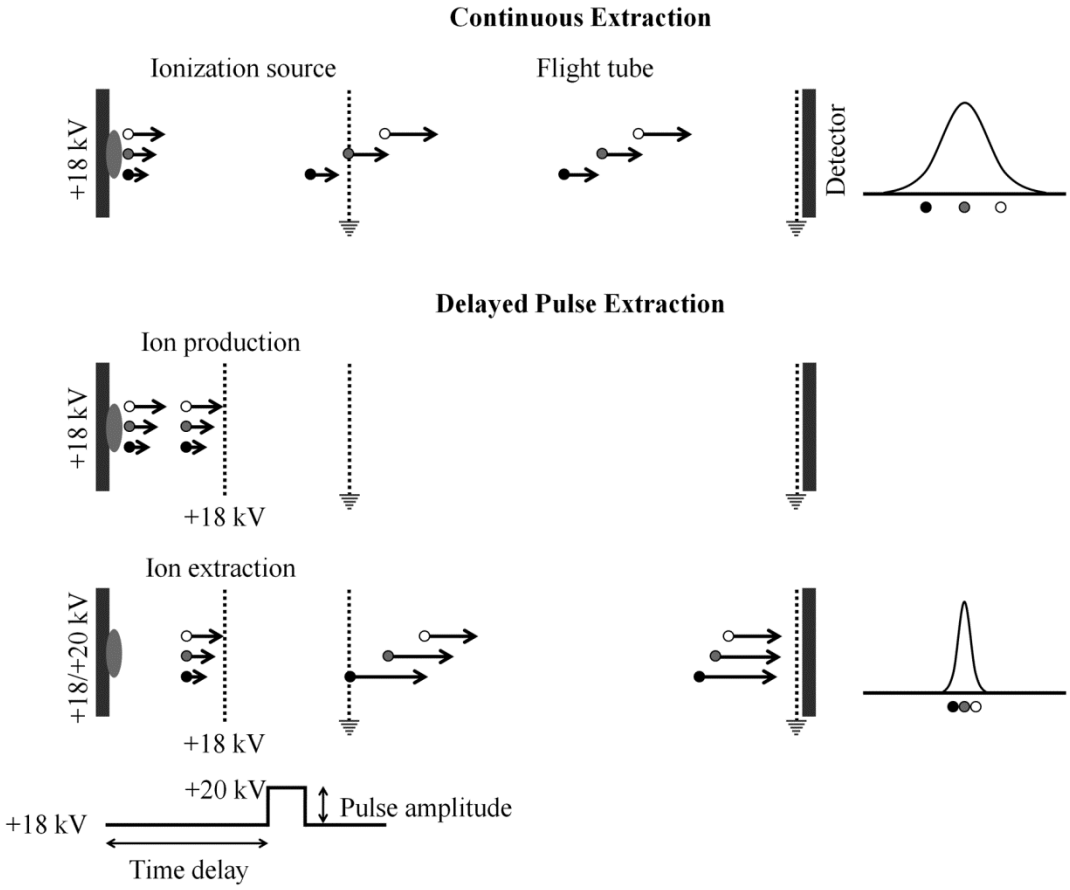


Figure 12.3 A diagrammatic representation of delayed extraction TOF

ii. Reflectrons: A simple reflectron is composed of a series of equally spaced grids. The reflectron is placed at the tube end opposite to the ion source. A potential is applied to the reflectron so as to reflect the incoming ions. A reflectron therefore acts as an ion mirror. The ions with higher kinetic energy will travel longer distance before reflecting back than those have smaller kinetic energy. The ions with higher kinetic energy are therefore made to travel longer distances thereby correcting for the spread in the peaks (Figure 12.4).

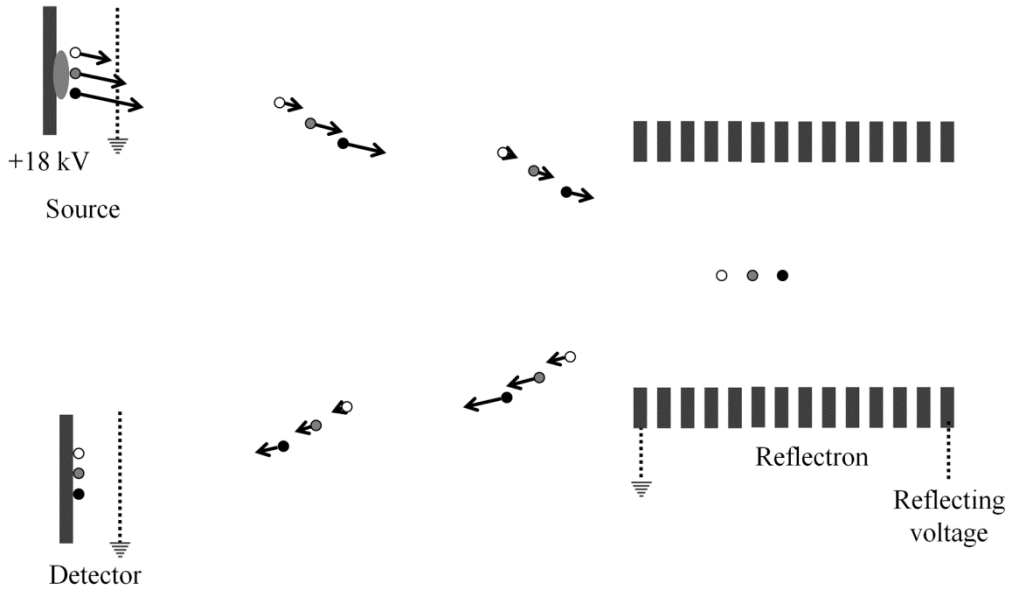


Figure 12.4 A diagrammatic representation of reflectron-mode TOF

Quadrupole analyzers

A quadrupole mass analyzer is made up of four rods arranged parallel to each other as shown in Figure 12.5. The principle of a quadrupole was proposed by Paul and Steinwegen in 1953 wherein hyperbolic cross-section of the rods was described as necessary. In practice, however, rods with circular cross section have also proved effective and have replaced the rods with hyperbolic cross-section in modern quadrupole detectors.

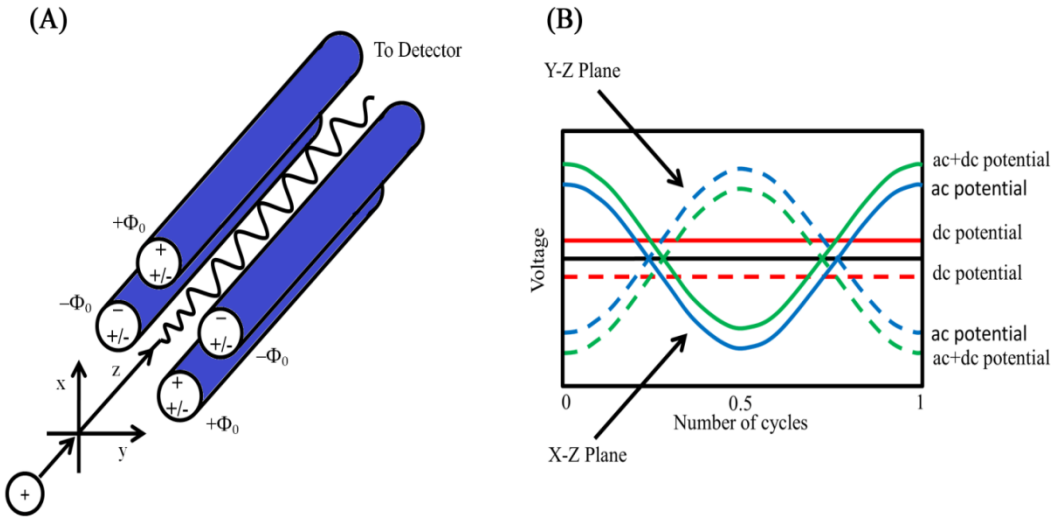


Figure 12.5 A quadrupole mass analyzer (A) and the potential on the rods as a function of time (B).

Consider the rods in x-z plane to be at a positive potential, U and the rods in y-z plane to be at a negative potential –U. An *a.c.* potential is subsequently applied to the rods such that the *a.c.* potential on the rods in x-z plane is 180° out of phase than that applied on the rods in the y-z plane. The potential on the rods in x-z and y-z plane can therefore be represented as

$$\Phi_{x-z} = U + V cos\omega t$$

$$\Phi_{y-z} = -U - V cos\omega t$$

The values of U range from 500 – 2000 volts while V ranges from 0 – 3000 V. The gas phase ions are accelerated and introduced into the quadrupole as a focused beam. To understand how the ions are separated inside a quadrupole, let us consider the rods in the x-z plane and those in the y-z planes separately. The rods in x-z plane have a positive *d.c.* potential and an *a.c.* potential that will periodically make the overall potential negative (Figure 12.5B). The ions will respond to the changes in the potential. If a cation is very heavy or the frequency of the *a.c.* potential is very high,

the cation remains largely unaffected and experiences only the average potential. This causes the ion to get focused towards the centre. If a cation is very light, it will readily respond to the changes in the potential and can accelerate towards the rods during negative potential and collide with them. Collision of the cation with the quadrupole rods during negative potential depends on the magnitude of the potential on the rods, frequency of the *a.c.* potential, mass of the cation, charge on the cation, and the position of the cation in the quadrupole. Let us now turn our attention towards the rods in the y-z plane. These rods have a negative average potential; the heavier cations will therefore get accelerated towards the rods and collide with them. Lighter cations will respond to the *a.c.* potential and get focused towards the centre. We can say that the x-z rods filter out the lower masses while y-z rods filter out the higher masses. The four rods together can therefore be used to allow the passage of a very small range of masses. A quadrupole therefore acts as a mass filter.

Ion trap

As the name suggests, ion trap mass analyzers trap the ions inside them. An ion trap can either be a 2D or a 3D ion trap. A 3D ion trap is basically a quadrupole in 3 dimensions. It has a circular electrode (also called a ring electrode) with two ellipsoid electrodes as its caps (Figure 12.6).

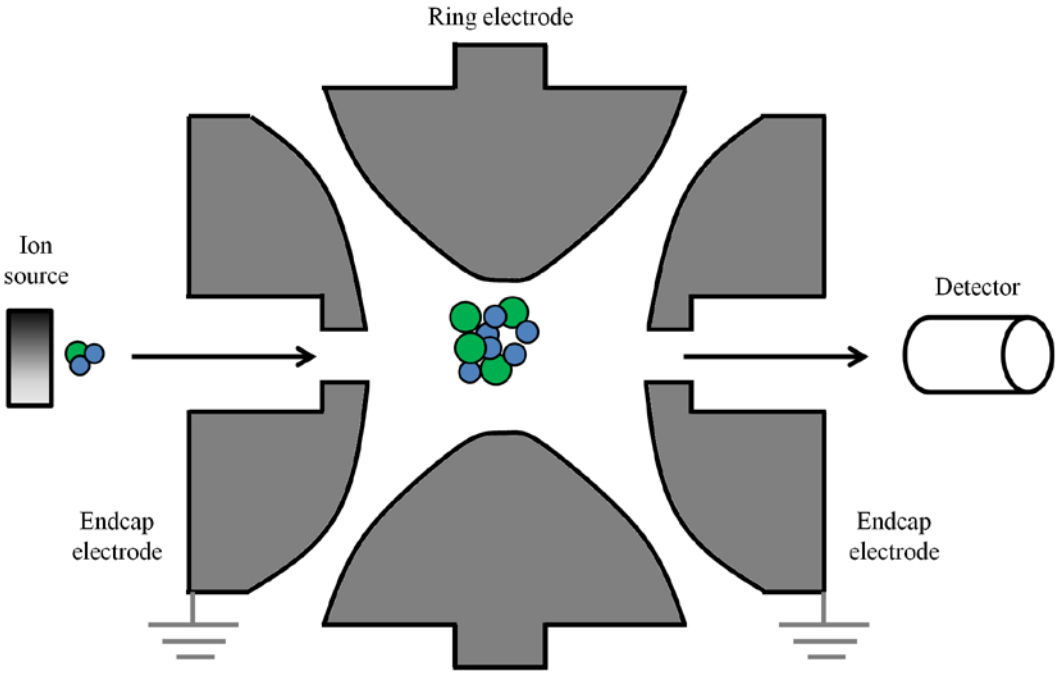


Figure 12.6 A diagrammatic representation of an ion trap

All the three electrodes have hyperbolic surfaces. Unlike linear quadrupole (quadrupole with four parallel rods) wherein electrostatic forces act in two axes, forces act in all the three axes in an ion trap. This implies that the stable trajectories of the ions cause them to be trapped. Ion traps provide higher sensitivity and are useful in tandem mass spectrometry; ions of desired mass can selectively be allowed to escape the trap by varying the ac potential; the escaped ions can then be analyzed by another mass analyzer attached in tandem.

Orbitrap

Orbitrap is an electrostatic ion trap. It has a barrel containing a spindle-shaped electrode at the centre. The spindle electrode is held at a constant negative voltage (-3200 V) for positive ion mode MS. Ions enter the orbitrap tangentially and get trapped by revolving around the spindle shaped electrode (Figure 12.7). The ions can be ejected out by applying the radiofrequencies of suitable frequency to the central electrode.

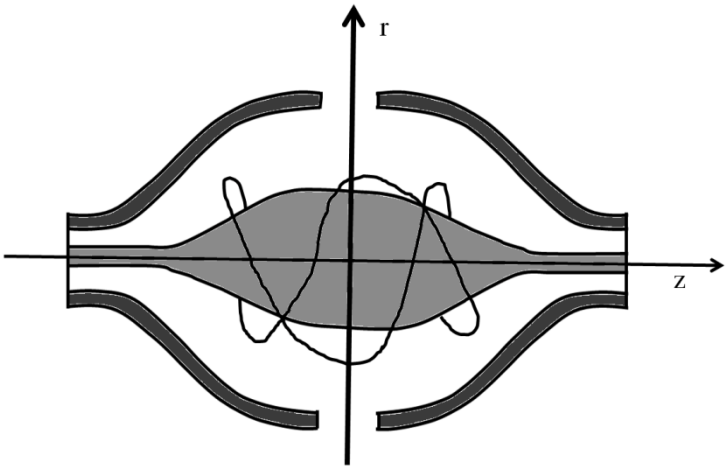


Figure 12.7 A schematic diagram of an orbitrap

Mass spectrometry coupled with chromatography

Mass spectrometers have been successfully coupled with the liquid and gas chromatographic methods. Chromatographic methods separate the compounds based on the differences in their physico-chemical properties. A complex mixture of compounds can be resolved into pure components using one or more chromatographic methods. Though excellent in separating the molecules, chromatographic methods do not allow identification of the unknown compounds in the mixture. The separated

components can be collected and identified using MS. It is also possible to couple the mass spectrometers with the chromatographic methods. Gas chromatography (GC) and liquid chromatography (LC) coupled with mass spectrometers have emerged as very powerful analytical tools. GC allows easy interfacing with the mass spectrometers; a gas chromatographic column can directly be coupled to the ionization source of the MS. Interfacing the LC with MS, however, is not as straightforward. We shall not be discussing the different types of interfaces of LC and MS. As it allows studying the large, polar, non-volatile, and thermolabile compounds, LC-MS is more widely used compared with GC-MS. Electrospray ionization is one of the softest ionization methods and probably the most widely used ionization method for LC-MS. LC-MS and GC-MS therefore allow determining the molecular masses of the eluants in real-time. The ions separated by MS can be fragmented in a collision cell; identification of these fragments by another mass analyzer allows identification of the compounds (tandem mass spectrometry). We shall see in the next lecture how fragmented ions help in the identification of the compounds.

Tandem mass spectrometry

Tandem mass spectrometry (also known as MS/MS) involves more than one mass analyzer. As we have just seen, incorporating more than one mass analyzer greatly enhances the capabilities of a mass spectrometer. Furthermore, it improves the sensitivity, mass resolution, and the mass accuracy of the spectrometer. The most common MS/MS experiment involves selecting an ion using first mass analyzer, which is then fragmented into daughter ions in a collision cell. The daughter ions are then detected by a second analyzer. It is possible to further fragment the daughter ions into granddaughter ions that are then analyzed by a third mass analyzer. It is in principle possible to do an MS^n experiment. We shall see in the next lecture how powerful tandem mass spectrometry is in identifying and characterizing the biomolecules.

Detectors

A variety of ion detectors presently exist, some of which are:

Electron multiplier: An electron multiplier is perhaps the most commonly used ion detector in mass spectrometers. It consists of a series of electrodes (dynodes). When an ion strikes the first dynode, it causes release of electrons from the dynode (the first dynode, therefore is a conversion dynode that converts the ion signal into electrons) that strike the second dynode releasing more electrons and so on. This cascading effect causes a large amplification in the electrical current (Figure 12.8A). Another design of the electron multiplier uses a continuous dynode (12.8B).

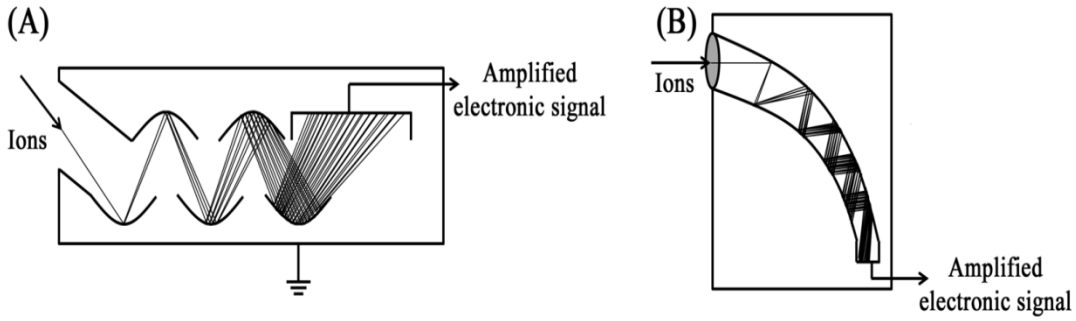


Figure 12.8 A diagrammatic representation of current amplification in an electron multiplier

Faraday cup: Faraday cup consists of a metallic cup that is connected to the earth through a resistor. An incoming ion strikes the cup and gets neutralized. This results in an electric current through the resistor that is proportional to the ion abundance.

Daly detector: A Daly detector is a type of electro-optical ion detector. The detector has one or more conversion dynodes that generate electrons in response to the ion strike. These electrons are accelerated towards a phosphor screen (scintillator) that generates photons in response to the electron strike. The photons, thus generated are detected by a photomultiplier tube (Figure 12.9).

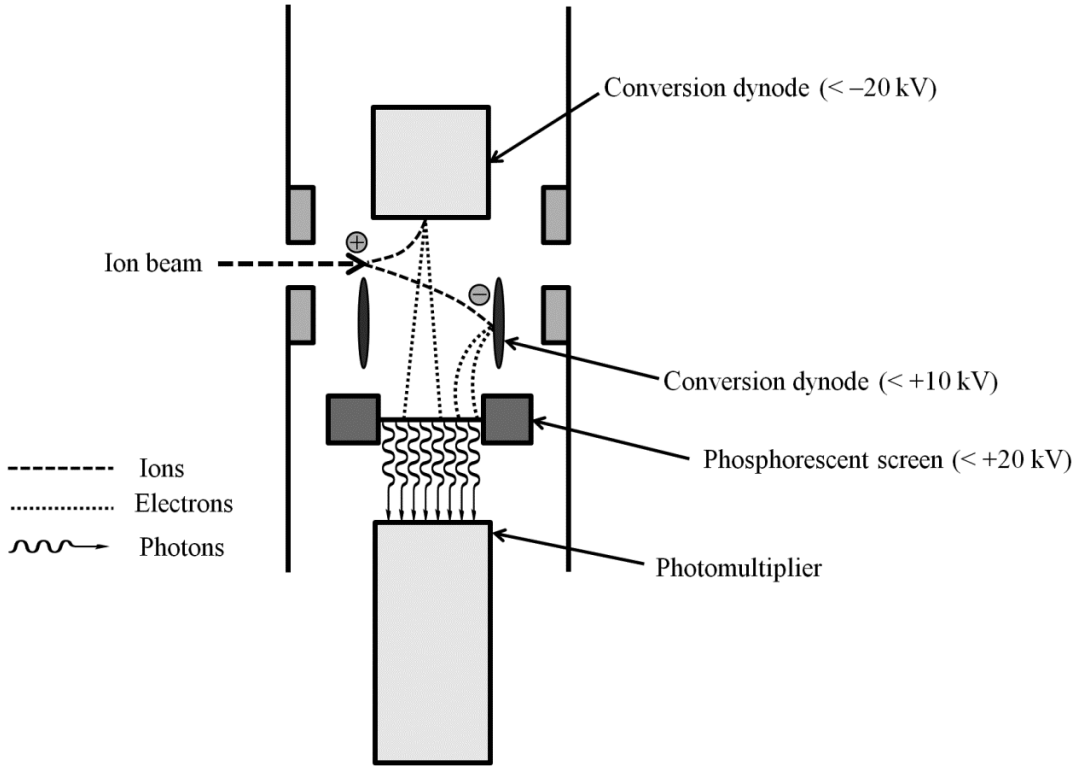


Figure 12.9 Schematic diagram showing working principle of a Daly detector

Focal-plane detectors: All the above-mentioned detectors fall in the category of point ion detectors *i.e.* the ions with different $\frac{m}{z}$ are resolved in time and detected at a single point. Focal-plane detectors detect the ions simultaneously; the ions of different $\frac{m}{z}$ are resolved in space, therefore strike at different points in array detectors (Figure 12.10).

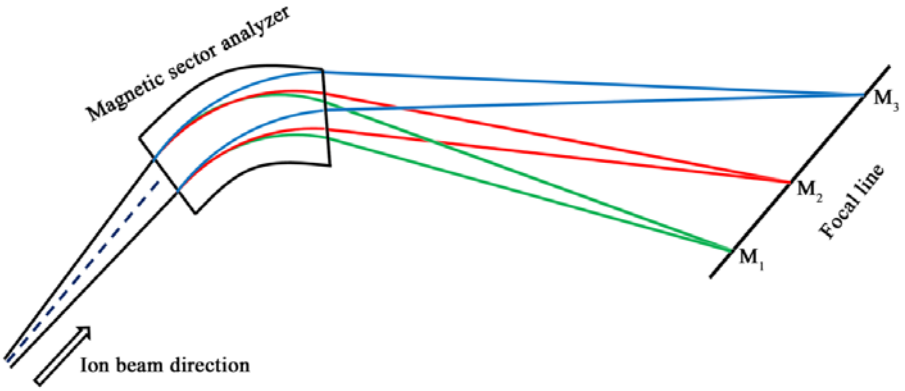


Figure 12.10 Schematic diagram showing a focal-plane detector for a magnetic sector mass analyzer

Lecture 13 Mass Spectrometry-III

We have already studied the various methods of ionization and mass analysis in lectures 10 and 11. This lecture discusses the properties of the mass spectra, their interpretation, and their applications, particularly in biomolecular analysis.

Characteristics of mass spectra

There are three basic characteristics of mass spectrometry: *exact mass*, *isotopic abundances*, and *fragmentation*.

Exact mass: When we talk of molecular weights in general chemistry, we typically refer either to the nominal mass or the average molecular mass. Nominal mass is calculated by adding the atomic masses of the predominant isotopes of all the elements rounded off to the nearest integer. Average molecular mass is the weighted average of the masses of all the isotopes (without rounding off). Stable isotopes of hydrogen, carbon, nitrogen, and oxygen as well as their relative abundances are listed in table 13.1.

Table 13.1 Natural abundance of the stable isotopes of selected elements

Isotope	Relative abundance	Exact mass	Nominal mass
^{1}H	99.985	1.0079	1
^{2}H (D)	0.015	2.0140	2
^{12}C	98.90	12.000	12
^{13}C	1.10	13.003	13
^{14}N	99.63	14.003	14
^{15}N	0.37	15.000	15
^{16}O	99.76	15.995	16
^{17}O	0.04	16.999	17
^{18}O	0.20	17.999	18

The nominal and average molecular masses of, say methane are 16 Da and 16.0428 Da, respectively. A mass spectrometer, however, detects the exact masses of the ions.

Isotopic abundances: Isotopic abundances are reflected in the high resolution mass spectra of the compounds and allow easy identification of small organic compounds. Let us take an example of methane. The different possible isotopologues of methane are listed in table 13.2 along with their natural abundances.

Table 13.2 Isotopologues of methane

Isotopologue	Relative abundances
$^{12}CH_4$	= 0.989 × 0.99985 × 0.99985 × 0.99985 × 0.99985 = 0.9884 = **98.84 %**
$^{12}CH_3D$	= 0.989 × 0.99985 × 0.99985 × 0.99985 × 0.00015 = 1.483×10^{-4} = **1.483×10^{-2} %**
$^{12}CH_2D_2$	= 0.989 × 0.99985 × 0.99985 × 0.00015 × 0.00015 = 2.247×10^{-8} = **2.247×10^{-6} %**
$^{12}CH_1D_3$	= 0.989 × 0.99985 × 0.00015 × 0.00015 × 0.00015 = 3.337×10^{-12} = **3.337×10^{-10} %**
$^{12}CD_4$	= 0.989 × 0.00015 × 0.00015 × 0.00015 × 0.00015 = 5.007×10^{-16} = **5.007×10^{-14} %**
$^{13}CH_4$	= 0.0110 × 0.99985 × 0.99985 × 0.99985 × 0.99985 = 0.011 = **1.1 %**
$^{13}CH_3D$	= 0.0110 × 0.99985 × 0.99985 × 0.99985 × 0.00015 = 1.649×10^{-6} = **1.649×10^{-4} %**
$^{13}CH_2D_2$	= 0.0110 × 0.99985 × 0.99985 × 0.00015 × 0.00015 = 2.474×10^{-10} = **2.474×10^{-8} %**
$^{13}CH_1D_3$	= 0.0110 × 0.99985 × 0.00015 × 0.00015 × 0.00015 = 3.712×10^{-14} = **3.712×10^{-12} %**
$^{13}CD_4$	= 0.0110 × 0.00015 × 0.00015 × 0.00015 × 0.00015 = 5.569×10^{-18} = **5.569×10^{-16} %**

$^{12}CH_4$ and $^{13}CH_4$ are the two predominant isotopologues of methane. Other isotopologues are too small in quantities to detect. A methane mass spectrum will therefore look like as shown in Figure 13.1

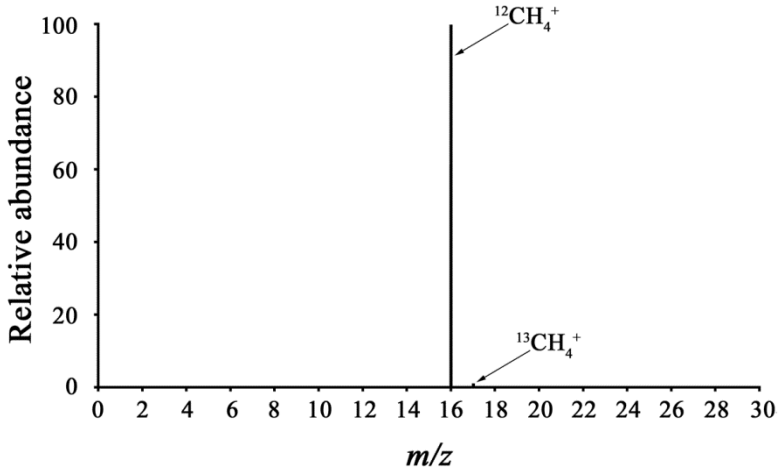
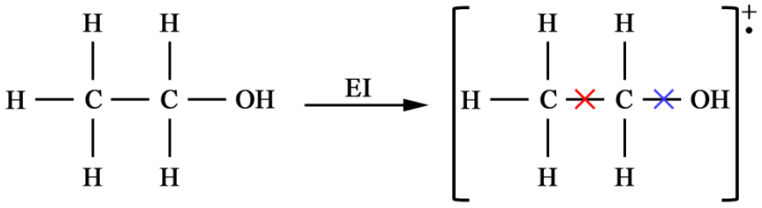


Figure 13.1 Electron ionization mass spectrum of methane showing the two predominant isotopologues

Fragmentation: We have already studied that electron ionization imparts large amount of energy to the cations that it generates. The radical cations thus generated undergo extensive fragmentation giving smaller cations. The fragments generated from molecular ions can provide important structural information about the molecules. Let us take ethanol as an example to see how this works:



The cross signs in the molecular radical cation represent the cleavage sites of the molecular ion. Cleavage between methyl and methylene carbons (red cross) can result in $[CH_3]^+$ or $[CH_2OH]^+$ ions while cleavage between methylene carbon and oxygen can result in $[C_2H_5]^+$ or $[OH]^+$ ions. The electron ionization mass spectrum for ethanol will therefore look like as shown in Figure 13.2. The idea behind identifying the structure of the molecules is to look at the differences between the peaks. A difference of 15 Da will be due to methyl loss, a difference of 17 is suggestive of the hydroxyl group, a difference of 29 can be due to the loss of an ethyl or aldehyde group. This information can be used to identify the molecules. In fact, there are softwares

available which can provide the possible molecular formulae of the compound when fed with the MS peaks.

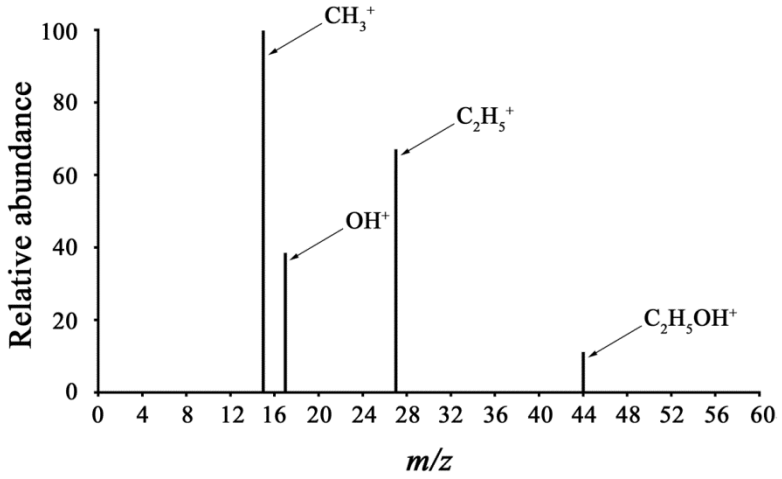


Figure 13.2 Electron ionization mass spectrum of ethanol

Analysis of biomolecules

Since the advent of MALDI and ESI ionization method, mass spectrometry has become a routine method for analyzing biomolecules. MS has been successfully utilized for obtaining a large amount of information about biomolecules including information that is difficult to obtain using other tools. Let us go through the various applications of MS in biomolecular analysis:

Molecular weight determination: Determination of molecular weight of a biomolecule is the most straightforward application of MS.

Structure verification and purity: Chemically synthesized molecules, such as peptides and oligonucleotides are often characterized by liquid chromatography and mass spectrometry. Suppose a chemically synthesized peptide, *DAKLRYFNQP* gives a MALDI mass spectrum as shown in Figure 13.3; the monoisotopic mass of the peptide is 1250.63 Da. The peak at 1180.59 is due to deletion of alanine.

THINK TANK??

In figure 13.3, where do you think the peaks at 1202.58 and 1273.62 m/z values arise from?

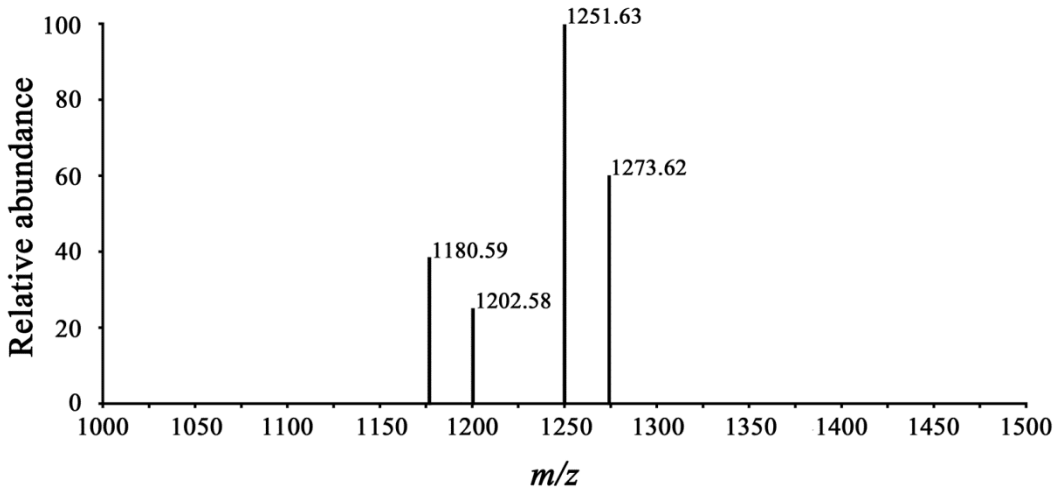


Figure 13.3 A hypothesized MALDI mass spectrum of the synthesized peptide, DAKLRYFNQP (theoretically calculated mass: 1250.63 Da)

Identification of chemical modifications: Biomolecules, especially proteins, can undergo a variety of chemical modifications such as phosphorylation, acetylation, methylation, fatty acylation, glycosylation, etc. These modifications are involved in biological processes like regulation of enzyme activity, signal transduction, gene expression, etc. It is therefore important to identify these species for understanding their function. Owing to its sensitivity and resolution, mass spectrometry has emerged as the method of choice for identification of small molecule modifications in biomolecules.

Protein sequencing:

Proteins are usually ionized using soft ionization techniques, MALDI and ESI. These methods yield quasimolecular ions that allow identification of proteins in complex mixtures. For determining their sequences, however, proteins need to be fragmented. The idea behind protein sequencing using MS is very straight forward and is summarized in Figure 13.4. Briefly, a protein quasimolecular ion is selected using a mass analyzer. The selected ion is then fragmented, typically in a collision cell (collision induced dissociation). Collision induced dissociation results in a large number of fragments that also have overlapping amino acid sequences. These daughter ions are then detected by a second mass analyzer. Mass of a fragment comprises the information for its amino acid composition. Masses of the overlapping fragments allow sequencing of the complete molecule.

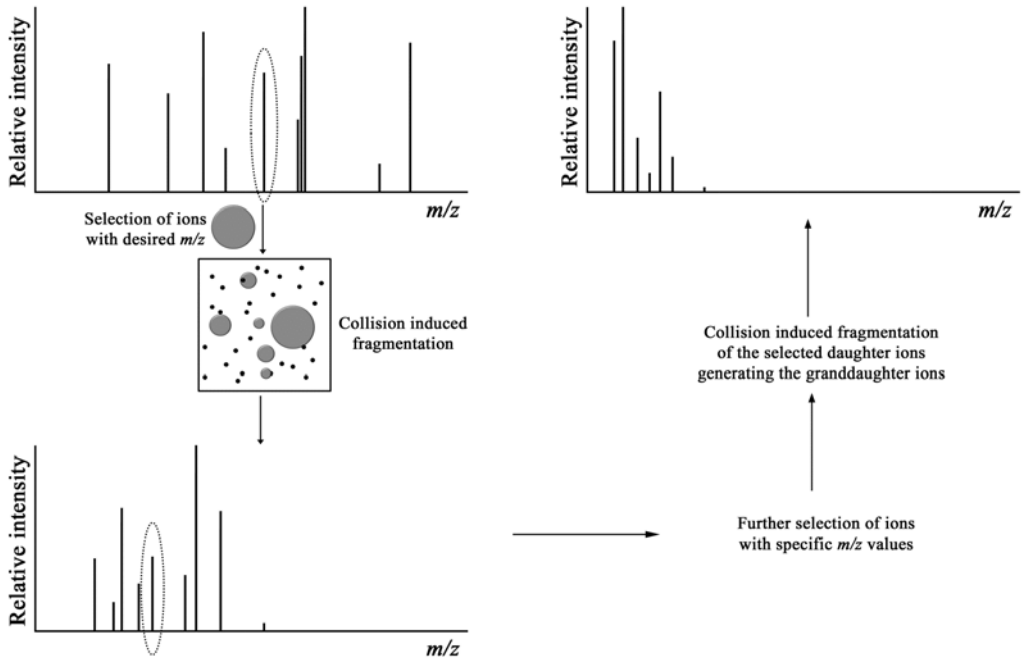
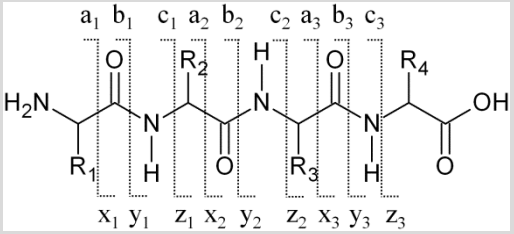


Figure 13.4 Principle of protein sequencing using tandem MS

Box 13.1: Peptide fragmentation

Fragmentation of peptides in a collision cell occurs in a well-defined manner; the fragmentation largely occurs along the peptide backbone with some side-chain fragmentation. The cleavage along the peptide backbone can occur at NH–C_α, C_α-CO, and CO–NH bonds. Each cleavage produces two species, one of which is charged and the other one neutral. As the charge can lie on any of the species produced, six different types of ions are generated. The nomenclature of these ions is shown in the figure below:



CO-NH is the most common cleave site. The difference in the masses of the adjacent b ions or y ions enables identification of the terminal amino acids thereby providing sequence information. Side-chain fragmentation is also useful as it provides information about side-chain modifications, if any.

Protein identification: Identification of a protein classically requires complete or partial protein sequence. A partial sequence can allow protein identification by comparing it with the sequences of the proteins available in protein sequence databases. A protein can therefore be identified by doing sequencing using MS. However, it may not be required to determine the sequence of a protein for its identification. A typical scheme for protein identification is shown in Figure 13.5.

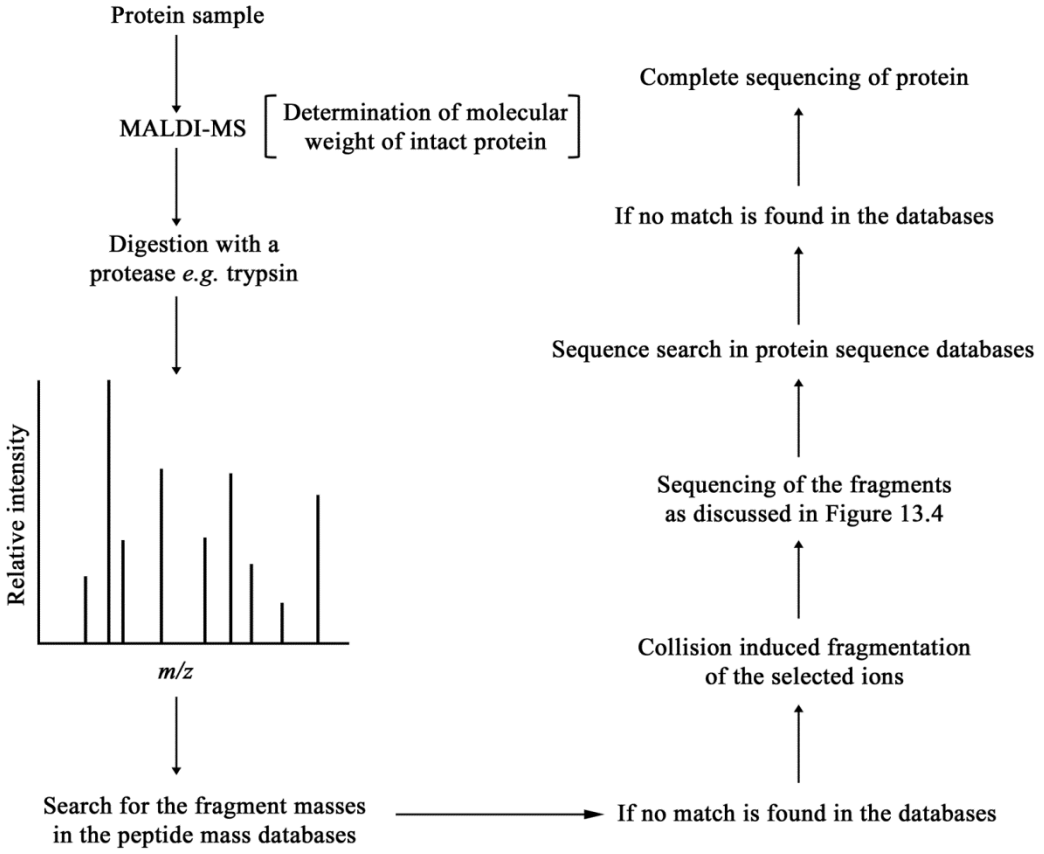


Figure 13.5 Protein identification using mass spectrometry

A protein is cleaved into smaller peptide fragments using a sequence specific enzyme, usually trypsin or chymotrypsin. Trypsin cleaves at the C-terminal side of lysine and arginine residues while chymotrypsin cleaves at the C-terminal side of aromatic amino acids, phenylalanine, tyrosine, and tryptophan. The masses of these fragments are then searched in peptide mass databases. This method is termed as the *peptide mass fingerprinting*. If no match is found in the databases, the peptides are sequenced using another MS as discussed earlier; the protein is identified by searching the

sequences of the fragments in the protein sequences databases. Protein identification using MS is central to the proteomic studies. A typical proteomic analysis is briefly summarized in Figure 13.6

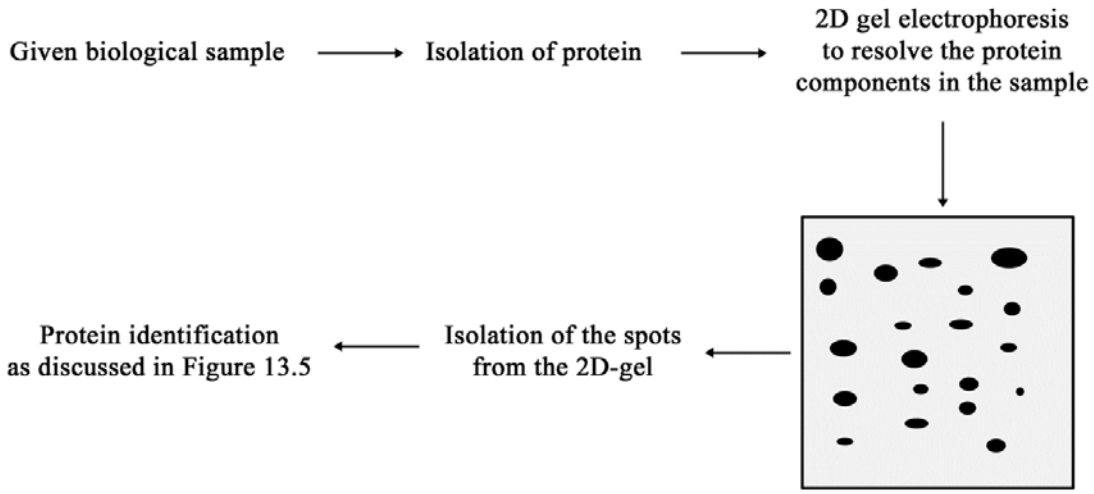


Figure 13.6 Outline of a proteomics experiment

Non-covalent protein complexes: Non-covalent interactions between the molecules are usually disrupted during ionization. ESI, however, has proved to be sufficiently soft for allowing detection of the protein complexes. It is, however, important to note that the complexes of biomolecules detected in gas phase may not represent the true solution complexes. The structures of the molecules in solution and the interactions they are involved in are determined by the complex interplay of hydrogen bonding, hydrophobic interactions, van der Waals forces, and electrostatic interactions. In gas phase under high vacuum, the electrostatic interactions are usually the most predominant ones. One therefore has to be careful while analyzing the data for the non-covalent complexes.

Three dimensional structures of proteins and peptides: Mass spectrometry can provide information about the three dimensional structure of a protein. Consider a protein that is being analyzed using ESI-MS. An unfolded protein usually shows a broader charge distribution with higher amount of charge due to larger solvent-accessible area and the exposed ionization sites. Another approach includes structural analysis of the proteins after subjecting them to the conformation sensitive reactions. Hydrogen-deuterium exchange (H/D exchange) is the most widely used such reaction.

The exposed regions on a protein will readily exchange their amide hydrogens with deuterium. The amides that are involved in backbone H-bonding (*i.e.* those involved in secondary structures) exchange their hydrogens very slowly. The folded regions in the protein can then be identified by analyzing the peptide fragments using tandem mass spectrometry. These approaches also allow study of the folding processes of proteins.

CHAPTER 2

Phase Rule

The 'phase rule' generalization was given by J.W. Gibbs in 1874 and further studied by H.W.B. Roozeboom 1884. The phase rule is able to predict the conditions necessary to be specified for a heterogeneous system to exhibit equilibrium. During the study of chemical systems, we usually deal with the systems containing two or more phases in equilibrium, which are called heterogeneous or polyphase systems. Phase rule was based on the basis of the principles of thermodynamics. The phase rule is able to predict qualitatively, by means of diagram, the effect of changing temperature, pressure, or concentration on a heterogeneous system in equilibrium. In this chapter we will study the phase rule and its various applications in the daily life.

1 GIBB'S PHASE RULE

Phase rule may be defined as:

When a heterogeneous system in equilibrium at a definite temperature and pressure, the number of degrees of freedom is equal to by 2 the difference in the number of components and the number of phases provided the equilibrium is not influenced by external factors such as gravity, electrical or magnetic forces, surface tension etc.

It is applicable for all the universally present heterogeneous systems.

Mathematically, the rule is written as

$$\boxed{F = C - P + 2}$$

Where F = Number of degrees of freedom,

C = Number of components

P = Number of phases of the system

For understanding the various applications of phase rule a clear understanding of the various terms, *phases (P), components (C)* and *degrees of freedom (F)* present in the phase rule, is essential which have their specific meanings.

[2] Components (C)

The minimum number of independently variable constituents in terms of which the composition of each phase of a heterogeneous system can be expressed directly or in the form of a chemical equation are called the components of system (C).

For example, a system consisting of a solution of sugar in water *(P = 1 i.e. solution phase)* is a two-component system because the solution phase present in the system consists of two constituents—water and sugar.

Some common examples related to the components are as follows:

(a) Consider the following system consisting of ice, water and vapour in equilibrium.

$$\text{Ice } (s) \rightleftharpoons \text{Water } (l) \rightleftharpoons \text{Vapour } (g)$$

The system consists of three phases *ice, water* and *vapour phase.* The chemical substance present in each phase is H_2O. Therefore, the composition of each phase is expressed in terms of H_2O. Hence, it is called *one-component system.*

(b) The saturated solution of sodium chloride consists of three *phases—solid sodium chloride, salt solution* and *water vapour* in equilibrium.

$$\text{NaCI } (s) \rightleftharpoons \text{NaCI } (aq) \rightleftharpoons \text{Water Vapour } (g)$$

$$\text{Ca } CO_3(s) \rightleftharpoons CaO(s) + CO_2(g)$$

The chemical composition of each phase of the system can be expressed if we consider two chemical constituents NaCI and water as shown below.

	Phase		*Components*
(i)	NaCI (s)	=	$NaCI + 0H_2O$
(ii)	NaCI(aq)	=	$yNaCI + xH_2O$
(iii)	$H_2O(g)$	=	$0NaCI + H_2O$

Hence, it is a *two-component system.*

(c) The system, $CuSO_4.5H_2O\,(s) \rightleftharpoons CuSO_4.3H_2O\,(s) + 2H_2O\,(g)$ is a three-phase and *two component system.* It requires two constituents $CuSO_4$ and H_2O to express the composition of each phase of the system.

[3] Degree of Freedom (F)

The smallest number of independently variable factors such as temperature, pressure and concentration which must be required in order to define the system completely are called the degree of freedom. Degree of freedom of a system is also known as variance.

When a system having no degree of freedom

$F = 0$ it is called *non-variant system* or *invariant system.*

When a system having only one degree of freedom

$F = 1$ it is a *univariant* or a *monovariant system.*

Similarly, a system having two degrees of freedom

$F = 2$ is a *bivariant system and so on.*

The term degree of freedom can be understood with the help of following examples:

(a) *The system ice* $\rightleftharpoons$ *water* $\rightleftharpoons$ *vapour* has no degree of freedom *(i.e.,* $F = 0$). The three phases of water *i.e.* ice, liquid water and vapour can exist together in equilibrium only at a particular-temperature and pressure (corresponding to the freezing point) and no factor is necessary to be specified to define the system. Hence, *a system consisting of ice, water and vapour in equilibrium has no degree of freedom i.e. it is a non-variant system.*

(b) *For a mixture of gases,* the number of degrees of freedom is three *(i.e.,* $F = 3$). Such a system can be completely defined when the temperature, pressure and composition are fixed. In this case, the remaining factor i.e., volume gets automatically fixed. For example, a gaseous mixture consisting of 70% N_2 and 30% O_2 at 22°C and 1 atm pressure is completely defined and does not require any other information for its description. Hence, *a system consisting of a mixture of gases has three degrees of freedom i.e. it is a trivariant system.*

(c) *For a saturated LiCl solution,* the number of degrees of freedom is one *(i.e.,* $F = 1$). This is because the system can be completely defined by specifying the temperature only. The other two factors *i.e.* composition and vapour pressure get automatically fixed when the temperature is fixed. *Hence, a system consisting of a saturated LiCl solution is a univariant system.*

2 DERIVATION OF PHASE RULE EQUATION

The Gibb's phase rule can be derived on the basis of thermodynamic principle as follows.

Let us consider a heterogeneous system consisting of $P(P_1, P_2, P_3 \ldots P)$ number of phases and $C(C_1, C_2, C_3 \ldots C)$ number of components in equilibrium. *Let us assume that the system is non-reacting i.e. the passage of a component from one phase to another does not involve any chemical reaction.* When the system is in equilibrium state it can be explained completely by specifying the following variables:

(i) Pressure (ii) Temperature

(iii) Composition of each phase.

(a) *Total number of variables required specifying the state of system:*

(i) Temperature: same for all phases

(ii) Pressure: same for all phases

(iii) Concentration

Independent concentration variables for one phase with respect to the C components $= C - 1$ [$\because$ Conc. of last component is independent]

$\therefore$ Independent concentration variables for P phases with respect to the C components $= P(C - 1)$

Total number of variables $= P(C - 1) + 2$ ...(1)

(b) *The total number of equilibria:*

The various phases present in the system can remain in equilibrium only when the *chemical potential* (μ) of each component is the same in each phases, *i.e.*

μ_1, P_1	=	μ_1, P_2	=	μ_1, P_3	=	...	=	μ_1, P	Component 1
μ_2, P_1	=	μ_2, P_2	=	μ_2, P_3	=	...	=	μ_2, P	Component 2
:		:		:				:	
:		:		:				:	
:		:		:				:	
μ_C, P_1	=	μ_C, P_2	=	μ_C, P_1	=	...	=	μ_C, P	Component C

(a) For each component the no of equilibria for P phases $= (P - 1)$

(b) For C component the no of equilibria for P phases $= C(P - 1)$

Total no. of equilibria involved (E) $= C(P - 1)$ ...(2)

From eq. 1 & 2 we get

$$F = [P(C - 1) + 2] - [C(P - 1)]$$

$$F = [CP - P + 2 - CP + C]$$

$$\boxed{F = C - P + 2}$$

This above equation is Gibb's phase rule equation.

Some conclusions from the phase rule equation:

(a) *For a system having a specified number of components, the greater the number of phases, the lesser is the number of degrees of freedom.* For example,

(i) When the system consists of only one phase, we have

$C = 1$ and $P = 1$

So, according to the phase rule,

$F = C - P + 2 = 1 - 1 + 2 = 2$. The system has two degrees of freedom.

(ii) When the system consists of two phases in equilibrium, we have

$C = 1$ and $P = 2$

$F = C - P + 2 = 1 - 2 + 2 = 1$. The system is *monovariant.*

(b) *A system having a given number of components and the maximum possible number of phases in equilibrium is non-variant.*

For a one component system, the maximum possible number of phases is three. When a one-component system has three phases in equilibrium, it has no degree of freedom or non-variant system.

(c) *For a system having a given number of phases, the larger the number of components, the greater will be the number of the degrees of freedom of the system.*

For example,

For one-component system: $C = 1, \quad P = 2$

$\therefore \quad F = C - P + 2 = 1 - 2 + 2 = 1$

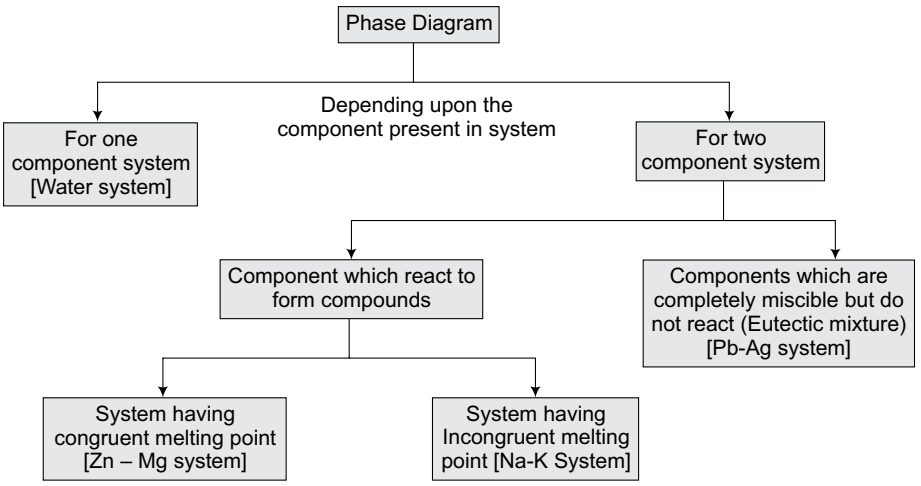
For two-component system: $C = 2, \quad P = 2$

$\therefore \quad F = C - P + 2 = 2 - 2 + 2 = 2$

The two-component system has a higher number of degrees of freedom.

3 PHASE DIAGRAMS

The graphical presentation giving the conditions of pressure and temperature under which the various phases are existing and transform from one phase to another is known as the phase diagram of the system. A phase diagram consists of *areas, curves* or *lines* and *points.*



4 PHASE RULE FOR ONE-COMPONENT SYSTEMS

The least number of phases possible in any system is one. So, according to the phase rule equation, a one-component system should have a maximum of two degrees of freedom.

When $C = 1, \quad P = 1$

So, $F = C - P + 2 = 1 - 1 + 2 = 2$

Hence, a one-component system requires a maximum of two variables to be fixed in order to define the system completely. The two variables are temperature and pressure. So, phase diagrams for one component system can be obtained by plotting P vs T.

In case of a one-component system, phase diagram consists of *areas, curves* or *lines* and *points* which provide the following informations regarding the system:

Point on a phase diagram represents a non-variant system.

Area represents a bivariant system

Curve or a *line* represents a univariant system.

Water system and the sulphur system are the example of one component systems.

[1] Water System

Water is a one component system which is chemically a single compound involved in the system. The three possible phases in this system are: ice (solid phase), water (liquid phase) and vapour (gaseous phase).

Hence, *water constitutes* a *three-phase, one-component system.* Since water is a three-phase system, it can have the following equilibria:

$$\text{ice} \rightleftharpoons \text{vapour},$$

$$\text{ice} \rightleftharpoons \text{water};$$

$$\text{water} \rightleftharpoons \text{vapour}$$

The existence of these equilibria at a particular stage depends upon the conditions of temperature and pressure, which are the variables of the system. If the values of vapour pressures at different temperatures are plotted against the corresponding temperatures, the phase diagram of the system is obtained.

The phase diagram of the water system is shown in Fig. 2.1. The explanation of the phase diagram of water system is as follows:

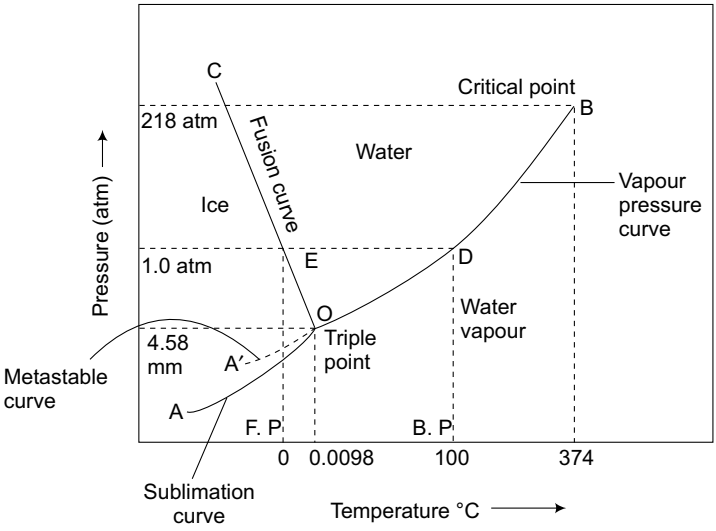


Fig. 2.1 Phase diagram of water system

(a) Curves

The phase diagram of the water system consists of three stable curves and one metastable curve, which are explained as follows:

(i) Curve *OB*: The curve *OB* is known as *vapour pressure curve of water* and tells about the vapour pressure of water at different temperatures. Along this curve, the two phases—*water* and *vapour* exist together in equilibrium.

At point *D,* the vapour pressure of water become equal to the atmospheric pressure (100°C), which represents the boiling point of water. The curve *OB* finishes at point B (temp. 374°C and pressure 218 atm) where the liquid water and vapour are indistinguishable and the system has only one phase. This point is called the *critical point.*

Applying the phase rule on this curve,

$$C = 1 \quad \text{and} \quad P = 2$$

$$F = C - P + 2 = 1 - 2 + 2 = 1$$

Hence, the curve represents a *univariant system.* This explains that only one factor (either temperature or pressure) is sufficient to be fixed in order to define the system.

(ii) Curve *OA*: *It is known as sublimation curve of ice* and gives the vapour pressure of solid ice at different temperatures. Along sublimation curve, the two phases *ice* and *vapour* exist together in equilibrium. The lower end of the curve *OA* extends to absolute zero (–273°C) where no vapour exists.

	Area	*Phase exits*	*Component*
(i)	Area *AOC*	ice	H_2O
(ii)	Area *COB*	water	H_2O
(iii)	Area below *BOA*	vapour	H_2O

Thus, for every area contains

$$C = 1 \quad \text{and} \quad P = 1$$

Therefore, applying phase rule on areas

$$F = C - P + 2 = 1 - 1 + 2 = 2$$

Hence, each area is a *bivariant system.* So, it becomes necessary to specify both the temperature and the pressure to define a one phase-system.

Table 2.1: Some salient features of the water system

Curve/ area/ point	*Name of the system*	*Phases in equilibrium*	*No. of phase (P)*	*Degree of the freedom (F)*
Curve *OB*	Vapourisation curve	Liquid & vapour	02	01(Univariant)
Curve *OA*	Sublimation curve	Solid & vapour	02	01(Univariant)
Curve *OC*	Fusion curve	Solid & liquid	02	01(Univariant)
Curve *OA′*	Metastable vaporization curve	Liquid & vapour	02	01(Univariant)
Area *AOC*		Ice	01	02(Bivariant)
Area *BOC*		Water	01	02(Bivariant)
Area *AOB*		Vapour	01	02(Bivariant)
Point *O*		Ice & water & vapour	03	0(Invariant)

5 TWO-COMPONENT SYSTEMS

When the two independent components are present in a heterogeneous system, the system is referred to as a *two-component system.* Hence, according to the phase rule, for a two-component system having one phase,

$$F = C - P + 2 = 2 - 1 + 2 = 3$$

Therefore, the two component system having one phase will have three degrees of freedom or three variables would be required to define the system. The three variables are pressure *(P),* temperature (*T*) and concentration *(C).* This will require a three-dimensional phase diagram for the study of a two-component system. However, in order to simplify the study, a two-component system is usually studied in the form of a condensed system. A condensed system can be studied by reducing a comparatively less important variable. This reduces the degree of freedom of the system by 1 and the system can easily be studied with the help of a two-dimensional phase diagram.

It can have a maximum of following four phases:

Solid lead, Solid silver, Solution of molten silver & lead and Vapours

The boiling points of silver and lead are considerably high and the vapour pressure of the system is very low. So, the vapour phase can be ignored and the system can be studied as a *condensed system.* This system thus can be easily studied with the help of a two dimensional $T - C$ diagram and the reduced phase rule equation, $F' = C - P + 1$, can be used. This system is generally studied at constant pressure (atmospheric). The phase diagram of Lead-Silver system is shown in Fig. 2.2.

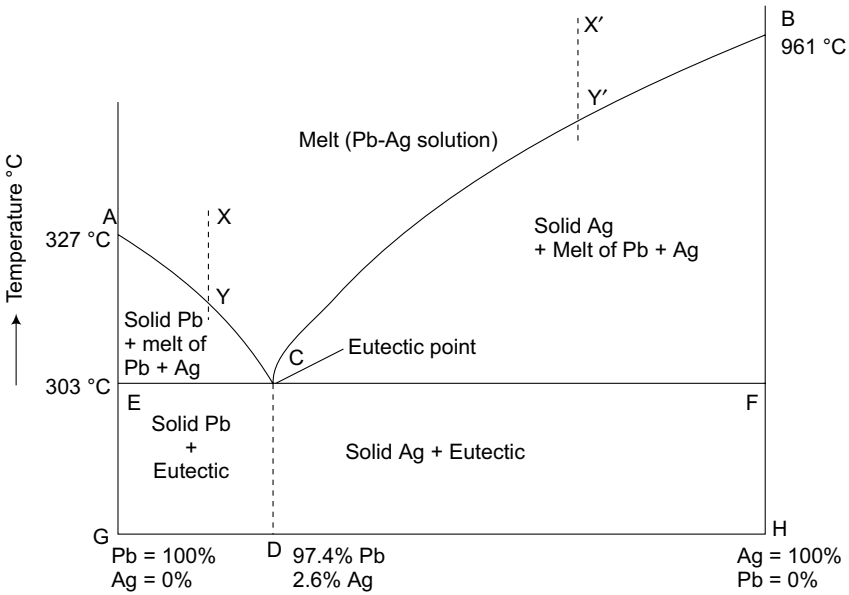


Fig. 2.2 Phase diagram of Pb-Ag system

(a) Curves

The phase diagram of the lead-silver system consists of following curves, which are explained as follows:

(i) Curve *AC (Freezing point curve of lead):* The *AC* curve shows the variation of the melting point of lead on addition of silver. The pure lead melts at 327°C (point *A*). Addition of silver lowers its melting point along curve *AC.* The added silver dissolves in molten lead to form Ag-Pb solution with the separation of some part of solid lead. Therefore, the two phases, solid lead and Ag-Pb solution remain together in equilibrium along the curve *AC.*

Hence, $P = 2$, (solid Pb and melt of Ag-Pb)

$C = 2$(Pb and Ag)

So, $C = 2$ and $P = 2$,

On applying the reduced phase rule

$$F' = C - P + 1 = 2 - 2 + 1 = 1$$ The system is univariant.

(iii) Area *BCF:* The area consists of two phases—solid Ag and a solution of Pb and Ag. Hence it is also univariant.

(iv) Area *DCFH*: This area also has the two phases which are solid Ag crystals and solid eutectic crystals. Hence $C = 2$ and $P = 2$, the system is *univariant.*

(iv) Area *CEGD*: The area also has the solid Pb crystals and solid eutectic crystals phases. The system is *univariant.*

Table 2.2: Some salient features of the Pb-Ag system.

Curve/ area/ point	*Name of the system*	*Phases in equilibrium*	*No. of phase (P)*	*Degree of freedom (F)*
Curve AC	Freezing curve of Pb	Pb & Melt (Pb+Ag Solution Pb & Ag)	02	01(Univariant)
Curve BC	Freezing curve of Ag	Ag & Melt Pb & Ag)	02	01(Univariant)
Area ACE	.....	Pb & melt	02	01(Univariant)
Area BCF	.....	Ag & melt	02	01(Univariant)
Area above ACB	.....	Liquid (melt)	01	02(Bivariant)
Area ECF	.....	Pb & Ag both in solid	02	01(Univariant)
Point O	Eutectic point	Pb, Ag & melt	03	0(invariant)

Desilverisation of Argentiferrous Lead (Pattinson's Process)

The process, which is used for the recovery of silver from argentiferrous lead is called *Pattinson's process* and involves the *desilverisation of lead* in accordance to the phase diagram of lead-silver system.

The argentiferrous lead contains a small percentage of silver (less than 0.1%). For its recovery, the argentiferrous lead is heated above its melting point when a liquid melt consisting of silver-lead solution is obtained. Now if the silver lead solution is cooled, then Pb continues to separate out and is regularly removed. In the end, a eutectic solution containing 2.6% Ag (corresponding to point *C*) is obtained. Thus, the above process increases the percentage of silver in the argentiferrous lead. Therefore, the eutectic mixture containing 2.6% silver can be treated for the recovery of silver profitably.

(B) Systems having Congruent Melting Point

A binary system is said to possess a congruent melting point when it melts at a sharp temperature to give a liquid of the same composition as that of the solid.

The components of a binary mixture at a certain stage enter into chemical combination and form one or more compounds (inter-metallic compounds) in *stochiometric* proportions. These compounds melt sharply at a constant temperature into a liquid having the same composition as that of the solid. The temperature at which such a compound melts is called the congruent melting point.

Some common examples of this type of system are zinc-magnesium system, mercury-thallium system, gold-tin system and ferric chloride-water system etc.

1 Zn-Mg system

Zn-Mg System is a two-component system and possess a congruent melting point. The phase diagram of Zn-Mg system is shown in Fig. 2.3. In this system, the two components are zinc and magnesium, which melt at 419°C and 650°C respectively which are represented as point *B* and *A* in the phase diagram of the system. Both metals enter into chemical combination and form an intermetallic compound $MgZn_2$ and melts at 590°C to give a liquid of the same composition. Hence, 590°C is the congruent melting point of the system.

In the reduced form, the system has the following four phases:

Solid magnesium, solid zinc, solid $MgZn_2$ and liquid solution of Zn and Mg.

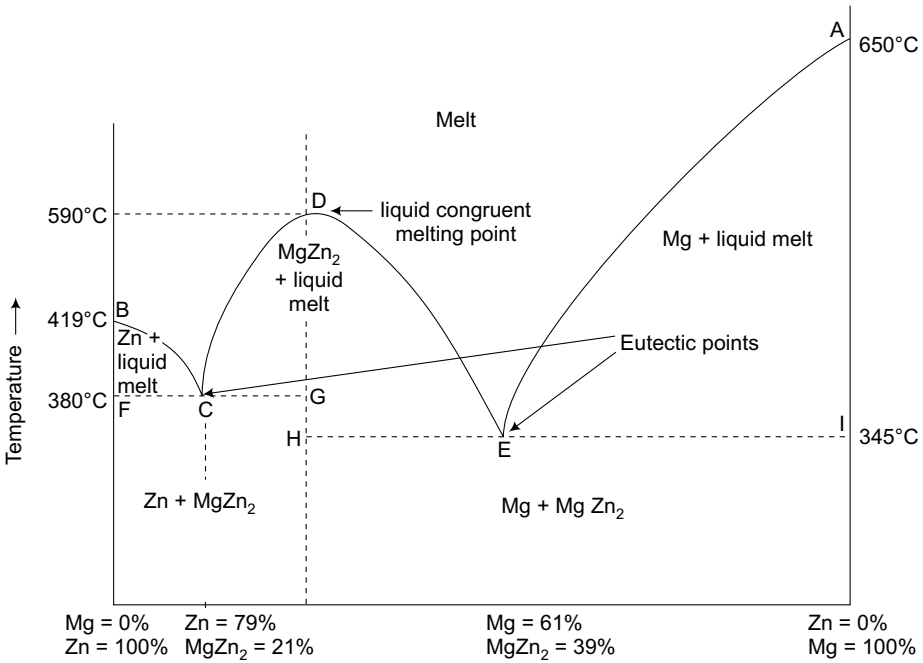


Fig. 2.3 Phase diagram of Zn-Mg system [Congruent melting point system]

On applying the reduced phase rule

$$F' = C - P + 1 = 1 - 2 + 1 = 0$$

Therefore, at point *D* constitutes a *non-variant* system.

(ii) Point *E* (Eutectic point): Point *E* represents the eutectic point of the system at a temperature of 345°C which is the least melting point of Mg-$MgZn_2$ system. Here, also three phases existing together in equilibrium at point *E* are solid Mg, solid $MgZn_2$ and liquid $MgZn_2$.

Hence, $C = 2$ and $P = 3$,

$F' = C - P + 1 = 2 - 3 + 1 = 0$ The system is *non-variant.*

(iii) Point *C* (Eutectic point): This point also represents the eutectic point (380°C) which is the least melting point of Zn-$MgZn_2$ system. At this point, the three phases—solid Zn, solid $MgZn_2$ and liquid $MgZn_2$ exist together in equilibrium. Therefore,

$$C = 2 \text{ and } P = 3$$

$$F' = C - P + l = 2 - 3 + 1 = 0$$

Hence, point *C* represents a *non-variant system.*

(c) Areas

The phase diagram of zinc-magnesium system consists of many areas. The area above the curve *BCDEA* constitutes a single phase system. The phase present in this area is a liquid melt consisting of a liquid solution of zinc and magnesium.

Hence $C = 2$ and $P = 1$,

$F' = C - P + 1 = 2 - 1 + 1 = 2$ that the system is *bivariant.*

Most of the other areas of the Zn-Mg system consists of two phases and they are *univariant systems* as represented in the phase diagram. These areas are explained in datail in table 2.3.

Table 2.3: Some salient features of the Zn-Mg system

Curve/ area/ point	*Phases in equilibrium*	*No. of phase(P)*	*Degree of the freedom (F)*
Curve BC (Freezing curve of Zn)	Zn & Melt (Containing Zn & $MgZn_2$)	02	01(Univariant)
Curve AE (Freezing curve of Mg)	Mg & Melt (Containing Zn & $MgZn_2$)	02	01(Univariant)
Curve CD	$MgZn_2$ & melt (Containing Mg & $MgZn_2$)	02	01(Univariant)
Curve DE	$MgZn_2$ & melt (Containing Zn & $MgZn_2$)	02	01(Univariant)
Curve CDE	$MgZn_2$ & melt	02	01(Univariant)
Area above BCDEA	Liquid (Melt of Zn, Mg & $MgZn_2$)	01	02(Bivariant)

Area BCF	Zn & Melt(Containing Zn & $MgZn_2$)	02	01(Univariant)
Area DCG	$MgZn_2$ & Melt(Containing Zn & $MgZn_2$)	02	01(Univariant)
Area DEH	$MgZn_2$ & Melt(Containing Mg & $MgZn_2$)	02	01(Univariant)
Area AEI	Mg & melt (containing Mg + $MgZn_2$)	02	01(Univariant)
Area below line FCG	Zn & $MgZn_2$ (both solid)	02	01(Univariant)
Area below line HEI	Mg & $MgZn_2$ (Both solid)	02	01(Univariant)
Point C (Eutectic)	Zn, $MgZn_2$ & Melt(Containing Zn & $MgZn_2$)	03	0(invariant)
Point E (Eutectic)	Mg, $MgZn_2$ & Melt(Containing Mg & $MgZn_2$)	03	0(invariant)

(C) Incongruent Melting Point System

There are several systems in which components combine together to form one or more compounds which are unstable and do not possess congruent melting points.

A system (compound) is said to possess incongruent melting point, if on heating it decomposes much below its melting point and forms a new solid phase and a solution having different composition from the corresponding solid state. It has no sharp melting point. The decomposition at this temperature is known as transition or meritectic or peritectic rection and the temperature (the incongruent melting point) is known as transition or meritectic or peritectic temperature.

$$\text{Original solid} \rightleftharpoons \text{new solid} + \text{solution (melt)}$$

Examples: Following are some examples of the binary systems which possess incongruent melting point:

(i) Gold-antimony system
(ii) Sodium-bismuth system
(iii) Sodium-potassium system
(iv) Sodium sulphate-water system
(v) Potassium chloride-copper chloride system

1 Na-K system

This is a two-component system having incongruent melting point. The melting points of sodium and potassium are 97.8°C and 63.8°C respectively which are shown in the phase diagram in Fig. 2.4. Both elements chemically combine together in the ratio of 2:1 to form a compound Na_2K. But this compound is unstable and decomposes into solid Na and melt at a temperature of 70°C. It is the incongruent melting point or peritectic temperature of this system. This system consists of four phases i.e. Solid K, Solid Na, Solid Na_2K and Liquid of Na and K.

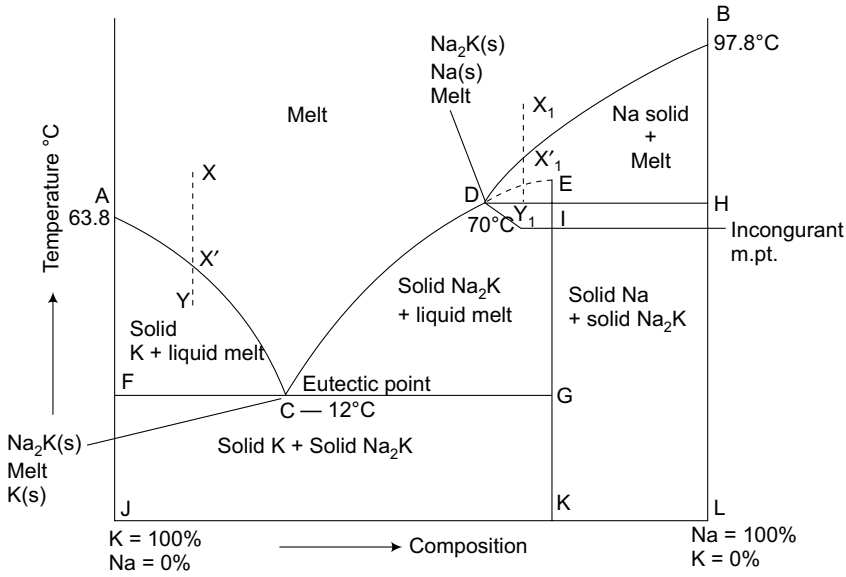


Fig. 2.4 Phase diagram of Na-K [Incongruent melting point system]

As the pressure does not have any effect on this type of equilibria hence the degree of freedom for such a system is reduced by one, So, reduced phase rule is applicable on the Na-K system. ($F' = C - P + 1$)

The phase diagram contains the following curves, points and areas.

1 Curves

(i) *Curve AC.* (Freezing point curve of potassium): This curve shows the lowering in freezing point of potassium by addition of sodium and continues till the point '*C*' is reached. Along this curve, potassium (K) separates out as solid phase. A new phase Na_2K separates out at point *C*. The two phases exist in equilibrium along this curve.

$P = 2$, [K(solid) and liquid (Na-K melt)] and $C = 2$

On applying the reduced phase rule

$F' = C - P + 1 = 2 - 2 + 1 = 1$ Hence, the system is univariant along this curve.

(ii) *Curve CD.* (Fusion curve of Na_2K): Along this curve the two phases exists in equilibrium. The Na_2K is stable along this curve. [If further the compound would be stable as having congruent melting point then the curve may be plotted up to the stable melting point *E* of the compound, which is shown in the phase diagram of the Na-K system].

$$Na_2K(s) \rightleftharpoons \text{Liquid}$$

$P = 2$ [Na_2K(s) and liquid] and $C = 2$

Hence $P = 3, \quad C = 2$

So, according to the phase rule

$$F' = C - P + 1 = 2 - 3 + 1 = 0$$

Thus, the system is invariant at point D.

3 Areas

(i) *Area above ACDB*: The area above the *ACDB* contains only liquid phase i.e. melt of Na, K and Na_2K.

Hence $P = 1, \quad C = 2$

On applying the reduced phase rule

$$F' = C - P + 1 = 2 - 1 + 1 = 2$$

Thus, the system is bivariant in the area above *ACDB*.

(ii) *Area ACF:* It consists of two phases solid K and liquid.

(iii) *Area ECG:* This area consists of two phases solid Na_2K and liquid.

(iv) *Area BDH:* It consists of the two phases, which exist in equilibrium i.e. solid Na and liquid.

(v) *Area below FCG:* It consists of solid K and solid Na_2K.

(vi) *Area IHLK:* This area consists of two phases solid Na and solid Na_2K.

All areas from (ii) to (vi) are having two phases and two components

Hence $P = 2, \quad C = 2$

On applying the reduced phase rule

$$F' = C - P + 1 = 2 - 2 + 1 = 1$$

$$F' = 1$$

Therefore, all the above areas represent univariant systems.

Table 2.4: Some salient features of the Na-K system.

Curve/ area/ point	*Phases in equilibrium*	*No. of phase(P)*	*Degree of the freedom (F)*
Curve AC (Freezing curve of K)	K & Melt (Containing K & Na_2K)	02	01(Univariant)
Curve BD (Freezing curve of Na)	Na & Melt (Containing Na & Na_2K)	02	01(Univariant)
Curve CD	Na_2K & Melt (Containing K & Na_2K)	02	01(Univariant)
Area above ACDB	Liquid (Melt of Na, K & Na_2K)	01	02(Bivariant)
Area ACF	K & Melt (Containing K & Na_2K)	02	01(Univariant)

Area BDH	Na & Melt (Containing Na & Na_2K)	02	01(Univariant)
Area CDIG	Na_2K & Melt (Containing K & Na_2K)	02	01(Univariant)
Area below line FGC	Na_2K & K (both solid)	02	01(Univariant)
Area below line IH	Na_2K & Na (Both solid)	02	01(Univariant)
Point C (Eutectic)	K, Na_2K & Melt (Containing K & Na_2K)	03	0(invariant)
Point I (Eutectic)	Na, Na_2K & Melt (Containing Na & Na_2K)	03	0(invariant)

6 APPLICATIONS OF THE PHASE RULE

Phase rule has wide applications in electronic industries, pharmaceutical science, medical science, etc. Some major applications of phase rule are as follows:

[1] Solders

Solder is an alloy, which is homogenous mixture having melting point lower than that of the corresponding metal pieces, which have to be joined together. Solders have compositions somewhat different from the eutectics so that the freezing occurs *over a range* of temperatures. The quality of solder depends upon the formation of a surface alloy between the solder and parts of metals being used. The selection of solder alloy is based upon the melting point desired and the pieces of metals to be joined. Some essential qualities of the solder are as follows:

(1) Melting point of the solder should be less than the material to be soldered.

(2) Solder should spread in liquid form and also form homogeneous mixture with the metals.

Some common solders which are available in the market are:

(i) 'Soft solder' alloy of Pb and Sn.

(ii) 'Plumber alloy' contains Pb = 67% and Sn = 33%.

(iii) 'Half-half alloy' contains Pb = 50% and Sn = 50%

Half-half alloy is commonly used for soldering the pipes with a bright surface finishing after soldering but very expensive. Also due to high contents of tin, it is not widely applicable for several electric appliances. The solder which contains about 60% Pb is used in the electrical wires.

[2] Safety Plug

Safety plugs are also known as the safety fuses. It is an alloy having low melting point, used to ensure the safe working and avoid accidents. Safety fuses are used in buildings to protect them against fires. One alloy is woods metal, which is used in the safety fuses. This alloy melts at 65°C and consists of the composition woods metal Bi = 50%, Pb = 25%, Sn = 12.5% and Cd = 12.5%.

and the frozen liquid (ice) is directly converted into a gaseous state. The removed water is stored again through the condensers. The material or tissue is left almost as a skeleton and original matter can be obtained by adding water. By this process shrinkage of flowers is eliminated or minimized.

All botanical samples, fruits and vegetables can be freeze-dried. This technique is also used in pharmaceutical industry, museums, taxidermy, floral industry and camping/hiking food processors.

QUESTIONS FOR EXAMINATION

[A] Short Answer Type Questions

1. What is the phase rule?
2. Define phase with some examples.
3. What are the components of a system? Explain with examples.
4. How many phases are present in a homogeneous system?
5. How many phases and components are present in each of the following systems:

   (i) $NH_4CI\ (s) \rightleftharpoons NH_3\ (g) + HCI\ (g)$

   (ii) An aqueous solution of salt (NaCl)

   (iii) Toluene in equilibrium with its vapour
6. Define degree of freedom of a system and explain with examples.
7. How many degrees of freedom are possible for a mixture of gases?
8. Calculate the maximum possible number of phases and degrees of freedom for a one-component system.
9. Is it possible to have a quadruple point (point having $P = 4$) in the phase diagram of a one-component system?
10. What is metastable equilibrium? Explain with reference to the water system.
11. What do you understand by a triple point? What is the variance of a one-component system at this point?
12. What is a condensed or reduced system? Explain with some example.
13. Can the phase rule equation in its original form be applied to a condensed system? Write the reduced phase rule equation.
14. What do you understand by a eutectic point?
15. Draw the phase diagram for Pb-Ag system.
16. What is the Pattinson's process?
17. Define congruent melting point and name a system, which forms a compound with such a melting point.

Unit IV

FUELS AND COMBUSTION

Fuels: Introduction – Classification of fuels – Coal - Analysis of coal (proximate and ultimate) – Carbonization – Manufacture of metallurgical coke (Otto Hoffmann method) – Petroleum – Manufacture of synthetic petrol (Bergius process) – knocking – Octane Number – Diesel Oil – Cetane Number – Natural Gas – Compressed Natural Gas (CNG) – Liquefied Petroleum Gases (LPG) – Power Alcohol and Biodiesel. **Combustion of Fuels:** Introduction – Calorific Value – Higher and Lower Calorific Values – Theoretical calculation of Calorific Value – Ignition Temperature – Spontaneous Ignition Temperature – Explosive Range – Flue Gas Analysis (ORSAT Method).

CHAPTER 6: Fuels

INTRODUCTION

The various types of fuels like liquid, solid and gaseous fuels are available for firing in boilers, furnaces and other combustion equipments. The selection of right type of fuel depends on various factors such as availability, storage, handling, pollution and landed cost of fuel.

The knowledge of the fuel properties helps in selecting the right fuel for the right purpose and efficient use of the fuel.

The following characteristics, determined by laboratory tests, are generally used for assessing the nature and quality of fuels.

CLASSIFICATION OF FUELS

Chemical Fuels: It is of two types.

(i) Primary Fuels : It occurs in nature as such. ex. coal, petroleum, natural gas.

(ii) Secondary Fuels: It is derived from primary fuels ex.: coke, gasoline, coal gas.

Primary Fuel or Natural Fuel

Primary Fuel or Natural Fuel

Solid	Liquid	Gaseous
Wood	Crude oil	Natural gas
Peat		
Lignite		
Coal		
Dung		

Secondary or Derived Fuel

Secondary or Derived Fuel

Solid	Liquid	Gaseous
Coke	Tar	Coal Gas
Charcoal	Kerosene	Water gas
Petroleum	Diesel	Oil gas
Coke	Petrol	Bio gas
Coal	Fuel oil	Blast furnace gas
Briquette	Synthetic Gasoline gas	Coke over gas

I. SOLID FUELS

COAL

Coal is a highly carbonaceous matter that has been formed as a result of alteration of vegetable matter (eg., plants) under certain favourable conditions. It is chiefly composed of C, H, N and O besides non-combustible inorganic matter.

The successive stages in the transformation of vegetable matter into coal are– wood, peat, lignite, bituminous coal, steam coal and anthracite. Anthracite is probably the purest form of coal and contains 95 % carbon.

CLASSIFICATION OF COAL

(a) Peat

1. Peat is the first stage in the formation of coal.
2. Its calorific value is about 4000-5400 k cal/kg.
3. It is an uneconomical fuel due to its high proportion of (80 -90%) moisture and lower calorific value.
4. It is a brown fibrous mass.

(b) Lignite

1. Lignite is an intermediate stage in the process of coal
2. Formation.
3. Its calorific value is about 6500-7100 kcal/kg
4. Due to the presence of high volatile content, it burns with long smoky flame.

(c) Bituminous Coal

Bituminous coal is further sub-classified on the basis of its carbon content into three types as:

1. Sub- bituminous coal,
2. Bituminous coal and
3. Semi-bituminous coal.

(d) Anthracite

1. Anthracite is the superior grade of coal.
2. Its volatile, moisture and ash contents are very less.
3. Its calorific value is about 8650 kcal/kg

ANALYSIS OF COAL

The quality of a coal is ascertained by the following two types of analysis are made.

Proximate Analysis

- *Proximate Analysis indicates the percentage by weight of the fixed carbon, volatiles, ash, and moisture content in coal.*

- The amounts of fixed carbon and volatile combustible matter directlycontribute to the heating value of coal.
- Fixed carbon acts as a main heat generator during burning. High volatile matter content indicates easy ignition of fuel.
- The ash content is important in the design of the furnace grate, combustion volume, pollution control equipment and ash handling systems of a furnace.

Significance of Various Parameters in Proximate Analysis

(a) Fixed Carbon

Fixed Carbon is the solid fuel left in the furnace after volatile matter is distilled off. It consists mostly of carbon but also contains some hydrogen, oxygen, sulphur and nitrogen not driven off with the gases. Fixed carbon gives a rough estimate of heating value of coal.

(b) Volatile Matter

Volatile Matters are the methane, hydrocarbons, hydrogen and carbon monoxide, and incombustible gases like carbon dioxide and nitrogen found in coal. Thus the volatile matter is an index of the gaseous fuels present. Typical range of volatile matter is 20 to 35%. The loss in weight of the sample is found out and the % of the volatile matter is calculated as:

$$\% \text{ of volatile matter in coal} = \frac{\text{loss in weight of the coal}}{\text{weight of air - dried coal}} \times 100$$

Volatile Matter

- Proportionately increases flame length, and helps in easier ignition of coal.
- Sets minimum limit on the furnace height and volume.
- Influences secondary air requirement and distribution aspects.
- Influences secondary oil support

(c) Ash Content

Ash is an impurity that will not burn. Typical range is 5 to 40%. After the analysis of volatile matter, the crusible with residual coal sample is heated without lid as $700 \pm 50^{o}C$ for 1/2 an hour in a muffle furnace. The loss in weight of the sample is found out and the % of ash content is calculated as:

$$\% \text{ of ash content in coal} = \frac{\text{weight of ash formed}}{\text{weight of air - dried coal}} \times 100$$

Ash

- Reduces handling and burning capacity.
- Increases handling costs.
- Affects combustion efficiency and boiler efficiency
- Causes clinkering and slagging.

(d) Moisture Content

Moisture in coal must be transported, handled and stored. Since it replaces combustible matter, it decreases the heat content per kg of coal. Typical range is 0.5 to 10% . The loss in weight of the sample is found out and the % of moisture is calculated as:

$$\% \text{ of moisture in coal} = \frac{\text{loss in weight of the coal}}{\text{weight of air - dried coal}} \times 100$$

Moisture

- Increases heat loss, due to evaporation and superheating of vapour
- Helps, to a limit, in binding fines.
- Aids radiation heat transfer.

(e) Sulphur Content

Typical range is 0.5 to 0.8% normally.

Sulphur

- Affects clinkering and slagging tendencies
- Corrodes chimney and other equipment such as air heaters and economizers
- Limits exit flue gas temperature

Ultimate Analysis

The ultimate analysis indicates the various elemental chemical constituents such as Carbon, Hydrogen, Oxygen, Sulphur, etc. It is useful in determining the quantity of air required for combustion and the volume and composition of the combustion gases. This information is required for the calculation of flame temperature and the flue duct design etc.

(a) Determination of carbon and hydrogen in coal

A known amount of coal is burnt in presence of oxygen thereby converting carbon and hydrogen of coal into –

i. $CO_2\ (C + O_2 \longrightarrow CO_2)$ and

ii. $H_2O (H_2 + 1/2\, O_2 \longrightarrow H_2O)$

respectively. The products of combustion CO_2 and H_2O are passing over weighed tubes of anhydrous $CaCl_2$ and KOH which absorb H_2O and CO_2 respectively.

The increase in the weight of $CaCl_2$ tube represents the weight of water formed while the increase in the weight of KOH tube represents the weight of CO_2 formed.

The percentage of carbon and hydrogen in coal can be calculated in the following way-

The weight of coal sample taken $= x\ g$

The increase in the weight of KOH tube $= y\ g$

The increase in the weight of $CaCl_2$ tube $= z\ g$

Consider the following reaction:

$$C(12) + O_2 \longrightarrow CO_2\ (44)$$

44 g of CO_2 contains 12 g of carbon

$$\text{Therefore } y\ g \text{ of } CO_2 \text{ contains} = \frac{y \times 12}{44\, x} \text{ g of carbon}$$

$$x\ g \text{ of coal contains} = \frac{12\, y}{44} \text{ g carbon}$$

$$\% \text{ of carbon in coal} = \frac{12\, y}{44\, x} \times 100 \quad \text{(or)}$$

$$\% \text{ of carbon in coal} = \frac{\text{increase in weight of KOH tube}}{\text{weight of coal sample taken}} \times \frac{12}{44} \times 100$$

Significance

It is the sum total of fixed carbon and the carbon present in the volatile matters like CO, CO_2, hydrocarbons. Thus, total carbon is always more than fixed carbon in any coal. High total carbon containing coal will have higher calorific value.

(b) Determination of hydrogen

Consider the following reaction:

$$H_2(2) + \frac{1}{2}O_2 \longrightarrow H_2O(18)$$

18 g of water contains 2 g of hydrogen.

z g of water contains $= \frac{2z}{18}$ of hydrogen

x g of coal contains $= \frac{2z}{18}$ g of hydrogen

% of hydrogen in coal $= \frac{2z}{18x} \times 100$ (or)

% of hydrogen $= \frac{\text{increase in weight of } CaCl_2 \text{ tube}}{\text{weight of coal sample taken}} \times \frac{2}{18} \times 100$

Significance

It increases the calorific value of the coal. It is associated with the volatile matter of the coal. When the coal containing more of hydrogen is heated, it combines with nitrogen present in coal forming ammonia. Ammonia is usually recovered as $(NH_4)_2SO_4$, a valuable fertilizer.

(c) Determination of Nitrogen

This is done by Kjeldhal's method: A known amount of powdered coal is heated with concentrated sulphuric acid in the presence of K_2SO_4 and $CuSO_4$ in a long necked Kjeldhal's fask. This converts nitrogen of coal to ammonium sulphate. When the clear solution is obtained (ie., the whole of nitrogen is converted into ammonium sulphate), it is heated with 50 % NaOH solution and the following reaction occurs:

$$(NH_4)_2SO_4 + 2\,NaOH \longrightarrow Na_2SO_4 + 2\,NH_3 + H_2O$$

The ammonia thus formed is distilled over and is absorbed in a known quantity of standard 0.1 NH_2SO_4 solution. The volume of unused 0.1 NH_2SO_4 is then determined by titrating against standard NaOH solution. Thus, the amount of acid neutralized by liberated ammonia from coal is determined using the formula.

$$\% \text{ of Nitrogen in Coal} = \frac{14\times \text{ volume of acid used}\times \text{ normality}}{1000\ x} \times 100$$

$$= \frac{1.4\times \text{ volume of acid used} \times \text{ normality}}{x}$$

Significance

Presence of nitrogen decreases the calorific value of the coal. However, when coal is carbonized, its N_2 and H_2 combine and form NH_3. Ammonia is recovered as $(NH_4)_2SO_4$, a valuable fertilizer.

(d) Determination of sulphur in coal

A known amount of coal is burnt completely in Bomb calorimeter in presence of oxygen.

Ash thus obtained contains sulphur of coal as sulphate which is extracted with dil. HC*l*.

The acid extract is then treated with B*a*Cl_2 solution to precipitate sulphate as B*a*SO_4. The precipitate is filtered, washed, dried and weighed. From the weight of B*a*SO_4, the percentage of sulphur in coal is calculated in the following way.

The weight of coal sample taken $= x\ g$

The weight of B*a*SO_4 precipitate $= y\ g$

Consider the following equations:

$$S(32) + 2O_2 \longrightarrow SO_4 \xrightarrow{BaCl_2} BaSO_4(233)$$

233 g of B*a*SO_4 contains 32 g of sulphur

Therefore, $y\ g$ of B*a*SO_4 contains $= \frac{32\ y}{233}$ g sulphur

Therefore, $x\ g$ of coal contains $= \frac{32\ y}{233}$ g sulphur

% of sulphur in the coal $= \frac{32\ y}{233} \times 100$

Significance

It increases the calorific value of the coal, yet it has the following undesirable effect.

The oxidation products of sulphur (SO_2, SO_3) especially in presence of moisture forms sulphuric acid which corrodes the equipment and pollutes the atmosphere.

(e) Determination of oxygen in coal

It is calculated indirectly in the following way-

% of oxygen in coal = 100 – % (C + H + N + S + ash).

Significance

The less the oxygen content, the better is the coal. As the oxygen content increases, its moisture holding capacity also increases.

CARBONIZATION

Carbonization (or carbonisation) is the term for the conversion of an organic substance into carbon or a carbon-containing residue through pyrolysis or destructive distillation. It is often used in organic chemistry with reference to the generation of coalgas and coal tar from raw coal.

Manufacture of Metallurgical Coke by Otto Hoffmann's Method

When bituminous coal (coal containing about 90 % carbon) is heated strongly in absence of air, the volatile matter escapes out and a while, lustrous, dense, strong, porous and coherent mass is left which is called metallurgical coke.

In order to (i) save the fuel for heating purpose and (ii) recover valuable by-products like coal gas, ammonia, benzol oil, tar etc. Otto Hoffmann developed a modern by-product coke oven. Here, the heating is done externally by a portion of coal gas produced during the process itself. It also utilizes the waste flue gases for heating the checker work bricks.

The oven consists of a number of narrow silica chambers, each about 10-12 m long, 3-4m tall and 0.4-0.45 m wide, erected side by side with vertical flues between them to form a sort of battery. Each chamber has a hole at the top to introduce the charge, a gas off take and a refractory lined cast iron door at each end for coke discharge. The oven works on heat regenerative principle ie. the waste gas produced during carbonization is utilized for heating. The ovens are charged from the top and closed to restrict the entry of air.

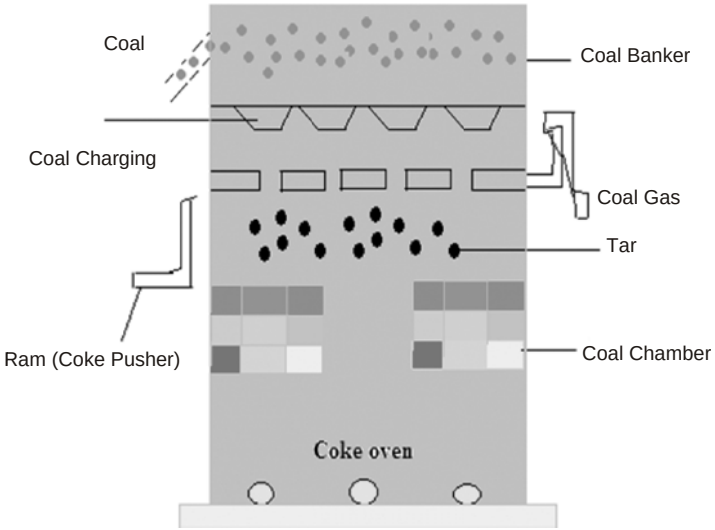


Figure 6.1 : Otto-Hoffmann's Coke Oven

Finely powdered, crushed coal is introduced through the charging hole at the top of the chambers which are then tightly closed at both ends to cut off the supply of air. The ovens are heated to 1200°C by burning producer gas. The air required for the combustion of the fuel is preheated in regenerators flanking the retorts, while the flue gases leave their acquired heat to one generator; the other generator is used for preheating the incoming air. The cycle goes on and the heating is continued until all the volatile matter has escaped.

It takes nearly 18 hours for carbonization of a charge. The heating of air-alone is required if the fuel gas is coal gas which has a high calorific value. If the fuel is producer gas or blast furnace gas, both air and fuel need to be preheated as they have low calorific value. When the carbonization is over, the red hot coke is pushed out into truck by a massive ram.

It is then quenched by spraying water (wet quenching). Alternatively, the red hot coke may be placed in a chamber and cooled by sending in inert gases from boilers. The inert gases are then circulated to boilers where they generate steam. This method is known as dry quenching. The dry quenched coke is cleaner, drier and stronger and contains lesser dust than the wet quenched. The yield is about 70%.

Characteristics of Metallurgical Coke

The most important industrial fuel is the metallurgical coke. This is used in the metallurgical industry, especially in the blast furnace. A good metallurgical coke must have following requirements:

- ***Purity:*** Low moisture and ash content are desirable in metallurgical coke. It must contain minimum percentage of phosphorous and sulphur.
- ***Porosity:*** High porosity is desirable in furnace coke to obtain high rate of combustion.
- ***Strength:*** The coke should be hard and strong to withstand pressure of ore, flux etc. in the furnace.
- ***Size:*** Metallurgical coke must be uniform and medium size.
- ***Calorific value:*** This should be high.
- ***Combustibility:*** It should burn easily.
- ***Reactivity:*** It refers to its ability to react with O_2, CO_2, steam and air. The metallurgical coke must have low reactivity.
- ***Cost:*** It must be cheap and readily available.

II. LIQUID FUELS

PETROLEUM

Petroleum or crude oil is a dark greenish -brown, viscous oil found deep in earth crust. It is composed mainly of various hydrocarbons (like straight-chain paraffins, cycloparaffins or naphthalenes, olefins and aromatics), together with small amounts of organic compounds containing oxygen, nitrogen and sulphur.

The average composition of crude oil is as follows:

Constituents	***Percentage (%)***
C	80-87
H	11-15
S	0.1-3.5
N + O	0.1-0.5

In countries like Germany and South Africa which do not have extensive petroleum deposits, motor fuels are derived from non-petroleum sources. Petroleum can be produced from coal by the following two methods.

6.7.1 Manufacutre of Synthetic Petrol (Bergius Process)

This method was first proposed by Bergius in Germany. It consists of converting low grade coals such as bitminuous coal into liquid and gaseous fuels by hydrogenating them in presence of iron oxide as catalyst.

The raw materials used in the process are coal dust, heavy oil and nickel oleate or tinoleate. A coal paste is prepared by mixing coal dust with heavy oil and catalyst. It is then pumped into the converter where the paste is heated to 450 °C under 200-250 atmosphere in pressure of hydrogen.

The reaction products mainly contain mixture of petroleum hydrocarbons

$$\text{Coal dust suspended in heavy oil} + H_2 \xrightarrow{450\ ^\circ C} \text{Mixture of hydrocarbons}$$

$$\downarrow \text{Condensation}$$

$$\text{Crude Oil}$$

Since the reaction is exothermic, the vapours leaving the converters are condensed in the condenser to give synthetic petroleum or crude oil. The oil is then fractionally distilled to give:

(i) Petrol, (ii) Middle Oil, (iii) Heavy Oil.

Middle Oil is again hydrogenated in presence of solid catalyst to produce more amount of petroleum Heavy oil is used for making paste with fresh coal dust which is required for this process.

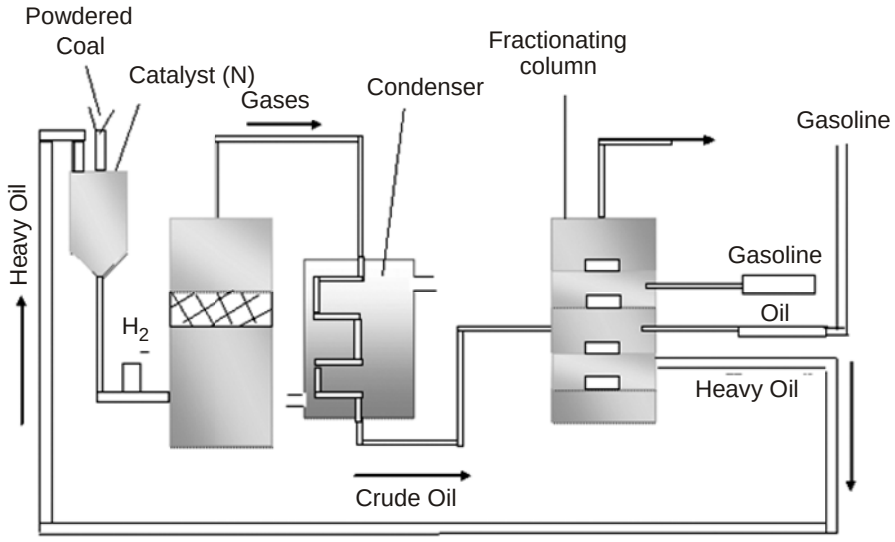


Figure 6.2 : Bergius Process

KNOCKING

In an internal combustion engine, a mixture of gasoline (petroleum) vapour and air is used as a fuel. After the initiation of the combustion reaction by spark in the cylinder, the flame should spread rapidly and smoothly through the gaseous mixture; thereby the expanding gas drives the piston down the cylinder. The ratio of the gaseous volume in the cylinder at the end of the suction stroke to the volume at the end of compression stroke of the piston is known as the compression ratio. The efficiency of an internal combustion engine increases with the compression ratio.

"Knocking is a kind of explosion due to rapid pressure rise occurring in an IC engine".

However, successful high compression ratio is dependent on the nature of the constituents present in the gasoline used. In certain circumstances, due to the presence of some constituents in the gasoline used, the rate of oxidation becomes so great that the last portion of the fuelair mixture gets ignited instantaneouslyproducing an explosive violence known as knocking. The knocking results in loss of efficiency, since this ultimately decreases the compression ratio. The phenomenon of knocking is not yet fully understood. However, it is noted that the tendency of fuel constituents to knock is in the following order:

Straight chain paraffins (n-paraffins) > branched chain paraffins (iso paraffins) > olefins> cycloparaffins (naphthalenes) > aromatics.

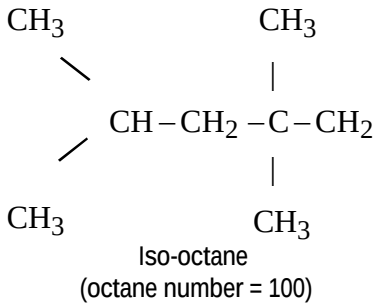
Thus, olefins of the same carbon chain length possess better antiknock properties than the corresponding paraffins and so on.

OCTANE NUMBER

The most common way of expressing the knocking characteristics of a combustion engine fuel is by octane number introduced by Edger in 1972. It has been found that n-heptane, $3–CH_2–CH_2–CH_2–CH_2–CH_2–CH_3$, knocks very badly and hence, its antiknock value has been arbitrarily given zero. On the other hand, iso-octane (2,2,4 –trimethylpentane) gives very little knocking, so its antiknock value has been given as 100.

"Thus, octane number (or rating) of a gasoline (or any other internal combustion engine fuel) is the percentage of iso-octane in a mixture of iso-octane and n-heptane", which matches the fuel under test in knocking characteristics.

Thus, if a sample of petrol gives as much of knocking as a mixture of 75 parts of iso-octane and 25 parts of n-heptane, then its octane number is taken as 75. The octane ratings of some common hydrocarbons are given in the table.



Iso-octane
(octane number = 100)

$CH_3 - CH_2 - CH_2 - CH_2 - CH_2 - CH_2 - CH_3$

n-heptane
(octane number = 0)

S.No	Hydrocarbon	Octane Number
1.	Benzene	100+
2.	Isopentane	90
3.	Cyclohexane	77
4.	2-metyl pentane	71
5.	n-pentane	62

Fuels with octane rating greater than 100 are quite common nowadays and they are rated by comparison with a blend of iso-octane with tetra ethyl lead (TEL) which greatly diminishes the knocking tendency of any hydrocarbon with which it is mixed. The value of octane number in such cases is determined by extrapolation.

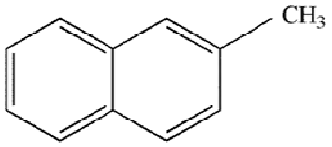
DIESEL OIL

- It is relatively a high boiling point fraction of petroleum obtained between 250 - 320°C.
- It is a mixture of hydrocarbons in terms of carbon atoms C_{15}-C_{18}
- Its calorific value is about 11,000 kcals/kg. It is used as fuel for compression ignition engine.
- Its antiknock value can be improved by doping with isoamyl nitrate.

CETANE NUMBER

In a diesel engine, the fuel is exploded not by a spark but by the application of heat and pressure. Dieselengine fuels consist oflonger chain hydrocarbons than internal combustion engine fuels. In other words, hydrocarbon molecules in a diesel fuel should be as far as possible the straight chain ones, with a minimum admixture of aromatics and side chain hydrocarbon molecules.

The suitability of a diesel fuel is determined by its cetane value which is the percentage of hexadecane in a mixture of hexadecane and 2-methyl naphthalene, which has the same ignition characteristics as the diesel fuel in question.



2-methyl naphthalene
Cetane number = 0

CH_3-$(CH_2)_{14}$-CH_3
n-hexadecane
Cetane number = 100

Thus, cetane number is defined as "the percentage of hexa decane present in a mixture of hexa decane and α–methyl naphthalene, which has the same ignition lag as the fuel under test".

The cetane number of a diesel fuel can be raised by the addition of small quantity of certain pre-ignition dopes like ethyl nitrite, isoamyl nitrite, and acetone peroxide. An oil of high octane number has a low cetane number and vice-versa.

Consequently, petroleum crude gives petrol of high octane number and diesel of low cetane number.

The cetane number decreases in the following order:

straight chain paraffins > cycloparaffins > olefins > branched paraffins > aromatics.

III. GASEOUS FUELS

NATURAL GAS

Methane is the main constituent of Natural gas and accounting for about 95% of the total volume. Other components are: Ethane, Propane, Butane, Pentane, Nitrogen, Carbon Dioxide, and traces of other gases. Very small amounts of sulphur compounds are also present. Since methane is the largest component of natural gas, generally properties of methane are used when comparing the properties of natural gas to other fuels.

Natural Gas is a high calorific value fuel requiring no storage facilities. It mixes with air readily and does not produce smoke or soot. It has no sulphur content. It is lighter than air and disperses into air easily in case of leak. A typical comparison of carbon contents in oil, coal and gas is given in the Table.

Table. Comparison of Chemical composition of various fuels			
Content	**Fuel oil**	**Coal**	**Natural gas**
Carbon	84	41.11	74
Hydrogen	12	2.76	25
Sulphur	3	0.41	-
Oxygen	1	9.89	Trace
Nitrogen	Trace	1.22	0.75
ash	Trace	38.63	-
water	Trace	5.98	-

Compressed Natural Gas (CNG)

CNG is natural gas compressed to a high pressure of about 1000 atmospheres. A steel cylinder containing 15 kg of CNG contains about 2×104 L or 20 m^3 of natural gas at 1 atmospheric pressure. It is derived from natural gas and the main constituent of CNG is methane.

The average composition of CNG is as follows:

Constituents	Percentage %
Methane	88.5
Ethane	5.5
Propane	3.7
Butane	1.8
Pentane	0.5

Properties

(i) CNG is comparatively much less pollution causing fuel as it produces less CO, ozone and hydrocarbons during combustion.

(ii) During it s combustion, no sulphur and nitrogen gases are evolved.

(iii) No carbon particles are ejected during combustion.

(iv) It is less expensive than petrol and diesel.

(v) The ignition temperature of CNG is 550

(vi) CNG is a better fuel than petrol/diesel for automobiles.

(vii) CNG requires more air for ignition.

Uses

As CNG is the cheapest, cleanest and least environmentally impacting alternative fuel. In Delhi, it is mandatory for all buses, taxis and auto to use CNG as a fuel.

Liquified Petroleum Gas (LPG)

LPG or bottled gas or refinery gas is obtained as a by-product during the cracking of heavy oils or from natural gas. LPG is dehydrated, desulphurised and traces of odorous organic sulphides (mercaptans) are added to give warning of gas leak. LPG is supplied under pressure in containers under the trade name like Indane, Bharat gas, etc. Its calorific value is about 27,800 kcal/m^3.

It consists of hydrocarbons of such volatility that they can exist as gas under atmospheric pressure, but can be readily liquefied under pressure. The main constituents of LPG are n-butane, isobutene, butylenes and propane, with little or no propylene and ethane.

The average composition of LPG is as follows:

Constituents	Percentage %
n-Butane	38.5
Iso Butane	37.0
Propane	24.5

Power Alcohol

When ethyl alcohol is used as fuel in internal combustion engine, it is called as "power alcohol". Generally ehty alcohol is used as its 5-25% mixture with petrol.

Advantages of Power Alcohol:

- Ethyl alcohol has good antiknocking property and its octane number is 90, while the octane number of petrol is about is 65. Therefore, addition of ethyl alcohol increases the octane number of petrol.

- Alcohol has property of absobing any traces of water if present in petrol.
- If specially designed engine with higher compression ratio is used, then disadvantage of lower Calorific value of ethyl alcohol can be overcome.
- Ethyl alcohol contains 'O' atoms, which helps for complete combustion of power alcohol and the polluting emissions of CO, hydrocarbon,particulates are reduced largely.
- Use of ethyl alcohol in petrol reduces our dependence on foreign countries for petrol and saves foreign considerably.
- Power alcohol is cheaper than petrol.

Disadvantages of Power Alcohol:

- Ethyl alcohol has calorific value 7000cal/gm much lower than calorific value of petrol 11500cal/gm. Use of power alcohol reduces power output upto 35%.
- Ethyl alcohol has high surface tension and its atomisation, especially at lower temperature, is difficult causing starting trouble.
- Ethyl alcohol may undergo oxidation reaction to form acetic acid, which corrodes engine parts.
- As ethyl alcoholcontains 'O' atoms, the amount of air require for complete combustion of power alochol is lesser and therefore carburettor and engine need to be modified, when only ethyl alcohol is used as fuel.

Biodiesel

A fuel derived from organic oils, such as vegetable oil, rather than petroleum. Biodiesel's use and production are increasing. It's typically used for aircraft, vehicles and as heating oil.

Vegetable oils comprise of 90–95% triglycerides with small amount of diglycerides, free fatty acids, phospholipids, etc. The viscosity of vegetable oils are higher and their molecular weights are in the range of 600 to 900, which are about 3 times higher than those of the diesel fuels.

Problems in using Vegetable Oils directly

(i) As the viscosity of vegetable oils are high, atomization is very poor and hence inefficient mixing of oil with air leads to incomplete combustion.

(ii) Oxidation and Thermal polymerization of vegetable oils cause deposit formation.

(iii) Their high velocity and consequent high flash point lead to more deposite formation.

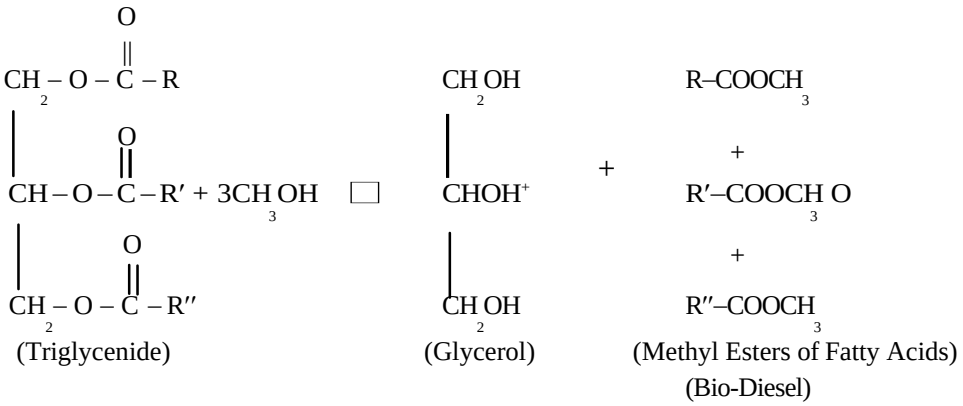
Manufacture: Trans-Esterification (or) Alcoholysis

The above problems are overcome by reducing the viscosity of the vegetable oils by the process known as **trans-esterification or alcoholysis**. Alcoholysis is nothing but displacement of alcohol from an ester by another alcohol.

It involves treatment of vegetable oil (sunflower oil, palm oil, soyabean oil, mustard oil, etc.) with excess of methanol in the presence of catalyst to give mono ethy esters of long chain fatty acid and glycerine. It is allowed to stand for some time and glycerine is separated.

Alcoholysis reaction is represented as:

$$-0.09H \times 587\ kcal / kg$$



(Triglycenide) (Glycerol) (Methyl Esters of Fatty Acids) (Bio-Diesel)

Methyl esters of fatty acids, thus formed, are called "Bio-diesel". **Bio diesel is defined as mono-alkyl esters of long chain fatty acids derived from vegetable oils or fats**.

Advantages

1. It can be produced from renewable, domestic resources.
2. Biodiesel is energy efficient (The total fossil fuel energy efficiency of biodiesel is 320% vs. 83% for petroleum diesel) (National Biodiesel Board, 1998)
3. It can be used directly in most diesel engine applications.
4. It can reduce global warming and tailpipe emissions (–41%)
5. It is nontoxic and biodegradable.
6. It is a good solvent and mayclean out fuel line and tank sediments. (Note that this may result in fuel filter clogging during initialuse)

Limitations

1. It contains approximately 8% less energy per gallon.
2. It generally has a higher cloud and pour point (will freeze at a higher temp) than conventional diesel.
3. It is not compatible with some hose and gasket materials, which may cause them to soften, degrade,and rupture.
4. It is not compatible with some metals and plastics.
5. It may increase nitrogen oxide emissions

CHAPTER 7: Combustion of Fuels

COMBUSTION OF FUELS

Fuel is a combustible substance, containing carbon as main constituent, which on proper burning gives large amount of heat, which can be used economically for domestic and industrial purposes. Eg., Wood, Charcoal, Coal, Kerosene, Petrol, Producer gas, Oilgas, LPG etc., During the process of combustion of a fuel (like coal), the atoms of carbon, hydrogen, etc. combine with oxygen with the simultaneous liberation of heat at a rapid rate.

$$\text{Fuel} + \text{Oxygen} \longrightarrow \text{Products} + \text{Heat}$$

Calorific Value

Calorific Value of a Fuel is *"the total quantity of heat liberated, when a unit mass (or volume) of the fuel is burnt completely"*.

Units of Heat:

(1) Calorie □ is the amount of heat required to raise the temperature of one gram of water through one degree centigrade (15-16 °C).

(2) Kilocalorie □ is equal to 1,000 calories. This is the unit of metric system and may be defined as "the quantity of heat required to raise the temperature of one kilogram of water through one degree centigrade. Thus, 1 kcal = 1,000 calories.

(3) British Thermal Unit (BTU) □ is defined as "the quantity of heat required to raise the temperature of one pound of water through one degree Fahrenheit (60-61°F). This is the English system unit.1 BTU = 252 cal = 0.252 kcal and 1 kcal = 3.968 BTU

(4) Centigrade Heat Unit (CHU) □ *i*s "the quantity of heat required to raise the temperature of 1 pound of water through one degree centigrade".

Thus, 1 kcal = 3.968 BTU = 2.2 CHU.

Higher or Gross Calorific Value (GCV)

It is the total amount of heat produced, when unit mass/volume of the fuel has been burnt completely and the products of combustion have been cooled to room temperature (15°C or 60°F).

It is explained that all fuels contain some hydrogen and when the calorific value of hydrogen containing fuel is determined experimentally, the hydrogen is converted into steam. If the products of combustion are condensed to the room temperature, the latent heat of condensation of steam also gets included in the measured heat which is then called GCV.

Lower or Net Calorific Value (NCV)

It is the net heat produced, when unit mass/volume of the fuel is burnt completely and the products are permitted to escape.

In actual practice of any fuel, the water vapour and moisture, etc., are not condensed and escape as such along with hot combustion gases. Hence, a lesser amount of heat is available.

$\therefore$ NCV = GCV – Latent heat of condensation of water vapour produced

= GCV – Mass of hydrogen $\times$ 9 $\times$ Latent heat of condensation of water vapour

Theoretical Calculation of Calorific Value

The calorific value of fuel can be *approximately* computed by noting the amounts of the constituents of the fuel. The higher calorific value of some of the chieg combustible constitutents of fuel are tabulated below:

Table : Calorific values of fuel constituents

Constituent	*Hydrogen*	*Carbon*	*Sulphur*
HCV (kcal/kg)	34,500	8,080	2,240

The oxygen, if present in the fuel, is assumed to be present in combined form with hydrogen, i.e., in the form of ***fixed hydrogen*** (H_2O). so, the amount of hydrogen available for combustion

= Total mass of hydrogen in fuel – Fixed hydrogen

$$= \text{Total mass of hydrogen in fuel} - \left(\frac{1}{8}\right) \text{Mass of oxygen in the fuel}$$

(ie., 8 parts of oxygen combine with one part of hydrogen to form H_2O)

Dulong's Formula for calorific value from the chemical composition of fuel is:

$$HCV = \frac{1}{100}\left[8,080\,C + 34,500\left(H - \frac{O}{8}\right) + 2240\,S\right] kcal/kg$$

where C,H,O and S are the percentages of carbon, hydrogen, oxygen and su;phur in the fuel respectively. In this formula, oxygen is ssumed to be present in combinatio with hydrogen as water, and

$$LCV = \left[HCV - \frac{9}{100} H \times 587\right] kcal/kg = [HCV - 0.09H \times 587]\ kcal/kg$$

This is based on the fact that 1 part of H by mass gives 9 parts of H_2O, and latent heat of stam is 587 kcal/kg.

PROBLEMS BASED ON CALORIFIC VALUE

Problem 1: Calculate the gross and net calorific values of coal having the following compositions, carbon = 85%, hydrogen = 8%, sulphur = 1%, nitrogen = 2%, ash = 4%, latent heat of steam = 587 cal/gm.

Solution:

(i) Gross Calorific Value (GCV)

$$= \frac{1}{100}\left[8080 \times \%\,c + 34500\left(\%\,H - \frac{\%\,O}{8}\right) + 2240 \times \%\,s\right] kcal/kg$$

$$= \frac{1}{100}\left[8080 \times 85 \times 34500\left(8 - \frac{0}{8}\right) + 2240 \times 1\right] kcal/kg$$

$$= \frac{1}{100}[6,86,800 + 2,76,000 + 2240] kcal/kg$$

$$= \frac{1}{100}[9, 65, 040] \text{kcal / kg}$$

$$= 9650.4 \text{ kcal/kg.}$$

(ii) Net Calorific Value (NCV)

$$= \text{GCV} - \frac{9}{100}\text{H} \times 587 \text{ kcal / kg}$$

$$= 9650.4 - \frac{9}{100} \times 8 \times 587 \text{ kcal / kg}$$

$$= 9650.4 - 422.64$$

$$= 9227.76 \text{ kcal / kg.}$$

Problem 2: Calculate the net and gross calorific value of a coal sample having following composition. C = 82%, H = 8%, 0 = 5%, N = 1.4% and ash = 3.6%.

Solution:

(i) GCV

$$= \frac{1}{100}\left[8080 \times \%\, c + 34500\left(\%\, \text{H} - \frac{\%0}{8}\right) + 2240 \times \%s\right] \text{kcal / kg}$$

$$= \frac{1}{100}\left[8080 \times 82 \times 34500\left(8 - \frac{5}{8}\right) + 0\right] \text{kcal / kg}$$

$$= \frac{1}{100}[662560 + 254437.3]$$

$$= 9169.8 \text{ kcal / kg .}$$

(ii) NCV

$$= \text{GCV} - \frac{9}{100}\text{H} \times 587 \text{ kcal / kg}$$

$$= 9169.98 - \frac{9}{100} \times 8 \times 587$$

$$= 8747.34 \text{ kcal/kg.}$$

Problem 3: Calculate the gross and net calorific valve of a fuel having following composition 82%, C, 8% H, 5% O, 2.5% S, 1.4% N and 2.1% ash.

Solution: We know that,

$$GCV = \frac{1}{100}\left[8080\ c + 34500\ (H - 0/8) + 22405\right] kcal/kg$$

$$= \frac{1}{100}\left[8080 \times 82 \times 34500\ (8 - 5/8) + 2240 \times 2.5\right]$$

$$= 9225.97 \text{ kcal/kg}$$

$$NCV = GCV - 0.09\ H \times 587\ kcal/kg$$

$$NCV = 9225.97 - 0.09 \times 8 \times 587$$

$$= 8803.3 \text{ kcal/kg.}$$

IGNITION TEMPERATURE (IT)

It is defined as ***"the lowest temperature to which the fuel must be heated, so that it starts burning smoothly".***

The ignition temperature of coal is about 300°C. In the case of liquid fuels, the ignition temperature is called the flash point, which ranges from 200–400°C. For gaseous fuels, the ignition temperature is in the order of 800°C.

Spontaneous Ignition Temperature (SIT)

When a gaseous mixture of fuel and oxidant is maintained at ambient temperature, reaction rates are extremely slow. Increasing the mixture of temperature, the reaction rate suddenly increases, giving rise to rapid combustion reactions. This condition is referred to as spontaneous ignition and the minimum temperature at which rapid combustion reactions are initiated is called **the *spontaneous ignition temperature [T_{sit}]*.**

The factors influencing the spontaneous ignition temperature of given mixture are the balance between heat release and heat loss, as well as the supply of reactants.

For liquid fuels, this parameter is determined using standardized tests, where liquid fuel is dropped into an open air container heated to a known temperature.

The spontaneous ignition temperature is defined as the lowest temperature at which visible or audible evidence of combustion is observed.

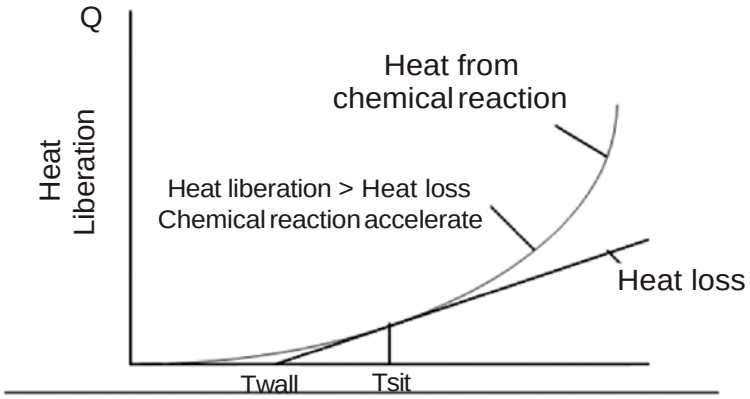


Figure 7.1 : Spontaneous Ignition Temperature

Explosive Range (or) Limits of Inflammability

Most of the gaseous fuels have two percentage limits called upper limit and lower limit. Those limits represent percentage by volume of fuel present in fuel-air mixture. The range covered by these limits is termed as explosive range of the fuel.

For continuous burning the amount of fuel present in the fuel-air mixture should not go below the lower limit or above the upper limit.

For example, the explosive range of petrol is 2-4.5. This means that when the concentration of petrol vapour in petrol-air mixture is between 2 and 4.5 by volume, the mixture will burn on ignition. When the concentration of petrol vapour in petrol-air mixture is below 2% (lower limit) or above 4.5% (upper limit) by volume, the mixture will not burn on ignition.

FLUE GAS ANALYSIS (ORSAT METHOD)

The mixture of gases like SO_2, CO_2, O_2, CO etc. coming out from the combustion chamber is called flue gas.

Analysis:

The flue gas analysis is carried out by using Orsat's apparatus. The analysis of flue gas generally deals with the determination of CO_2, O_2 and CO by absorbing them in the respective solution of KOH, alkaline pyrogallol and ammonium cuprous chloride.

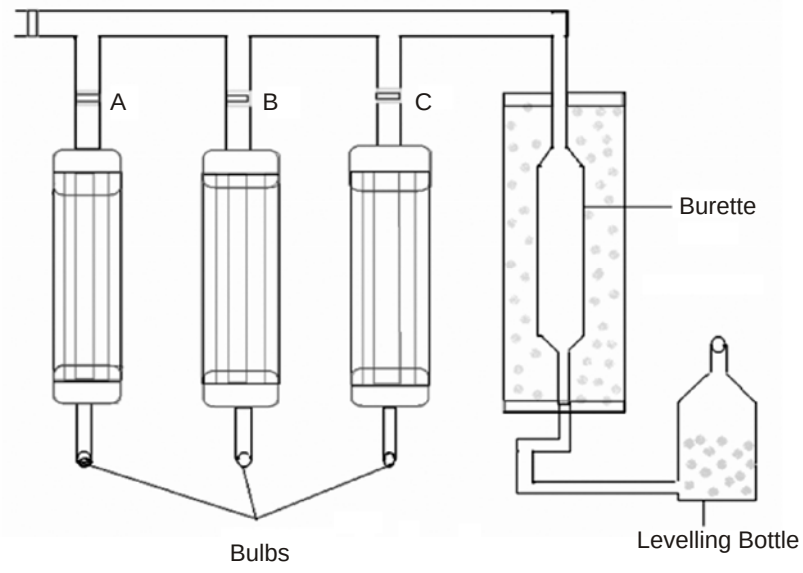


Figure 7.2 : Orsat's Apparatus

Description of Orsat's Apparatus:

Orsat's apparatus consists of a horizontal tube having 3 way stopcock at one end and a water jacketed measuring burette at the other end. The horizontal tube is connected to three different absorption bulbs for the absorption of CO_2, O_2 and CO respectively. The lower end of the burette is connected to the leveling bottle by means of rubber tube.

The level of water in the leveling bottle (water reservoir) can be raised or lowered by raising or lowering the water reservoir. By changing the level of water, the flue gas can be moved into various parts of the apparatus during analysis.

It is essential to follow the order of absorbing the gases- CO_2 first; O_2 second and CO last. This is because the absorbent used for O_2 (ie., alkaline pyrogallol) can also absorb some amount of CO_2 and the percentage of CO_2 left would be less.

Importance of Flue Gas Analysis:

(i) The analysis gives the idea of whether a combustion process is complete or not.

(ii) The C and H present in a fuel undergo combustion forming CO_2 and H_2O respectively. Any N present is not at all involved in the combustion. ie., the products of combustion are CO_2, H_2O and N_2.

(iii) If analysis of a flue gas indicates the presence of CO; it is suggestive of incomplete combustion (Wastage of heat is inferred)

(iv) If there is considerable amount of oxygen, it shows that there is excess supply of O_2 although combustion would have been complete.

(a) Absorption of CO_2

Flue gas is passed into the bulb A via its stopcock by raising the water reservoir. CO_2 present in the flue gas is absorbed by KOH (usually 250 g KOH in 500 mL distilled water). The gas is again sent to the burette and then again sent to bulb A. This process is repeated several times, by raising or lowering of water reservoir so as to ensure complete absorption of CO_2 in KOH. Now, the stopcock of bulb A is closed. The volume of residual gases in the burette is taken by equalizing the water level both in the burette and in the water reservoir. The difference between original volume and the volume of the gases after CO_2 absorption gives the volume of CO_2 absorbed.

(b) Absorption of O_2

Stopcock of bulb A is closed and bulb B is opened. Oxygen present in the flue gas is absorbed by alkaline pyrogallol (25 g pyrogallol +200g KOH in 500 mL distilled water). The absorption process is same as in bulb A.

(c) Absorption of CO

Now the stopcock of bulb B is closed and stopcock of bulb C is opened. Carbon monoxide present in the flue gas is absorbed by ammoniacal cuprous chloride (100 g Cu_2Cl_2 + 125 mL liquid NH_3 + 375 mL water). Here also absorption process is same as in bulb A.

Since the total volume of the gas taken for analysis is 100 mL, the volume of the constituents are their percentage.

The residual gas after the above three determinations is taken as nitrogen.

Further, as the content of CO in the flue gas would be very low, it should be measured quite carefully.

❑ ❑ ❑

UNIT II CORROSION AND CORROSION CONTROL

Chemical corrosion – Pilling-Bedworth rule – Electrochemical corrosion – Different types – Galvanic corrosion – Differential aeration corrosion - Factors influencing corrosion – Corrosion control – Sacrificial anode and impressed cathodic current methods – Corrosion inhibitors – Protective coatings – Paints – Constituents and functions – Metallic coatings – Electroplating (Au) and Electroless (Ni) plating.

INTRODUCTION:

Corrosion is an undesirable process. Due to corrosion there is limitation of progress in many areas. The cost of replacement of materials and equipments lost through corrosion is unlimited.

Metals and alloys are used as fabrication or construction materials in engineering. If the metals or alloy structures are not properly maintained, they deteriorate slowly by the action of atmospheric gases, moisture and other chemicals. This phenomenon of destruction of metals and alloys is known as corrosion.

Corrosion of metals is defined as the spontaneous destruction of metals in the course of their chemical, electrochemical or biochemical interactions with the environment. Thus, it is exactly the reverse of extraction of metals from ores.

Example: Rusting of iron

A layer of reddish scale and powder of oxide (Fe_3O_4) is formed on the surface of iron metal.

A green film of basic carbonate [$CuCO_3 + Cu(OH)_2$] is formed on the surface of copper, when it is exposed to moist-air containing carbon dioxide.

CONSEQUENCES (EFFECTS) OF CORROSION:

The economic and social consequences of corrosion include

i) Due to formation of corrosion product over the machinery, the efficiency of the machine gets failure leads to plant shut down.

ii) The products contamination or loss of products due to corrosion.

iii) The corroded equipment must be replaced

iv) Preventive maintenance like metallic coating or organic coating is required.

v) Corrosion releases the toxic products.

vi) Health (eg., from pollution due to a corrosion product or due to the escaping chemical from a corroded equipment).

CAUSES OF CORROSION:

In nature, metals occur in two different forms.

1) Native State (2) Combined State

Native State: The metals exist as such in the earth crust then the metals are present in a native state. Native state means free or uncombined state. These metals are non-reactive in nature. They are noble metals which have very good corrosion resistance. Example: Au, Pt, Ag, etc.,

Combined State: Except noble metals, all other metals are highly reactive in nature which undergoes reaction with their environment to form stable compounds called ores and minerals. This is the combined state of metals. Example: Fe_2O_3, ZnO, PbS, $CaCO_3$, etc.,

Metallic Corrosion: The metals are extracted from their metallic compounds (ores). During the extraction, ores are reduced to their metallic states by applying energy in the form of various processes. In the pure metallic state, the metals are unstable as they are considered in excited state (higher energy state). Therefore as soon as the metals are extracted from their ores, the reverse process begins and form metallic compounds, which are thermodynamically stable (lower energy state). Hence, when metals are used in various forms, they are exposed to environment, the exposed metal surface begin to decay (conversion to more stable compound). This is the basic reason for metallic corrosion.

$$\text{Metal} \underset{\text{Metallurgy-Reduction}}{\overset{\text{Corrosion-Oxidation}}{\rightleftharpoons}} \text{Metallic Compound} + \text{Energy}$$

Although corroded metal is thermodynamically more stable than pure metal but due to corrosion, useful properties of a metal like malleability, ductility, hardness, luster and electrical conductivity are lost.

CLASSIFICATION OR THEORIES OF CORROSION

Based on the environment, corrosion is classified into

(i) Dry or Chemical Corrosion (ii) Wet or Electrochemical Corrosion

DRY or CHEMICAL CORROSION:

This type of corrosion is due to the direct chemical attack of metal surfaces by the atmospheric gases such as oxygen, halogen, hydrogen sulphide, sulphur dioxide, nitrogen or anhydrous inorganic liquid, etc. The chemical corrosion is defined as the direct chemical attack of metals by the atmospheric gases present in the environment.

Example: (i) Silver materials undergo chemical corrosion by Atmospheric H_2S gas .

(ii) Iron metal undergo chemical corrosion by HCl gas.

TYPES OF DRY or CHEMICAL CORROSION:

1. Corrosion by Oxygen or Oxidation corrosion
2. Corrosion by Hydrogen
3. Liquid Metal Corrosion

CORROSION BY OXYGEN or OXIDATION CORROSION:

Oxidation Corrosion is brought about by the direct attack of oxygen at low or high temperature on metal surfaces in the absence of moisture. Alkali metals (Li, Na, K etc.,) and alkaline earth metals (Mg, Ca, Sn, etc.,) are rapidly oxidized at low temperature. At high temperature, almost all metals (except Ag, Au and Pt) are oxidized. The reactions of oxidation corrosion are as follows:

Mechanism:

1) Oxidation takes place at the surface of the metal forming metal ions M2+

$$M \rightarrow M^{2+} + 2e^-$$

2) Oxygen is converted to oxide ion (O^{2-}) due to the transfer of electrons from metal.

$$n/2\ O_2 + 2n\ e^- \rightarrow n\ O^{2-}$$

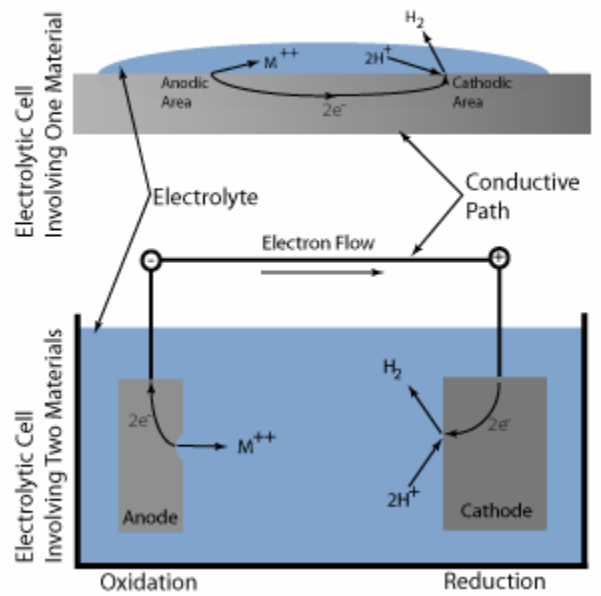
3) The overall reaction is of oxide ion reacts with the metal ions to form metal oxide film.

$$2\ M + n/2\ O_2 \rightarrow 2\ M^{n+} + nO^{2-}$$

The Nature of the Oxide formed plays an important part in oxidation corrosion process.

Metal + Oxygen → Metal oxide (corrosion product)

When oxidation starts, a thin layer of oxide is formed on the metal surface and the nature of this film decides the further action. If the film is



(i) Stable layer:

A Stable layer is fine grained in structure and can get adhered tightly to the parent metal surface. Hence, such layer can be of impervious nature (ie., which cuts-off penetration of attaching oxygen to the underlying metal). Such a film behaves as protective coating in nature, thereby shielding the metal surface. The oxide films on Al, Sn, Pb, Cu, Pt, etc., are stable, tightly adhering and impervious in nature.

(ii) Unstable oxide layer:

This is formed on the surface of noble metals such as Ag, Au, Pt. As the metallic state is more stable than oxide, it decomposes back into the metal and oxygen. Hence, oxidation corrosion is not possible with noble metals.

(iii) Volatile oxide layer:

The oxide layer film volatilizes as soon as it is formed. Hence, always a fresh metal surface is available for further attack. This causes continuous corrosion. MoO_3 is volatile in nature.

(iv) Porous layer:

The layer having pores or cracks. In such a case, the atmospheric oxygen have access to the underlying surface of metal, through the pores or cracks of the layer, thereby the corrosion continues unobstructed, till the entire metal is completely converted into its oxide.

Pilling-Bedworth rule: According to it "an oxide is protective or non-porous, if the volume of the oxide is atleast as great as the volume of the metal from which it is formed". On the other hand, "if the volume of the oxide is less than the volume of metal, the oxide layer is porous (or non-continuous) and hence, non-protective, because it cannot prevent the access of oxygen to the fresh metal surface below".

Thus, alkali and alkaline earth metals (like Li, K, Na, Mg) form oxides of volume less than the volume of metals. Consequently, the oxide layer faces stress and strains, thereby developing cracks and pores in its structure. Porous oxide scale permits free access of oxygen to the underlying metal surface (through cracks and pores) for fresh action and thus, corrosion continues non-stop.

Metals like Aluminium forms oxide, whose volume is greater than the volume of metal. Consequently, an extremely tightly-adhering non-porous layer is formed. Due to the absence of any pores or cracks in the oxide film, the rate of oxidation rapidly decreases to zero.

Corrosion by other gases (by hydrogen):

1) Hydrogen Embrittlement:

Loss in ductility of a material in the presence of hydrogen is known as hydrogen embrittlement .

Mechanism:

This type of corrosion occurs when a metal is exposed to hydrogen environment. Iron liberates atomic hydrogen with hydrogen sulphide in the following way.

$$Fe + H_2S \rightarrow FeS + 2H$$

Hydrogen diffuses into the metal matrix in this atomic form and gets collected in the voids present inside the metal. Further, diffusion of atomic hydrogen makes them combine with each other and forms hydrogen gas.

$$H + H \rightarrow H_2\uparrow$$

Collection of these gases in the voids develops very high pressure, causing cracking or blistering of metal.

2) Decarburisation:

The presence of carbon in steel gives sufficient strength to it. But when steel is exposed to hydrogen environment at high temperature, atomic hydrogen is formed.

$$H_2 \xrightarrow{Heat} 2H$$

Atomic hydrogen reacts with the carbon of the steel and produces methane gas.

$$C + 4H \rightarrow CH_4$$

Hence, the carbon content in steel is decreases. The process of decrease in carbon content in steel is known as decarburization.

Collection of methane gas in the voids of steel develops high pressure, which causes cracking. Thus, steel loses its strength.

3) Liquid metal corrosion:

This is due to chemical action of flowing liquid metal at high temperatures on solid metal or alloy. Such corrosion occur in devices used for nuclear power. The corrosion reaction involves either: (i) dissolution of a solid metal by a liquid metal or (ii) internal penetration of the liquid metal into the solid metal. Both these modes of corrosion cause weakening of the solid metal.

WET OR ELECTROCHEMICAL CORROSION

Electrochemical corrosion involves:

i) The formation of anodic and cathodic areas or parts in contact with each other

ii) Presence of a conducting medium

iii) Corrosion of anodic areas only and

iv) Formation of corrosion product somewhere between anodic and cathodic areas.This involves flow of electron-current between the anodic and cathodic areas.

At anodic area oxidation reaction takes place (liberation of free electron), so anodic metal is destroyed by either dissolving or assuming combined state (such as oxide, etc.). Hence corrosion always occurs at anodic areas.

M (metal)	→	$M^{n+} + n e^-$
M^{n+} (metal ion)	→	Dissolves in solution
	→	forms compounds such as oxide

At cathodic area, reduction reaction takes place (gain of electrons), usually cathode reactions do not affect the cathode, since most metals cannot be further reduced. So at cathodic part, dissolved constituents in the conducting medium accepts the electrons to form some ions like OH- and O2-.

Cathodic reaction consumes electrons with either by

(a) evolution of hydrogen or

(b) absorption of oxygen, depending on the nature of the corrosive environment

Hydorgen Evolution Type:

All metals above hydrogen in the electrochemical series have a tendency to get dissolved in acidic solution with simultaneous evolution of hydrogen.

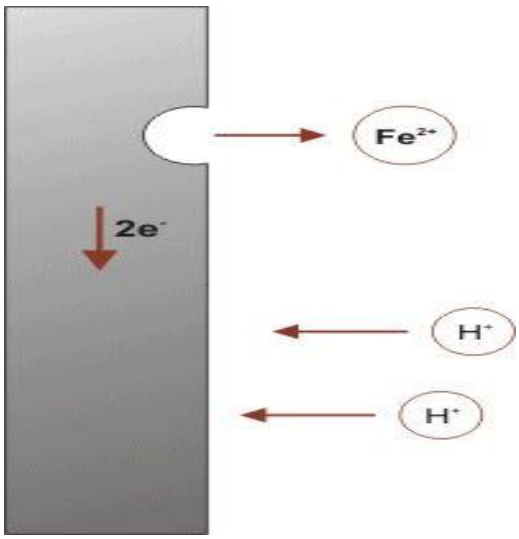
It occurs in acidic environment. Consider the example of iron

At anode: $Fe \rightarrow Fe^{2+} + 2e^-$

These electrons flow through the metal, from anode to cathode, where H+ ions of acidic solution are eliminated as hydrogen gas.

At cathode: $2H^+ + 2e^- \rightarrow H_2\uparrow$

The overall reaction is: $Fe + 2H^+ \rightarrow Fe^{2+} + H_2$



Oxygen Absorption Type:

Rusting of iron in neutral aqueous solution of electrolytes (like NaCl solution) in the presence of atmospheric oxygen is a common example of this type of corrosion. The surface of iron is usually coated with a thin film of iron oxide. However, if this iron oxide film develops some cracks, anodic areas are created on the surface; while the well metal parts acts as cathodes.

At Anode: Metal dissolves as ferrous ions with liberation of electrons.

$$Fe \rightarrow Fe^{2+} + 2e^-$$

At Cathode: The liberated electrons are intercepted by the dissolved oxygen.

$$\tfrac{1}{2} O_2 + H_2O + 2e^- \rightarrow 2OH^-$$

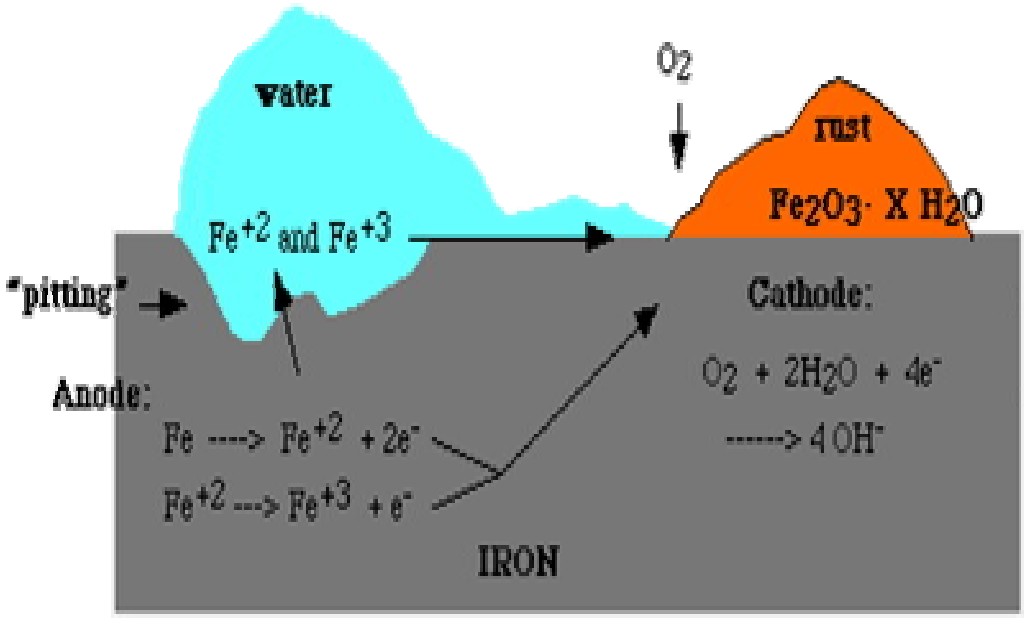
The Fe2+ ions and OH- ions diffuse and when they meet, ferrous hydroxide is precipitated.

$$Fe^{2+} + 2OH^- \rightarrow Fe(OH)_2$$

(i) If enough oxygen is present, ferrous hydroxide is easily oxidized to ferric hydroxide.

$$4Fe(OH)_2 + O_2 + 2H_2O \rightarrow 4Fe(OH)_3$$ (Yellow rust $Fe_2O_3.H_2O$)

(ii) If the supply of oxygen is limited, the corrosion product may be even black anhydrous magnetite, Fe_3O_4.



Difference between (dry) chemical and (wet) electrochemical corrosion:

Sl. No.	Chemical Corrosion	Electrochemical Corrosion
1.	It occurs in dry condition.	It occurs in the presence of moisture or electrolyte.
2.	It is due to the direct chemical attack of the metal by the environment.	It is due to the formation of a large number of anodic and cathodic areas.
3.	Even a homogeneous metal surface gets corroded.	Heterogeneous (bimetallic) surface alone gets corroded.
4.	Corrosion products accumulate at the place of corrosion	Corrosion occurs at the anode while the products are formed elsewhere.
5.	It is a self controlled process.	It is a continuous process.
6.	It adopts adsorption mechanism.	It follows electrochemical reaction.
7.	Formation of mild scale on iron surface is an example.	Rusting of iron in moist atmosphere is an example.

TYPES OF ELECTROCHEMICAL CORROSION

The electrochemical corrosion is classified into the following two types:

(i) Galvanic (or Bimetallic) Corrosion

(ii) Differential aeration or concentration cell corrosion.

Galvanic Corrosion:

When two dissimilar metals (eg., zinc and copper) are electrically connected and exposed to an electrolyte, the metal higher in electrochemical series undergoes corrosion. In this process, the more active metal (with more negative electrode potential) acts as a anode while the less active metal (with less negative electrode potential) acts as cathode.

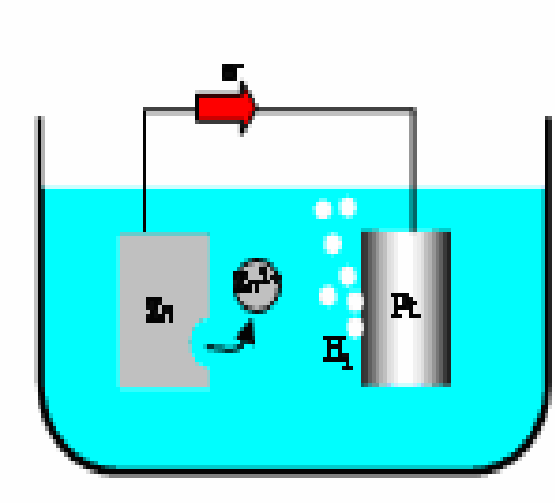
In the above example, zinc (higher in electrochemical series) forms the anode and is attacked and gets dissolved; whereas copper (lower in electrochemical series or more noble)acts as cathode.

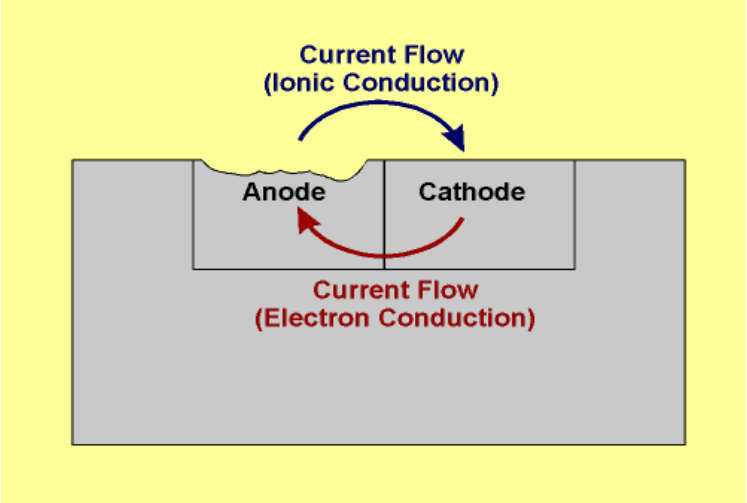
Mechanism: In acidic solution, the corrosion occurs by the hydrogen evolution process; while in neutral or slightly alkaline solution, oxygen absorption occurs. The electron-current flows from the anode metal, zinc to the cathode metal, copper.

$$Zn \longrightarrow Zn^{2+} + 2e^{-} \text{ (Oxidation)}$$

Thus it is evident that the corrosion occurs at the anode metal; while the cathodic part is protected from the attack.

Example: (i) Steel screws in a brass marine hardware (ii)Lead-antimony solder around copper wise; (iii) a steel propeller shaft in bronze bearing (iv Steel pipe connected to copper plumbing.





Concentration Cell Corrosion:

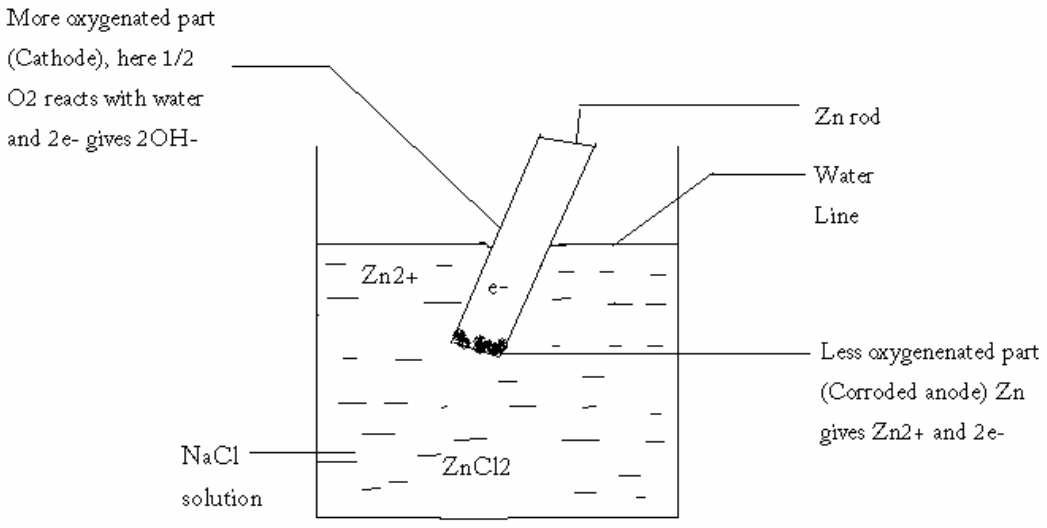
It is due to electrochemical attack on the metal surface, exposed to an electrolyte of varying concentrations or of varying aeration.

It occurs when one part of metal is exposed to a different air concentration from the other part. This causes a difference in potential between differently aerated areas. It has been found experimentally that poor-oxygenated parts are anodic.

Examples: i) The metal part immersed in water or in a conducting liquid is called water line corrosion.

ii) The metal part partially buried in soil.

Explanation: If a metal is partially immersed in a conducting solution the metal part above the solution is more aerated and becomes cathodic. The metal part inside the solution is less aerated and thus becomes anodic and suffers corrosion.



At anode: Corrosion occurs (less aerated) $M \longrightarrow M^{2+} + 2e^-$

At cathode: OH- ions are produced (more aerated) $\frac{1}{2} O_2 + H_2O + 2e^- \longrightarrow 2OH^-$

Examples for this type of corrosion are

1) Pitting or localized corrosion
2) Crevice corrosion
3) Pipeline corrosion
4) Corrosion on wire fence

Pitting Corrosion:

Pitting is a localized attack, which results in the formation of a hole around which the metal is relatively unattacked.

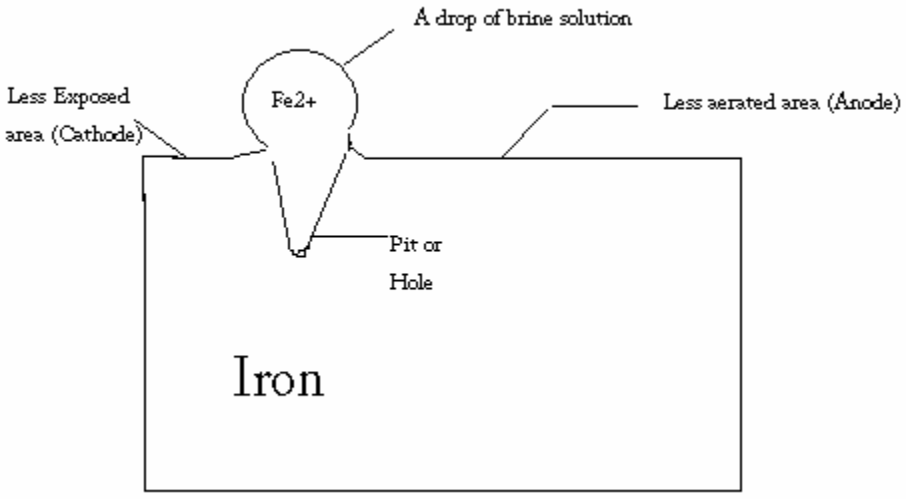
The mechanism of this corrosion involves setting up of differential aeration or concentration cell.

Metal area covered by a drop of water, dust, sand, scale etc. is the aeration or concentration cell.

Pitting corrosion is explained by considering a drop of water or brine solution (aqueous solution of NaCl) on a metal surface, (especially iron).

The area covered by the drop of salt solution as less oxygen and acts as anode. This area suffers corrosion, the uncovered area acts as cathode due to high oxygen content.

It has been found that the rate of corrosion will be more when the area of cathode is larger and the area of the anode is smaller. Hence there is more material around the small anodic area results in the formation hole or pit.



At anode: Fe is oxidized to Fe2+ and releases electrons.

$$Fe \longrightarrow Fe^{2+} + 2e^-$$

At cathode: Oxygen is converted to hydroxide ion

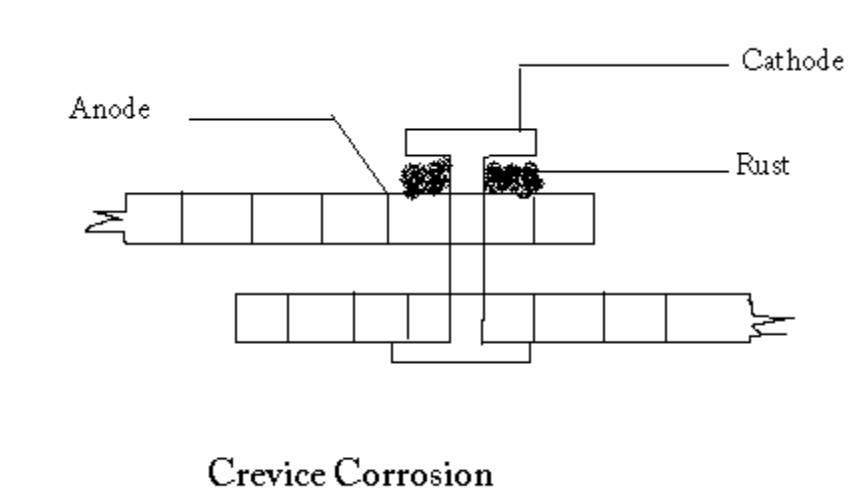
$$\tfrac{1}{2} O_2 + H_2O + 2e^- \longrightarrow 2OH^-$$

The net reaction is $Fe + 2OH^- \longrightarrow Fe(OH)_2$

The above mechanisms can be confirmed by using ferroxyl indicator (a mixture containing phenolphthalein and potassium ferricyanide). Since OH- ions are formed at the cathode, this area imparts pink colour with phenolphthalein indicator. At the anode, iron is oxidized to Fe2+ which combines with ferricyanide and shows blue colour.

Crevice corrosion:

If a crevice (a crack forming a narrow opening) between metallic and non-metallic material is in contact with a liquid, the crevice becomes anodic region and undergoes corrosion. Hence, oxygen supply to the crevice is less. The exposed area has high oxygen supply and acts as cathode.



Crevice Corrosion

Bolts, nuts, rivets, joints are examples for this type of corrosion.

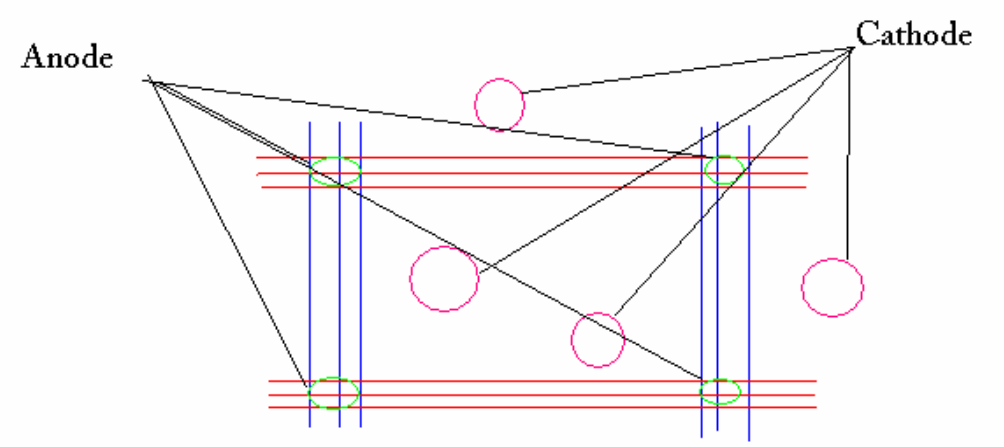
Pipeline corrosion:

Buried pipelines or cables passing from one type of soil (clay less aerated) to another soil (sand more aerated) may get corroded due to differential aeration.

Corrosion in wire fence:

A wire fence is one in which the areas where the wires cross (anodic) are less aerated than the rest of the fence (cathodic). Hence corrosion takes place at the wire crossing.

Corrosion occurring under metal washers and lead pipeline passing through clay to cinders(ash) are other examples.



Corrosion in wire fence

FACTORS INFLUENCING CORROSION

There are two factors that influence the rate of corrosion. Hence a knowledge of these factors and the mechanism with which they affect the corrosion rate is essential because the rate of corrosion is different in different atmosphere.

1. Nature of the metal
2. Nature of the corroding environment

Nature of the metal:

a) **Physical state:** The rate of corrosion is influenced by physical state of the metal (such as grain size, orientation of crystals, stress, etc). The smaller the grain size of the metal or alloy, the greater will be its solubility and hence greater will be its corrosion. Moreover, areas under stress, even in a pure metal, tend to be anodic and corrosion takes place at these areas.

b) **Purity of metal:** Impurities in a metal cause heterogeneity and form minute/tiny electrochemical cells (at the exposed parts), and the anodic parts get corroded. The cent percent pure metal will not undergo any type of corrosion. For example, the rate of corrosion of aluminium in hydrochloric acid with increase in the percentage impurity is noted.

% purity of aluminium	99.99	99.97	99.2
Relative rate of corrosion	1	1000	30000

c) **Over voltage:** The over voltage of a metal in a corrosive environment is inversely proportional to corrosion rate. For example, the over voltage of hydrogen is 0.7 v when zinc metal is placed in 1 M sulphuric acid and the rate of corrosion is low. When we add small amount of copper sulphate to dilute sulphuric acid, the hydrogen over voltage is reduced to 0.33 V. This results in the increased rate of corrosion of zinc metal.

d) **Nature of surface film:** In aerated atmosphere, practically all metals get covered with a thin surface film (thickness=a few angstroms) of metal oxide. The ratio of the volumes of the metal oxide to the metal is known as a specific volume ratio. Greater the specific volume ratio, lesser is the oxidation corrosion rate. The specific volume ratios of Ni, Cr and W are 1.6, 2.0 and 3.6 respectively. Consequently the rate of oxidation of tungsten is least, even at elevated temperatures..

e) **Relative areas of the anodic and cathodic parts:** When two dissimilar metals or alloys are in contact, the corrosion of the anodic part is directly proportional to the ratio of areas of the cathodic part and the anodic part.

Corrosion is more rapid and severe, and highly localized, if the anodic area is small (eg., a small steel pipe fitted in a large copper tank), because the current density at a smaller anodic area is much greater and the demand for electrons can be met by smaller anodic areas only by undergoing corrosion more brisky.

f) Position in galvanic series:

g) Passive character of metal:

h) Solubility of corrosion products:

i) Volatility of corrosion products:

Nature of the Corroding Environment:

a) Temperature: The rate of corrosion is directly proportional to temperature ie., rise in temperature increases the rate of corrosion. This is because the rate of diffusion of ions increases with rise in temperature.

b) Humidity of air: The rate of corrosion will be more when the relative humidity of the environment is high. The moisture acts as a solvent for oxygen, carbon dioxide, sulphur dioxide etc. in the air to produce the electrolyte which is required for setting up a corrosion cell.

c) Presence of impurities in atmosphere: Atmosphere in industrial areas contains corrosive gases like CO_2, H_2S, SO_2 and fumes of HCl, H_2SO_4 etc. In presence of these gases, the acidity of the liquid adjacent to the metal surfaces increases and its electrical conductivity also increases, thereby the rate of corrosion increases.

d) Presence of suspended particles in atmosphere: In case of atmospheric corrosion: (i) if the suspended particles are chemically active in nature (like NaCl, Ammonium sulphate), they absorb moisture and act as strong electrolytes, thereby causing enhanced corrosion; (ii) if the suspended particles are chemically inactive in nature (eg., charcoal), they absorb both sulphur gases and moisture and slowly enhance corrosion rate.

e) Influence of pH: Generally acidic media (ie., pH<7) are more corrosive than alkaline and neutral media. However, amphoteric metals (like Al, Zd, Pb, etc.) dissolve in alkaline solutions as complex ions. The corrosion rate of iron in oxygen-free water is slow, until the pH is below 5. The corresponding corrosion rate in presence of oxygen is much higher. Consequently corrosion of metals, readily attacked by acid, can be reduced by

increasing the pH of the attacking environment, eg., Zn (which is rapidly corroded, even in weakly acidic solutions such as carbonic acid suffers minimum corrosion at pH=11.

f) Nature of ions present:
g) Conductance of the corroding medium:
h) Formation of oxygen concentration cell:
i) Flow velocity of process stream:
j) Polarization of electrodes:

CORROSION CONTROL (PROTECTION AGAINST CORROSION)

As the corrosion process is very harmful and losses incurred are tremendous, it becomes necessary to minimize or control corrosion of metals. Corrosion can be stopped completely only under ideal conditions. But the attainment of ideal conditions is not possible. However, it is possible only to minimize corrosion considerably. Since the types of corrosion are so numerous and the conditions under which corrosion occurs are so different, diverse methods are used to control corrosion. As the corrosion is a reaction between the metal or alloy and the environment, any method of corrosion control must be aimed at either modifying the metal or the environment.

a. Choice of metals and alloys:

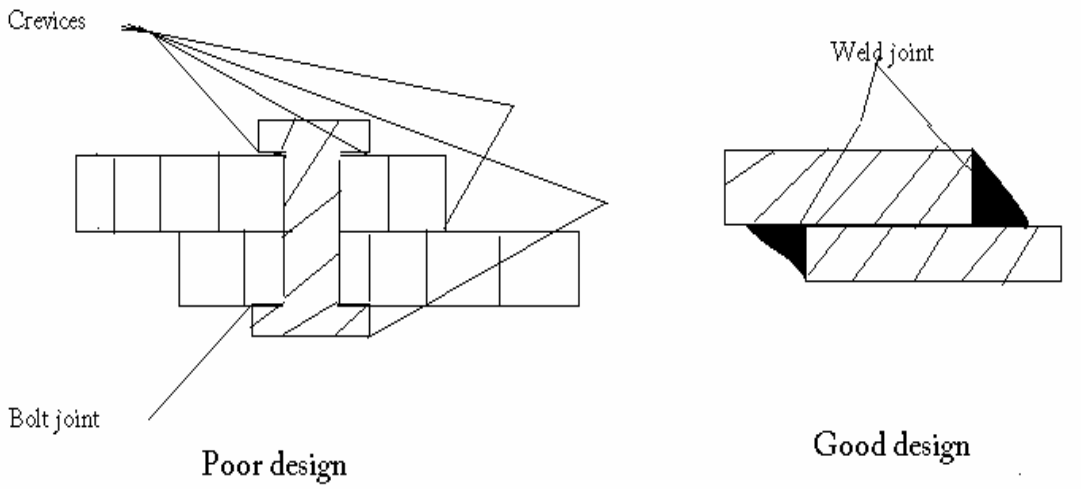
1. The first choice is to use noble metals such as gold and platinum. They are most resistant to corrosion. As they are precious, they cannot be used for general purposes.
2. The next choice is to use purest possible metal. But in many cases, it is not possible to produce a metal of high chemical purity. Hence, even a trace amount of impurity leads to corrosion.
3. Thus, the next choice is the use of corrosion resistant alloys. Several corrosion resistant alloys have been developed for specific purposes and environment. For example, a) Stainless steel containing chromium produce an exceptionally coherent oxide film which protects the steel from further attack. (b) Cupro-nickel (70% Cu + 30%Ni) alloys are now used for condenser tubes and for bubble trays used in fractionating column in oil refineries. (c) Highly stressed Nimonic alloys (Ni-Cr-Mo alloys) used in gas turbines are very resistant to hot gases.

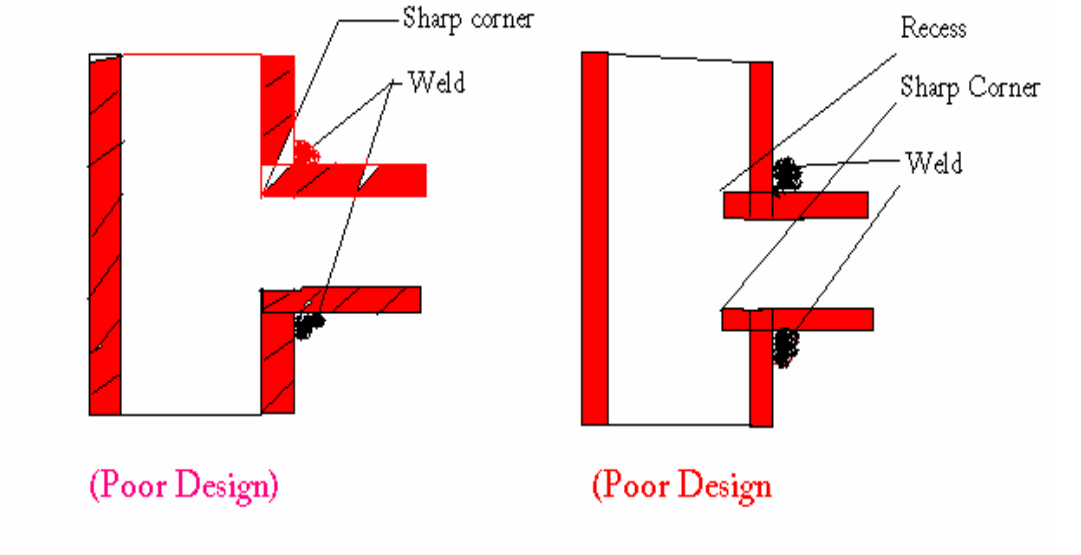
b. Proper Designing:

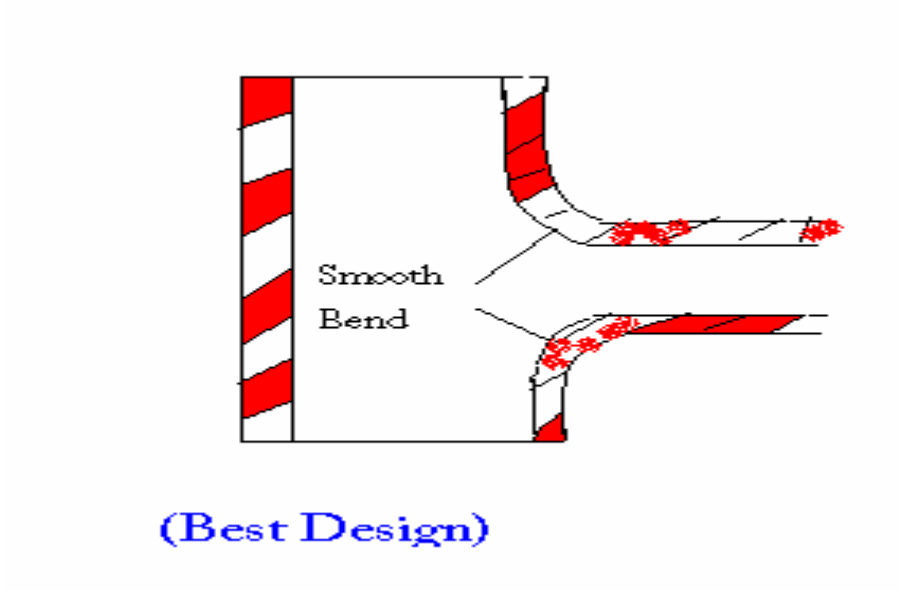
Proper geometrical design plays a vital role in the control of corrosion of equipments and structures. The general guidelines of the design of materials and components to control corrosion are the following:

a. Use always simple design and structure
b. The design must avoid more complicated shapes having more angles, edges, corners etc.
c. Avoid the contact of dissimilar metals as they may lead to galvanic type corrosion. To overcome this, insulation can be used.

d. When two dissimilar metals are to be in contact, the anodic area must be as large as possible and the cathodic area should be as small as possible.
e. As far as possible, crevices (gap or crack) should be avoided between adjacent parts of a structure.
f. Bolts and rivets should be replaced by proper welding
g. Metal washers should be replaced by rubber or plastic washers as they do not adsorb water. They also act as insulation.
h. Corrosion in pipelines can be prevented by using smooth bends.
i. Heat treatment like annealing minimizes the stress corrosion.
j. A good design of water storage container is the one from which water can be drained and cleaned easily. Such a design avoids accumulation of dirt etc.







(Best Design)

CATHODIC PROTECTION:

The reduction or prevention of corrosion by making metallic structure as cathode in the electrolytic cell is called cathodic protection. Since there will not be any anodic area on the metal, corrosion does not occur. There are two methods of applying cathodic protection to metallic structures.

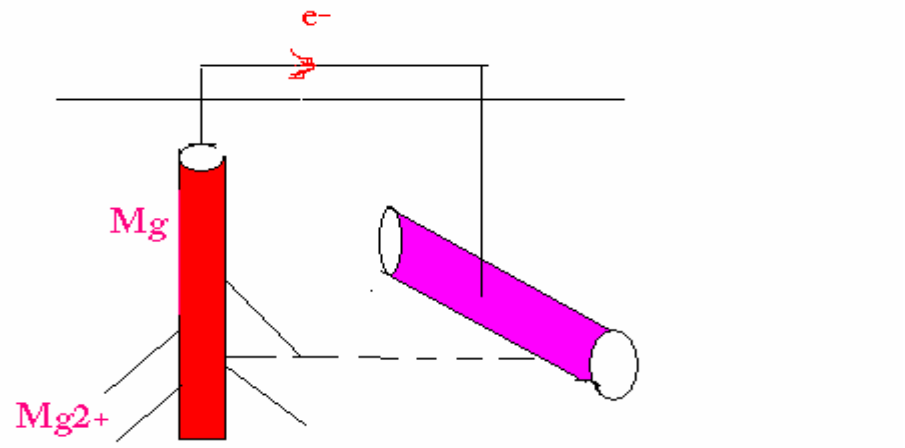
a) Sacrificial anodic protection (galvanic protection)
b) Impressed current cathodic protection

SACRIFICIAL ANODIC PROTECTION METHOD

In this method, the metallic structure to be protected is made cathode by connection it with more active metal (anodic metal). Hence, all the corrosion will concentrate only on the active metal. The parent structure is thus protected. The more active metal so employed is called sacrificial anode. The corroded sacrificial anode block is replaced by a fresh one. Metals commonly employed as sacrificial anodes are magnesium, zinc, aluminium and their alloys. Magnesium has the most negative potential and can provide highest current output and hence is widely used in high resistivity electrolytes like soil.

Applications:

1. Protection as buried pipelines, underground cables from soil corrosion.
2. Protection from marine corrosion of cables, ship hulls, piers etc.
3. Insertion of magnesium sheets into the domestic water boilers to prevent the formation of rust.
4. Calcium metal is employed to minimize engine corrosion.



In cathodic protection, an anode of a more strongly reducing metal is sacrificed to maintain the integrity of the protected object (eg., a pipeline, bridge, ship hull or boat).

Advantages:

1. Low installation and operating cost.
2. Capacity to protect complex structures.
3. Applied to wide range of severe corrodents.

Limitations:

1. High starting current is required.
2. Uncoated parts cannot be protected.
3. Limited driving potential, hence, not applicable for large objects.

IMPRESSED CURRENT CATHODIC PROTECTION METHOD

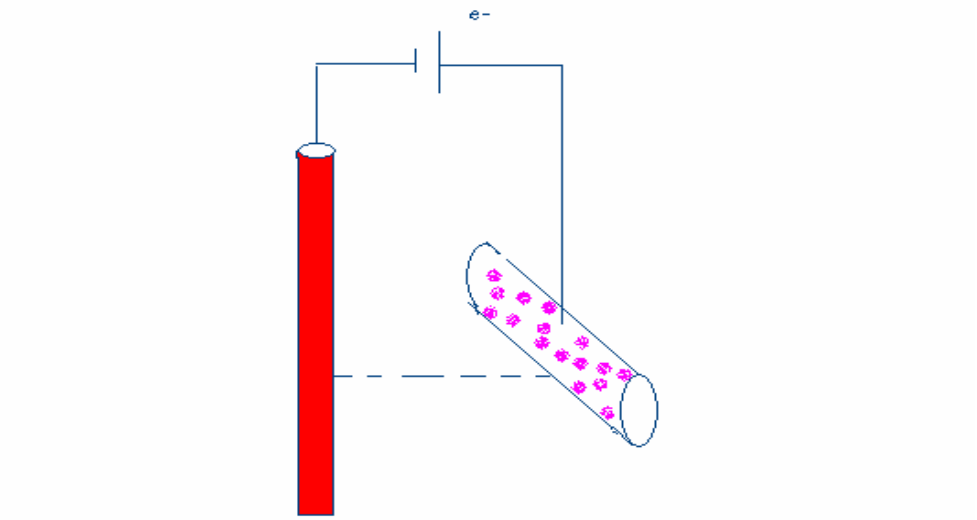
In this method, an impressed current is applied in opposite direction to nullify the corrosion current and convert the corroding metal from anode to cathode.

Usually the impressed current is derived from a direct current sources (like battery or rectifier on AC line) with an imsoluble, inert anode (like graphite, scrap iron, stainless steel, platinum or high silica iron).

A sufficient DC current is applied to an inert anode, buried in the soil (or immersed in the corroding medium) and connected to the metallic structure to be protected. The anode is, usually, a back fill, composed of coke breeze or gypsum, so as to increase the electrical contact with the surrounding soil.

Impressed current cathodic protection has been applied to open water box coolers, water tanks, buried oil or water pipes, condensers, transmission line towers, marine piers, laid up ships etc.

This kind of protection technique is particularly useful for large structures for long term operations.



In Impressed-current cathodic protection, electrons are supplied from an external cell so that the object itself becomes cathodic and is not oxidized.

Comparison of Sacrificial anode method with Impressed current cathodic method:

Sl. No.	Sacrificial Anode method	Impressed Current method
1/	External power supply is not required.	External power supply is required.
2/	The cost of investment is low.	The cost of investment is high.
3/	This requires periodic replacement of sacrificial anode.	Replacement is not required as anodes are stable.
4/	Soil and microbiological corrosion effects are not considered.	Soil and microbiological corrosion effects are taken into account.
5/	This is the most economical method especially when short term protection is required.	This is well suited for large structures and long term operations.
6/	This is a suitable method when the current requirement and the resisitivity of the electrolytes are relatively low.	This is a suitable method even when the current requirement and the resistivity of the electrolytes are high.

MODIFYING THE ENVIRONMENT-CORROSION CONTROL

Environment plays a major role in the corrosion of metals. Hence, we can prevent corrosion to a great extent by modifying the environment. Some of the methods are

i) Deaeration:

Fresh water contains dissolved oxygen. The presence of increased amount of oxygen is harmful and increases the corrosion rate. Deaeration involves the removal of dissolved oxygen by increase of temperature together with mechanical agitation. It also removes dissolved carbon dioxide in water

ii) By using inhibitors:

Inhibitors are organic or inorganic substances which decrease the rate of corrosion. Usually the inhibitors are added in small quantities to the corrosive medium. Inhibitors are classified into

1) Anodic inhibitors (chemical passivators)
2) Cathodic inhibitors (adsorption inhibitors)
3) Vapour phase inhibitors (volatile corrosion inhibitors)

Anodic Inhibitors:

Inhibitors which retard the corrosion of metals by forming a sparingly soluble compound with a newly produced metal cations. This compound will then adsorb on the corroding metal surface forming a passive film or barrier. Anodic inhibitors are used to repair

a) the crack of the oxide film over the metal surface
b) the pitting corrosion
c) the porous oxide film formed on the metal surface.

Examples: Chromate, phosphate, tungstate, nitrate, molybdate etc.

Cathodic Inhibitors:

Depending on the nature of the cathodic reaction in an electrochemical corrosion, cathodic inhibitors are classified into

a) **In an acidic solution:** the main cathodic reaction is the liberation of hydrogen gas, the corrosion can be controlled by slowing down the diffusion of H+ ions through the cathode. Eg., Amines, Mercaptans, Thiourea etc.

$$2H^+ + 2e^- \longrightarrow H_2 \uparrow$$

b) **In a neutral solution:** in a neutral solution, the cathodic reaction is the adsorption of oxygen or formation of hydroxyl ions.

The corrosion is therefore controlled either by eliminating oxygen from the corroding medium or by retarding its diffusion to the cathodic area.

The dissolved oxygen can be eliminated by adding reducing agents like Na_2SO_3.

The diffusion of oxygen can be controlled by adding inhibitors like Mg, Zn or Ni salts.

Eg., Na_2SO_3, N_2H_4, Salts of Mg, Zn or Ni.

$$\tfrac{1}{2} O_2 + H_2O + 2e^- \longrightarrow 2OH^-$$

Vapour phase inhibitors:

These are organic inhibitors which are readily vapourised and form a protective layer on the metal surface.

These are conveniently used to prevent corrosion in closed spaces, storage containers, packing materials, sophisticated equipments etc.

Examples are Dicyclohexylammonium nitrate, dicyclohexyl ammonium chromate, benzotriazole, phenylthiourea etc.

ANODIC PROTECTION

This is an electrochemical method of corrosion control in which an external potential control system, called potentiostat, is used to produce and maintain a thin non corroding, passive film on a metal or an alloy. The use of potentiostat is to shift corrosion potential into passive potential so that the corrosion of the metal is stopped.

The potential of the object (say acid storage tank) to be protected is controlled by potential controller (potentiostat) so that under certain potential range, the object becomes passive and prevents further corrosion. This potential range depends upon the relationship between the metal and the environment.

Applications:

1. Used in acid coolers in dilute sulphuric acid plants
2. used in storage tanks for sulphuric acid
3. used in chromium in contact with hydrofluoric acid

Limitations:

1. This method cannot be applied in the case of corrosive medium containing aggressive chloride.
2. This cannot be applied if protection breaks down at any point, it is difficult to reestablish.

PROTECTIVE COATINGS

Introduction:

In order to protect metals from corrosion, it is necessary to cover the surface by means of protective coatings. These coatings act as a physical barrier between the coated metal surface and the environment. They afford decorative appeal and impart special properties like hardness, oxidation resistance and thermal insulation.

Classification:

Protective coatings can be broadly classified into two types. They are

1. Inorganic coatings **2) Organic coatings**

Inorganic coatings are further classified into two types. They are

i) Metallic coating:

1. Hot dipping- Galvanising, Tinning
2. Metal cladding
3. Cementation-Sherardising, Chromising, Calorising
4. Electroplating.

ii) Non-metallic coating:

1. Surface coating or chemical conversion coating – Chromate coating, Phosphate coating and Oxide coating.
2. Anodising
3. Enamel coating or Vitreous or Porcelain coating.

Organic coatings consists of

Paints, Varnishes, Lacquers and Enamels.

PAINTS

Paint is a viscous, opaque (not clear), mechanical dispersion mixture of one or more pigments (dye) in a vehicle (drying oil).

Requisites of a good paint:

A good paint should the following properties, it should

1) have a high hiding power
2) form a good and uniform film on the metal surface
3) the film should not crack on drying
4) give a glossy film
5) the film produced should be washable
6) give a stable and decent colour on the metal surface
7) have good resistance to the atmospheric conditions
8) be fluid enough to spread easily over the surface
9) possess high adhesion capacity to the material over which it is intended to be used
10) dry quickly or in a reasonable duration.
11) the colour of the paint should not fade.

Constituents of paint:

a) Pigment
b) Vehicle or medium or drying oil
c) Thinner
d) Driers
e) Fillers or Extenders
f) Plasticizers
g) Antiskinning agents

a) Pigment:

It is a solid substance which imparts colour to the paint. It is an essential constituent of a paint. Its functions are to

i) Give opacity(cloudiness) and colour to the film
ii) Provide strength to the paint

iii) Provide an aesthetical appeal
iv) Give protection to the paint film by reflecting UV light.
v) Increase weather resistance of the film
vi) Provide resistance to paint film against abrasion.

The most commonly used pigments in paints and the compounds required as as follows:

White pigments	-	White lead, ZnO, $BaSO_4$, TiO_2, ZrO_2
Blue pigments	-	Prussian blue, ultramarine blue
Black pigments	-	Graphite, carbon black, lamp black
Red pigments	-	Red lead, Fe_3O_4, carmine
Green pigments	-	Chromium oxide, chrome green
Brown pigments	-	Burnt umber, ochre
Yellow pigments	-	Chrome yellow, lead chromate

b) Vehicle or drying oil or medium:

Vehicle is a liquid substance and film forming material. It holds all the ingredients of a paint in liquid suspension. Eg., linseed oil, tung oil.

Functions:

i) To hold the pigment on the metal surface
ii) to form the protective film by evaporation or by other means.
iii) to impart water repellency, durability and toughness to the film
iv) to improve the adhesion of the film

c) Thinners

Thinners are volatile substances which evaporate easily after application of the paint. They are added to the paints for reducing the viscosity of the paints so that they can be easily applied to the metal surface. Eg., Dipentine, turpentine, toluol, xylol.

Functions:

i) To reduce the viscosity of the paint
ii) To dissolve vehicle and the additives in the vehicle
iii) To suspend the pigments
iv) To increase the penetration power of the vehicle
v) To increase the elasticity of the paint film
vi) To help the drying of the paint film.

d) Driers:

These are the substances used to accelerate the process of drying. They are oxygen carrier catalysts. Eg., Naphthenates, linoleates, borates, resonates and tungstates of heavy metals (Pb, Zn, Co, Mn).

Functions:

i) To accelerate the drying of the oil film through oxidation, polymerization and condensation

ii) To improve the drying quality of the oil film.

e) Extenders or Fillers:

These are the inert materials which improve the properties of the paint. Eg., Gypsum, chalk, silica, talc, clay, $CaCO_3$, $CaSO_4$.

Functions:

i) To fill the voids (empty space or any curved area) in the film
ii) To act as a carrier for the pigment color.
iii) To reduce the cost of the paint
iv) To increase the durability of the paint
v) To reduce the cracking of dry paint
vi) To increase random arrangement of pigment particles.

f) Plasticisers:

These are added to the paint to provide elasticity to the film and to minimize its crack. Eg., Triphenyl phosphate, dibutyl tartarate, tributyl phthalate, tricresyl phosphate, diamyl phthalate.

g) Antiskinning agents:

These are sometimes added to some paints to prevent gelling and skinning of the finished product. Eg., Polyhydroxy phenols.

METALLIC COATINGS:

Corrosion of metals can be prevented or controlled by using methods like galvanization, tinning, metal cladding, electroplating, cementation, anodizing, phosphate coating, enamelling, electroless plating. Some of the methods are

1) Hot dipping:

It is used for producing a coating of low-melting metals such as Zn (m.p.=419 deg C), Sn (m.p.=232 deg C), Pb, Al etc., on iron, steel and copper which have relatively higher melting points. The process in immersing the base metal in a bath of the molten coating-metal, covered by a molten flux layer (usually zinc chloride).

2) Galvanizing:

It is the process of coating iron or steel sheets with a thin coat of zinc to prevent them from rusting. The process is iron or steel article is first cleaned with dil. Sulphuric acid and washed with distilled water and dried. The dried metal is dipped in bath of molten zinc, now the thin layer of zinc is coated on the iron or steel article.

3) Metal cladding:

It is the process by which a dense, homogeneous layer of coating metal is bonded firmly and permanently to the base metal on one or both sides. Corrosion resistant metals like nickel, copper,

lead, silver, platinum and alloys like SS, nickel alloys, copper alloys, lead alloys can be used as cladding materials.

4) Tinning:

It is a method of coating tin over the iron or steel articles. The process is first treating steel sheet in dilute sulphuric acid and it is passed through a flux ($ZnCl_2$), next steel passes through a tank of molten tin and finally through a series of rollers from underneath (bottom of) the surface of a layer of palm oil.

ELECTROPLATING OR ELECTRODEPOSITION

Electroplating is a coating technique. It is the most important and most frequently applied industrial method of producing metallic coating.

Electroplating is the process by which the coating metal is deposited on the base metal by passing a direct current through an electrolytic solution containing the soluble salt of the coating metal.

The base metal to be plated is made cathode whereas the anode is either made of the coating metal itself or an inert material of good electrical conductivity (like graphite).

Objectives:

Electroplating is carried out for

1) Decoration or better appearance
2) Increasing the resistance to corrosion of the coated metal.
3) Improving the hardness of the metal
4) Increasing the resistance to chemical attack
5) Electro refining.

Procedure:

The article is to be plated first treated with organic solvent like carbon tetrachloride, acetone, tetrachloro ethylene to remove oils, greases etc. Then it is made free from surface scale, oxides, etc. by treating with dil. HCl or H_2SO_4 (acid pickling). The cleaned article is then made as the cathode of the electrolytic cell. The anode is either the coating metal itself or an inert material of good electrical conductivity. The electrolyte is a solution of soluble salt of the coating metal.

When direct current is passed, coating metal ions migrate to the cathode and get deposited there. Thus, a thin layer of coating metal is obtained on the article made as the cathode.

In order to get strong, adherent and smooth deposit, certain types of additives (glue, gelatin, boric acid) are added to the electrolytic bath.

In order to improve the brightness of the deposit, brightening agents are added in the electrolytic bath.

The favourable conditions for a good electrodeposit are

i) Optimum temperature ii) Optimum current density iii) Low metal ion concentrations.

Gold Electroplating:

Anode: Gold
Cathode: Metal article
Electrolyte: Gold + KCN
Temperature: 60 deg C.
Current density (mA cm-2): 1-10

Use:

i) This is used for electrical and electronic applications.
ii)It is used for high quality decorations and high oxidation resistant coatings
iii)Usually for ornamental jewellery, a very thin goldcoating (about 1x10-4 cm) is given.

ELECTROLESS PLATING

Principle

Electroless plating is a newer technique of depositing a noble metal from its salt solution on a catalytically active surface of the metal to be protected by using a suitable reducing agent without using electrical energy.

The reducing agent reduces the metal ions into metal which gets plated over the catalytically activated surface giving a uniform and thin coating.

Metal ions + reducing agents $\longrightarrow$ Metal (deposited) + Oxidised product

ELECTROLESS NICKEL PLATING:

Pretreatment and activation of the surface:

The surface to be plated is first degreased by using organic solvents or alkali, followed by acid treatment.

i) The surface of the stainless steel is activated by dipping in hot solutionof 50 % dilute sulphuric acid.

ii) The surface of magnesium alloy is activated by thin coating of zinc or copper over it.

iii) Metals (Al, Cu, Fe) and alloys like brass can be directly nickel plated without activation.

iv) Non metallic articles like plastics, glass are activated by dipping them in the solution containing $SnCl_2$ +HCl, followed by dipping in palladium chloride solution. On drying, a thin layer of palladium is formed on the surface.

Preparation of plating bath:

The details of preparation of plating bath are:

Sl. No.	Nature of the compound	Name of the compound	Quantity (g / L)
1 /	Coating solution	Nickel chloride ($NiCl_2$)	20
2 /	Reducing agent	Sodium hypophosphite (NaH_2PO_2)	20
3 /	Complexing agent	Sodium succinate	15
4 /	Buffer	Sodium acetate	10
5 /	Optimum pH	4.5	--
6 /	Optimum temperature	93 deg C	--

Procedure:

The pretreated object is immersed in the plating bath for the required time. The following reactions occur and nickel gets coated over the object.

At cathode: $Ni^{2+} + 2\,e^- \longrightarrow Ni$

At anode: $H_2PO_2^- + H_2O \longrightarrow H_2PO_3^- + 2H^+ + 2\,e^-$

Net reaction: $Ni^{2+} + H_2PO_2^- + H_2O \longrightarrow Ni + H_2PO_3^- + 2H^+$

Applications:

1) It is used extensively in electronic appliances.
2) It is used in domestic as well as automotive fields (eg., jewellery, tops of perfume bottles).
3) Its polymers are used in decorative and functional works.
4) Its plastic cabinets are used in digital as well as electronic instruments.

References:

1) Jain and Jain, Engineering Chemistry, 15th Edition, Dhanpat Rai Publishing Co., New Delhi.
2) S.S. Dara, Engineering Chemistry, 1st Edition, S. Chand & Co, New Delhi.

Model Questions (2 Marks)

1) Define corrosion.
2) What is meant by rusting of iron?
3) What is wet corrosion?
4) State Pilling- Bedworth rule.
5) Define water line corrosion.
6) What is pitting corrosion?
7) What is galvanic corrosion?
8) The rate of metallic corrosion increases with increase in temperature. Give reason.
9) Differentiate chemical and electrochemical corrosion.
10) What is differential aeration corrosion?
11) Mention the factors influencing corrosion.
12) What is corrosion control and why is it required?
13) Write a small note on cathodic protection.
14) Write a small note on anodic protection.
15) What should be the nature of the corrosion product to prevent further corrosion?
16) What are the important constituents of paint?
17) Bolt and nut made of the same metal is preferred in practice. Why?
18) What is metal cladding?
19) Why coating of zinc on iron is called sacrificial anode?
20) During electroplating, pH of bath is strictly maintained. Give reasons.
21) Give any three functions of pigments in paints.

Model Questions (6 Marks)

1) What is corrosion of metals? Explain the mechanism of oxidation corrosion.
2) What are the factors that affect electrochemical corrosion rate? Discuss.
3) Differentiate chemical and electrochemical corrosion. Mention any four factors that affect electrochemical corrosion.
4) Describe the mechanism of electrochemical corrosion by hydrogen evolution and oxygen adsorption.
5) Explain water line corrosion.
6) How is galvanic corrosion occur>
7) Deposition of oil or dust on metal surfaces for a long period is undesirable. Give reasons.
8) Describe the mechanism of differential aeration corrosion taking pitting as example.
9) Explain the electrochemical theory of corrosion with suitable example.
10) Discuss the mechanism of chemical and electrochemical corrosion.
11) Explain the following:
    i) hydrogen embrittlement
    ii) decarburation
    iii) liquid metal corrosion
    iv) water line corrosion
    v) pitting corrosion

vi) crevice corrosion
vii) pipeline corrosion

12) Substantiate the statement that nature of the environment affects corrosion.
13) What is sacrificial anode? Mention its role in the prevention of corrosion.
14) Write short note on corrosion control by impressed current method.
15) What are corrosion inhibitors? How do they function?
16) Explain how corrosion of metals controlled by sacrificial anode technique.
17) Write a note on pitting corrosion and cathodic protection.
18) Mention the constituents of a paint. Explain the function of the various constituents.
19) Describe the mechanisms of drying of an oil.
20) How the Hot dipping process is carried out?

INTRODUCTION TO NANOMATERIALS

Introduction:

Nanomaterials are cornerstones of nanoscience and nanotechnology. Nanostructure science and technology is a broad and interdisciplinary area of research and development activity that has been growing explosively worldwide in the past few years. It has the potential for revolutionizing the ways in which materials and products are created and the range and nature of functionalities that can be accessed. It is already having a significant commercial impact, which will assuredly increase in the future.

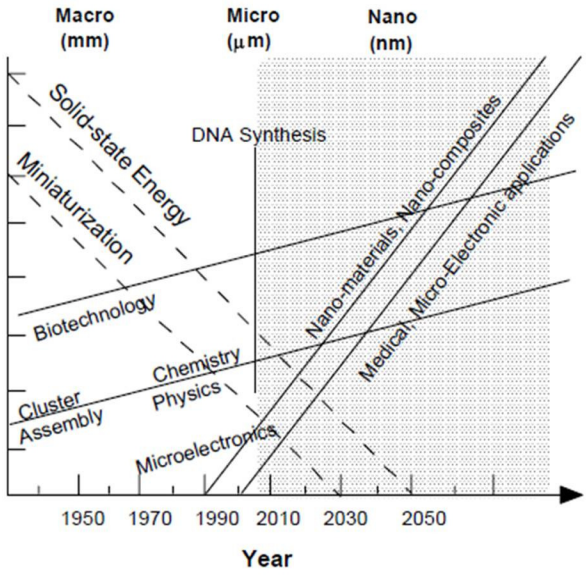


Fig. (1): Evolution of science & technology and the future

What are nanomaterials?

Nanoscale materials are defined as a set of substances where at least one dimension is less than approximately 100 nanometers. A nanometer is one millionth of a millimeter - approximately 100,000 times smaller than the diameter of a human hair. Nanomaterials are of interest because at this scale unique optical, magnetic, electrical, and other properties emerge. These emergent properties have the potential for great impacts in electronics, medicine, and other fields.

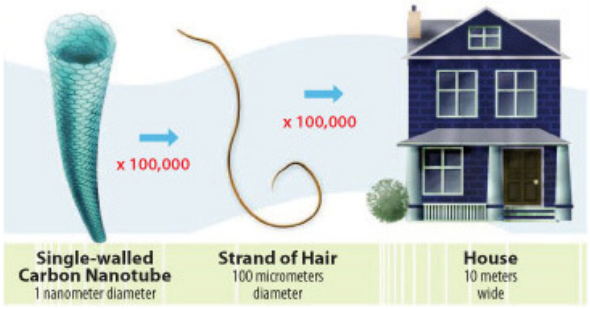


Fig. (2): Nanomaterial (For example: Carbon nanotube)

Where are nanomaterials found?

Some nanomaterials occur naturally, but of particular interest are engineered nanomaterials (EN), which are designed for, and already being used in many commercial products and processes. They can be found in such things as sunscreens, cosmetics, sporting goods, stain-resistant clothing, tires, electronics, as well as many other everyday items, and are used in medicine for purposes of diagnosis, imaging and drug delivery.

Engineered nanomaterials are resources designed at the molecular (nanometre) level to take advantage of their small size and novel properties which are generally not seen in their conventional, bulk counterparts. The two main reasons why materials at the nano scale can have different properties are increased relative surface area and new quantum effects. Nanomaterials have a much greater surface area to volume ratio than their conventional forms, which can lead to greater chemical reactivity and affect their strength. Also at the nano scale, quantum effects can become much more important in determining the materials properties and characteristics, leading to novel optical, electrical and magnetic behaviours.

Nanomaterials are already in commercial use, with some having been available for several years or decades. The range of commercial products available today is very broad, including stain-resistant and wrinkle-free textiles, cosmetics, sunscreens, electronics, paints and varnishes. Nanocoatings and nanocomposites are finding uses in diverse consumer products, such as windows, sports equipment, bicycles and automobiles. There are novel UV-blocking coatings on glass bottles which protect beverages from damage by sunlight, and longer-lasting tennis balls using butyl-rubber/nano-clay composites. Nanoscale titanium dioxide, for instance, is finding applications in cosmetics, sun-block creams and self-cleaning windows, and nanoscale silica is being used as filler in a range of products, including cosmetics and dental fillings.

History of Nanomaterials:

The history of nanomaterials began immediately after the big bang when Nanostructures were formed in the early meteorites. Nature later evolved many other Nanostructures like seashells, skeletons etc. Nanoscaled smoke particles were formed during the use of fire by early humans. The scientific story of nanomaterials however began much later. One of the first scientific report is the colloidal gold particles synthesized by Michael Faraday as early as 1857. Nanostructured catalysts have also been investigated for over 70 years. By the early 1940's, precipitated and fumed silica nanoparticles were being manufactured and sold in USA and Germany as substitutes for ultrafine carbon black for rubber reinforcements.

Nanosized amorphous silica particles have found large-scale applications in many every-day consumer products, ranging from non-diary coffee creamer to automobile tires, optical fibers and catalyst supports. In the 1960s and 1970's metallic nanopowders for magnetic recording tapes were developed. In 1976, for the first time, nanocrystals produced by the now popular inert- gas evaporation technique was published by Granqvist and Buhrman. Recently it has been found that the Maya blue paint is a nanostructured hybrid material. The origin of its color and its resistance to acids and biocorrosion are still not understood but studies of authentic samples from Jaina Island show that the material is made of needle-shaped palygorskite (clay) crystals that form a superlattice with a period of 1.4 nm, with intercalates of amorphous silicate substrate containing

inclusions of metal (Mg) nanoparticles. The beautiful tone of the blue color is obtained only when both these nanoparticles and the superlattice are present, as has been shown by the fabrication of synthetic samples.

Today nanophase engineering expands in a rapidly growing number of structural and functional materials, both inorganic and organic, allowing to manipulate mechanical, catalytic, electric, magnetic, optical and electronic functions. The production of nanophase or cluster-assembled materials is usually based upon the creation of separated small clusters which then are fused into a bulk-like material or on their embedding into compact liquid or solid matrix materials. e.g. nanophase silicon, which differs from normal silicon in physical and electronic properties, could be applied to macroscopic semiconductor processes to create new devices. For instance, when ordinary glass is doped with quantized semiconductor "colloids," it becomes a high performance optical medium with potential applications in optical computing.

Classification of Nanomaterials:

Nanomaterials have extremely small size which having at least one dimension 100 nm or less. Nanomaterials can be nanoscale in one dimension (eg. surface films), two dimensions (eg. strands or fibres), or three dimensions (eg. particles). They can exist in single, fused, aggregated or agglomerated forms with spherical, tubular, and irregular shapes. Common types of nanomaterials include nanotubes, dendrimers, quantum dots and fullerenes. Nanomaterials have applications in the field of nano technology, and displays different physical chemical characteristics from normal chemicals (i.e., silver nano, carbon nanotube, fullerene, photocatalyst, carbon nano, silica).

According to Siegel, Nanostructured materials are classified as Zero dimensional, one dimensional, two dimensional, three dimensional nanostructures.

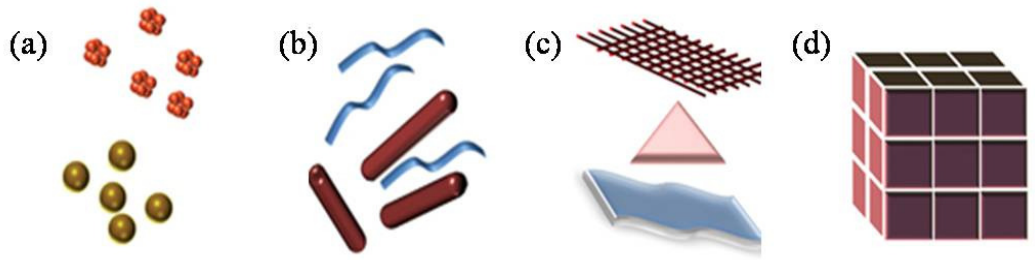


Fig (3): Classification of Nanomaterials (a) 0D spheres and clusters; (b) 1D nanofibers, nanowires, and nanorods; (c) 2D nanofilms, nanoplates, and networks; (d) 3D nanomaterials.

Nanomaterials are materials which are characterized by an ultra fine grain size (< 50 nm) or by a dimensionality limited to 50 nm. Nanomaterials can be created with various modulation dimensionalities as defined by Richard W. Siegel: zero (atomic clusters, filaments and cluster assemblies), one (multilayers), two (ultrafine-grained overlayers or buried layers), and three (nanophase materials consisting of equiaxed nanometer sized grains) as shown in the above figure 3.

Why so much interest in nanomaterials?

These materials have created a high interest in recent years by virtue of their unusual mechanical, electrical, optical and magnetic properties. Some examples are given below:

- Nanophase ceramics are of particular interest because they are more ductile at elevated temperatures as compared to the coarse-grained ceramics.

- Nanostructured semiconductors are known to show various non-linear optical properties. Semiconductor Q-particles also show quantum confinement effects which may lead to special properties, like the luminescence in silicon powders and silicon germanium quantum dots as infrared optoelectronic devices. Nanostructured semiconductors are used as window layers in solar cells.

- Nanosized metallic powders have been used for the production of gas tight materials, dense parts and porous coatings. Cold welding properties combined with the ductility make them suitable for metal-metal bonding especially in the electronic industry.

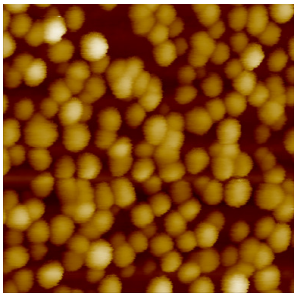
- Single nanosized magnetic particles are mono-domains and one expects that also in magnetic nanophase materials the grains correspond with domains, while boundaries on the contrary to disordered walls. Very small particles have special atomic structures with discrete electronic states, which give rise to special properties in addition to the super-paramagnetism behaviour. Magnetic nanocomposites have been used for mechanical force transfer (ferrofluids), for high density information storage and magnetic refrigeration.

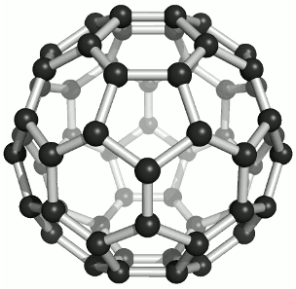
- Nanostructured metal clusters and colloids of mono- or plurimetallic composition have a special impact in catalytic applications. They may serve as precursors for new type of heterogeneous catalysts (Cortex-catalysts) and have been shown to offer substantial advantages concerning activity, selectivity and lifetime in chemical transformations and electrocatalysis (fuel cells). Enantioselective catalysis was also achieved using chiral modifiers on the surface of nanoscale metal particles.

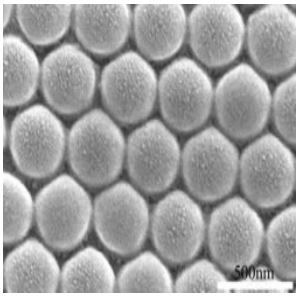
- Nanostructured metal-oxide thin films are receiving a growing attention for the realization of gas sensors (NOx, CO, CO_2, CH_4 and aromatic hydrocarbons) with enhanced sensitivity and selectivity. Nanostructured metal-oxide (MnO_2) finds application for rechargeable batteries for cars or consumer goods. Nanocrystalline silicon films for highly transparent contacts in thin film solar cell and nano-structured titanium oxide porous films for its high transmission and significant surface area enhancement leading to strong absorption in dye sensitized solar cells.

- Polymer based composites with a high content of inorganic particles leading to a high dielectric constant are interesting materials for photonic band gap structure.

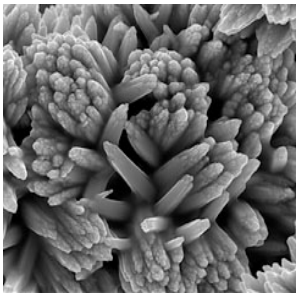
Examples of Nanomaterials:

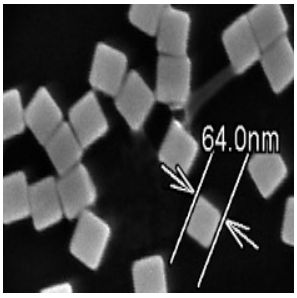

Au nanoparticle


Buckminsterfullerene



FePt nanosphere


Titanium nanoflower



Silver nanocubes

SnO2 nanoflower

NANOMATERIAL SYNTHESIS AND PROCESSING

We are dealing with very fine structures: a nanometer is a billionth of a meter. This indeed allows us to think in both the 'bottom up' or the 'top down' approaches to synthesize nanomaterials, i.e. either to assemble atoms together or to dis-assemble (break, or dissociate) bulk solids into finer pieces until they are constituted of only a few atoms. This domain is a pure example of interdisciplinary work encompassing physics, chemistry, and engineering upto medicine.

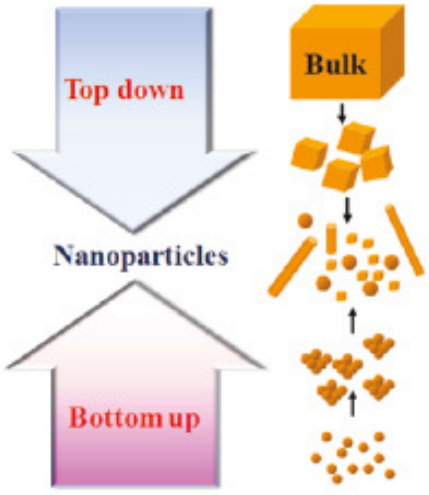


Fig. (4): Schematic illustration of the preparative methods of nanoparticles.

Methods for creating nanostructures:

There are many different ways of creating nanostructures: of course, macromolecules or nanoparticles or buckyballs or nanotubes and so on can be synthesized artificially for certain specific materials. They can also be arranged by methods based on equilibrium or near-equilibrium thermodynamics such as methods of self-organization and self-assembly (sometimes also called bio-mimetic processes). Using these methods, synthesized materials can be arranged into useful shapes so that finally the material can be applied to a certain application.

Mechanical grinding:

Mechanical attrition is a typical example of 'top down' method of synthesis of nanomaterials, where the material is prepared not by cluster assembly but by the structural decomposition of coarser-grained structures as the result of severe plastic deformation. This has become a popular method to make nanocrystalline materials because of its simplicity, the relatively inexpensive equipment needed, and the applicability to essentially the synthesis of all classes of materials. The major advantage often quoted is the possibility for easily scaling up to tonnage quantities of material for various applications. Similarly, the serious problems that are usually cited are;

1. contamination from milling media and/or atmosphere, and
2. to consolidate the powder product without coarsening the nanocrystalline microstructure.

In fact, the contamination problem is often given as a reason to dismiss the method, at least for some materials. Here we will review the mechanisms presently believed responsible for formation of nanocrystalline structures by mechanical attrition of single phase powders, mechanical alloying of dissimilar powders, and mechanical crystallisation of amorphous materials. The two important problems of contamination and powder consolidation will be briefly considered.

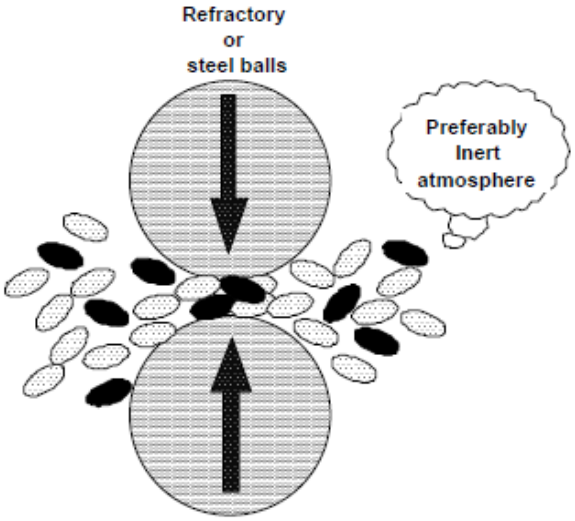


Fig. (5): Schematic representation of the principle of mechanical milling

Mechanical milling is typically achieved using high energy shaker, planetary ball, or tumbler mills. The energy transferred to the powder from refractory or steel balls depends on the

rotational (vibrational) speed, size and number of the balls, ratio of the ball to powder mass, the time of milling and the milling atmosphere. Nanoparticles are produced by the shear action during grinding.

Milling in cryogenic liquids can greatly increase the brittleness of the powders influencing the fracture process. As with any process that produces fine particles, an adequate step to prevent oxidation is necessary. Hence this process is very restrictive for the production of non-oxide materials since then it requires that the milling take place in an inert atmosphere and that the powder particles be handled in an appropriate vacuum system or glove box. This method of synthesis is suitable for producing amorphous or nanocrystalline alloy particles, elemental or compound powders. If the mechanical milling imparts sufficient energy to the constituent powders a homogeneous alloy can be formed. Based on the energy of the milling process and thermodynamic properties of the constituents the alloy can be rendered amorphous by this processing.

Wet Chemical Synthesis of Nanomaterials

In principle we can classify the wet chemical synthesis of nanomaterials into two broad groups:

1. The top down method: where single crystals are etched in an aqueous solution for producing nanomaterials, For example, the synthesis of porous silicon by electrochemical etching.
2. The bottom up method: consisting of sol-gel method, precipitation etc. where materials containing the desired precursors are mixed in a controlled fashion to form a colloidal solution.

Sol-gel process

The sol-gel process, involves the evolution of inorganic networks through the formation of a colloidal suspension (**sol**) and gelation of the sol to form a network in a continuous liquid phase (**gel**). The precursors for synthesizing these colloids consist usually of a metal or metalloid element surrounded by various reactive ligands. The starting material is processed to form a dispersible oxide and forms a sol in contact with water or dilute acid. Removal of the liquid from the sol yields the gel, and the sol/gel transition controls the particle size and shape. Calcination of the gel produces the oxide.

Sol-gel processing refers to the hydrolysis and condensation of alkoxide-based precursors such as $Si(OEt)_4$ (tetraethyl orthosilicate, or TEOS). The reactions involved in the sol-gel chemistry based on the hydrolysis and condensation of metal alkoxides M(OR)**z** can be described as follows:

$$MOR + H_2O \rightarrow MOH + ROH \text{ (hydrolysis)}$$
$$MOH + ROM \rightarrow M\text{-}O\text{-}M + ROH \text{ (condensation)}$$

Sol-gel method of synthesizing nanomaterials is very popular amongst chemists and is widely employed to prepare oxide materials. The sol-gel process can be characterized by a series of distinct steps.

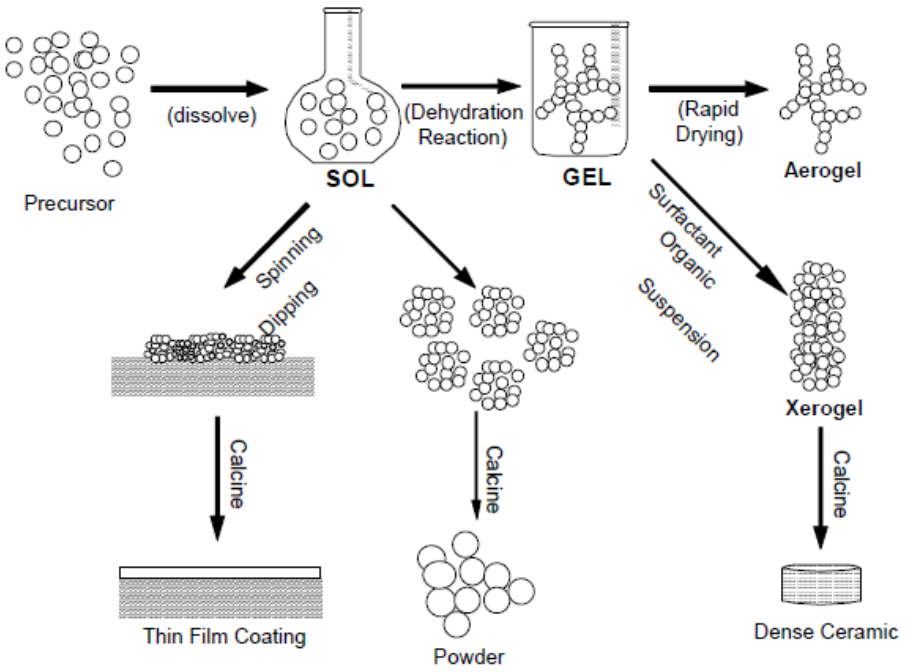


Fig. (6): Schematic representation of sol-gel process of synthesis of nanomaterials.

1. Formation of different stable solutions of the alkoxide or solvated metal precursor.
2. Gelation resulting from the formation of an oxide- or alcohol- bridged network (the gel) by a polycondensation reaction that results in a dramatic increase in the viscocity of the solution.
3. Aging of the gel (Syneresis), during which the polycondensation reactions continue until the gel transforms into a solid mass, accompanied by contraction of the gel network and expulsion of solvent from gel pores. Ostwald ripening (also referred to as coarsening, is the phenomenon by which smaller particles are consumed by larger particles during the growth process) and phase transformations may occur concurrently with syneresis. The aging process of gels can exceed 7 days and is critical to the prevention of cracks in gels that have been cast.
4. Drying of the gel, when water and other volatile liquids are removed from the gel network. This process is complicated due to fundamental changes in the structure of the gel. The drying process has itself been broken into four distinct steps: (i) the constant rate period, (ii) the critical point, (iii) the falling rate period, (iv) the second falling rate period. If isolated by thermal evaporation, the resulting monolith is termed a *xerogel*. If the solvent (such as water) is extracted under supercritical or near super critical conditions, the product is an *aerogel*.
5. Dehydration, during which surface- bound M-OH groups are removed, there by stabilizing the gel against rehydration. This is normally achieved by calcining the monolith at temperatures up to 800^0C.
6. Densification and decomposition of the gels at high temperatures (T>800^0C). The pores of the gel network are collapsed, and remaining organic species are volatilized. The typical steps that are involved in sol-gel processing are shown in the schematic diagram below.

The interest in this synthesis method arises due to the possibility of synthesizing nonmetallic inorganic materials like glasses, glass ceramics or ceramic materials at very low temperatures compared to the high temperature process required by melting glass or firing ceramics.
The major difficulties to overcome in developing a successful bottom-up approach is controlling the growth of the particles and then stopping the newly formed particles from agglomerating. Other technical issues are ensuring the reactions are complete so that no unwanted reactant is left on the product and completely removing any growth aids that may have been used in the process. Also production rates of nano powders are very low by this process. The main advantage is one can get monosized nano particles by any bottom up approach.

Gas Phase synthesis of nanomaterials:

The gas-phase synthesis methods are of increasing interest because they allow elegant way to control process parameters in order to be able to produce size, shape and chemical composition controlled nanostructures. Before we discuss a few selected pathways for gas-phase formation of nanomaterials, some general aspects of gas-phase synthesis needs to be discussed. In conventional chemical vapour deposition (CVD) synthesis, gaseous products either are allowed to react homogeneously or heterogeneously depending on a particular application.

1. In homogeneous CVD, particles form in the gas phase and diffuse towards a cold surface due to thermophoretic forces, and can either be scrapped of from the cold surface to give nano-powders, or deposited onto a substrate to yield what is called ***'particulate films'***.

2. In heterogeneous CVD, the solid is formed on the substrate surface, which catalyses the reaction and a dense film is formed.

In order to form nanomaterials several modified CVD methods have been developed. Gas phase processes have inherent advantages, some of which are noted here:

- An excellent control of size, shape, crystallinity and chemical composition
- Highly pure materials can be obtained
- Multicomonent systems are relatively easy to form
- Easy control of the reaction mechanisms

Most of the synthesis routes are based on the production of small clusters that can aggregate to form nano particles (condensation). Condensation occurs only when the vapour is supersaturated and in these processes homogeneous nucleation in the gas phase is utilised to form particles. This can be achieved both by physical and chemical methods.

Furnace:

The simplest fashion to produce nanoparticles is by heating the desired material in a heat-resistant crucible containing the desired material. This method is appropriate only for materials that have a high vapour pressure at the heated temperatures that can be as high as 2000°C. Energy is normally introduced into the precursor by arc heating, electron-beam heating or Joule heating. The atoms are evaporated into an atmosphere, which is either inert (e.g. He) or reactive (so as to form a compound). To carry out reactive synthesis, materials with very low vapour

pressure have to be fed into the furnace in the form of a suitable precursor such as organometallics, which decompose in the furnace to produce a condensable material. The hot atoms of the evaporated matter lose energy by collision with the atoms of the cold gas and undergo condensation into small clusters via homogeneous nucleation. In case a compound is being synthesized, these precursors react in the gas phase and form a compound with the material that is separately injected in the reaction chamber. The clusters would continue to grow if they remain in the supersaturated region. To control their size, they need to be rapidly removed from the supersaturated environment by a carrier gas. The cluster size and its distribution are controlled by only three parameters:

1) the rate of evaporation (energy input),
2) the rate of condensation (energy removal), and
3) the rate of gas flow (cluster removal).

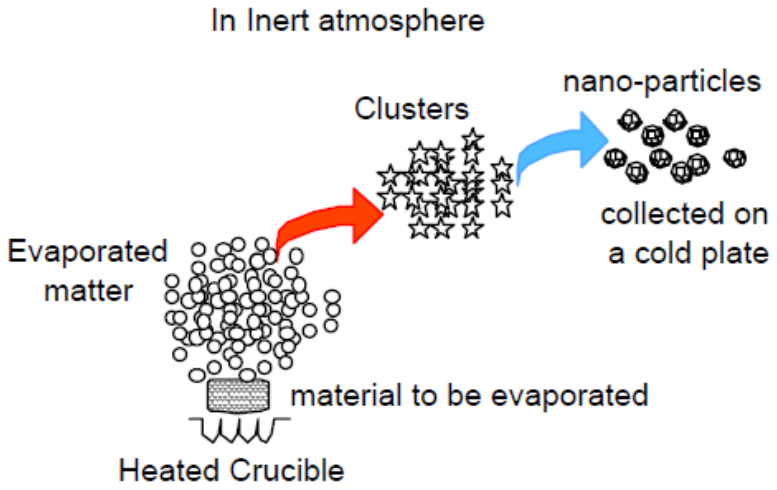


Fig. (7): Schematic representation of gas phase process of synthesis of single phase nanomaterials from a heated crucible.

Because of its inherent simplicity, it is possible to scale up this process from laboratory (mg/day) to industrial scales (tons/day).

Flame assisted ultrasonic spray pyrolysis

In this process, precusrsors are nebulized and then unwanted components are burnt in a flame to get the required material, eg. ZrO_2 has been obtained by this method from a precursor of Zr $(CH_3$ CH_2 $CH_2O)_4$. Flame hydrolysis that is a variant of this process is used for the manufacture of fused silica. In the process, silicon tetrachloride is heated in an oxy-hydrogen flame to give highly dispersed silica. The resulting white amorphous powder consists of spherical particles with sizes in the range 7-40 nm. The combustion flame synthesis, in which the burning of a gas mixture, e.g. acetylene and oxygen or hydrogen and oxygen, supplies the energy to initiate the pyrolysis of precursor compounds, is widely used for the industrial production of powders in large quantities, such as carbon black, fumed silica and titanium dioxide. However, since the gas pressure during the reaction is high, highly agglomerated powders are produced which is

disadvantageous for subsequent processing. The basic idea of low pressure combustion flame synthesis is to extend the pressure range to the pressures used in gas phase synthesis and thus to reduce or avoid the agglomeration. Low pressure flames have been extensively used by aerosol scientists to study particle formation in the flame.

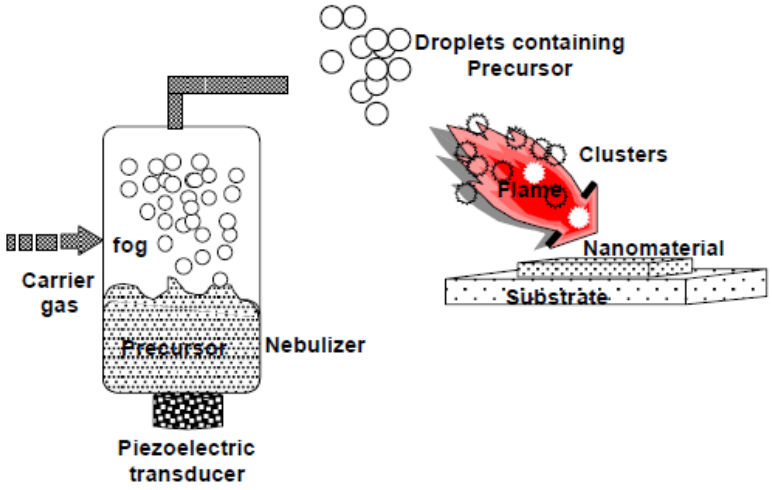


Fig. (8): Flame assisted ultrasonic spray pyrolysis

A key for the formation of nanoparticles with narrow size distributions is the exact control of the flame in order to obtain a flat flame front. Under these conditions the thermal history, i.e. time and temperature, of each particle formed is identical and narrow distributions result. However, due to the oxidative atmosphere in the flame, this synthesis process is limited to the formation of oxides in the reactor zone.

Gas Condensation Processing (GPC)

In this technique, a metallic or inorganic material, e.g. a suboxide, is vaporised using thermal evaporation sources such as crucibles, electron beam evaporation devices or sputtering sources in an atmosphere of 1-50 mbar He (or another inert gas like Ar, Ne, Kr). Cluster form in the vicinity of the source by homogenous nucleation in the gas phase and grow by coalescence and incorporation of atoms from the gas phase. The cluster or particle size depends critically on the residence time of the particles in the growth system and can be influenced by the gas pressure, the kind of inert gas, i.e. He, Ar or Kr, and on the evaporation rate/vapour pressure of the evaporating material. With increasing gas pressure, vapour pressure and mass of the inert gas used the average particle size of the nanoparticles increases. Lognormal size distributions have been found experimentally and have been explained theoretically by the growth mechanisms of the particles. Even in more complex processes such as the low pressure combustion flame synthesis where a number of chemical reactions are involved the size distributions are determined to be lognormal.

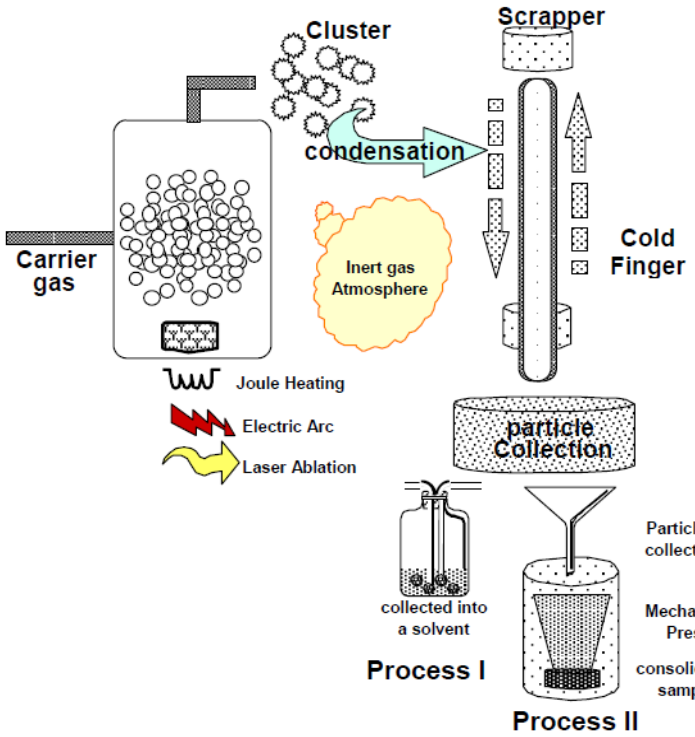


Fig. (9): Schematic representation of typical set-up for gas condensation synthesis of nanomaterials followed by consolidation in a mechanical press or collection in an appropriate solvent media.

Originally, a rotating cylindrical device cooled with liquid nitrogen was employed for the particle collection: the nanoparticles in the size range from 2-50 nm are extracted from the gas flow by thermophoretic forces and deposited loosely on the surface of the collection device as a powder of low density and no agglomeration. Subsequenly, the nanoparticles are removed from the surface of the cylinder by means of a scraper in the form of a metallic plate. In addition to this cold finger device several techniques known from aerosol science have now been implemented for the use in gas condensation systems such as corona discharge, etc. These methods allow for the continuous operation of the collection device and are better suited for larger scale synthesis of nanopowders. However, these methods can only be used in a system designed for gas flow, i.e. a dynamic vacuum is generated by means of both continuous pumping and gas inlet via mass flow controller. A major advantage over convectional gas flow is the improved control of the particle sizes. It has been found that the particle size distributions in gas flow systems, which are also lognormal, are shifted towards smaller average values with an appreciable reduction of the standard deviation of the distribution. Depending on the flow rate of the He-gas, particle sizes are reduced by 80% and standard deviations by 18%.

The synthesis of nanocrystalline pure metals is relatively straightforward as long as evaporation can be done from refractory metal crucibles (W, Ta or Mo). If metals with high melting points or metals which react with the crucibles, are to be prepared, sputtering, i.e. for W and Zr, or laser or electron beam evaporation has to be used. Synthesis of alloys or intermetallic compounds by thermal evaporation can only be done in the exceptional cases that the vapour pressures of the elements are similar. As an alternative, sputtering from an alloy or mixed target can be

employed. Composite materials such as Cu/Bi or W/Ga have been synthesised by simultaneous evaporation from two separate crucibles onto a rotating collection device. It has been found that excellent intermixing on the scale of the particle size can be obtained.

However, control of the composition of the elements has been difficult and reproducibility is poor. Nanocrystalline oxide powders are formed by controlled postoxidation of primary nanoparticles of a pure metal (e.g. Ti to TiO_2) or a suboxide (e.g. ZrO to ZrO_2). Although the gas condensation method including the variations have been widely employed to prepared a variety of metallic and ceramic materials, quantities have so far been limited to a laboratory scale. The quantities of metals are below 1 g/day, while quantities of oxides can be as high as 20 g/day for simple oxides such as CeO_2 or ZrO_2. These quantities are sufficient for materials testing but not for industrial production. However, it should be mentioned that the scale-up of the gas condensation method for industrial production of nanocrystalline oxides by a company called nanophase technologies has been successful.

Chemical Vapour Condensation (CVC)

As shown schematically in Figure, the evaporative source used in GPC is replaced by a hot wall reactor in the Chemical Vapour Condensation or the CVC process. Depending on the processing parameters nucleation of nanoparticles is observed during chemical vapour deposition (CVC) of thin films and poses a major problem in obtaining good film qualities. The original idea of the novel CVC process which is schematically shown below where, it was intended to adjust the parameter field during the synthesis in order to suppress film formation and enhance homogeneous nucleation of particles in the gas flow. It is readily found that the residence time of the precursor in the reactor determines if films or particles are formed. In a certain range of residence time both particle and film formation can be obtained.

Adjusting the residence time of the precursor molecules by changing the gas flow rate, the pressure difference between the precursor delivery system and the main chamber occurs. Then the temperature of the hot wall reactor results in the fertile production of nanosized particles of metals and ceramics instead of thin films as in CVD processing. In the simplest form a metal organic precursor is introduced into the hot zone of the reactor using mass flow controller. Besides the increased quantities in this continuous process compared to GPC has been demonstrated that a wider range of ceramics including nitrides and carbides can be synthesised. Additionally, more complex oxides such as $BaTiO_3$ or composite structures can be formed as well. Appropriate precursor compounds can be readily found in the CVD literature. The extension to production of nanoparticles requires the determination of a modified parameter field in order to promote particle formation instead of film formation. In addition to the formation of single phase nanoparticles by CVC of a single precursor the reactor allows the synthesis of

1. mixtures of nanoparticles of two phases or doped nanoparticles by supplying two precursors at the front end of the reactor, and

2. coated nanoparticles, i.e., n-ZrO_2 coated with n-Al_2O_3 or vice versa, by supplying a second precursor at a second stage of the reactor. In this case nanoparticles which have

been formed by homogeneous nucleation are coated by heterogeneous nucleation in a second stage of the reactor.

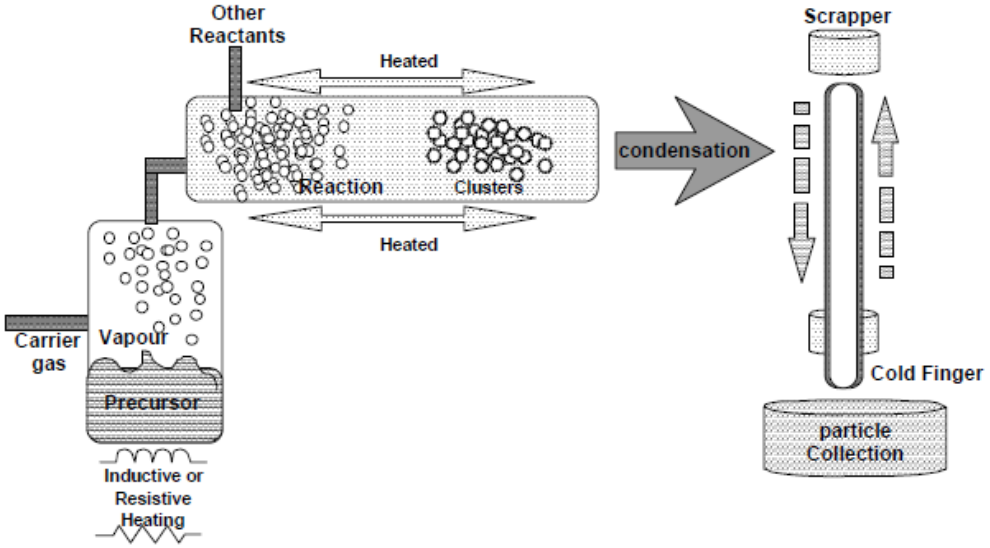


Fig. (10): A schematic of a typical CVC reactor

Because CVC processing is continuous, the production capabilities are much larger than in GPC processing. Quantities in excess of 20 g/hr have been readily produced with a small scale laboratory reactor. A further expansion can be envisaged by simply enlarging the diameter of the hot wall reactor and the mass flow through the reactor.

Sputtered Plasma Processing:

In this method is yet again a variation of the gas-condensation method excepting the fact that the source material is a sputtering target and this target is sputtered using rare gases and the constituents are allowed to agglomerate to produce nanomaterial. Both dc (direct current) and rf (radio-frequency) sputtering has been used to synthesize nanoparticles. Again reactive sputtering or multitarget sputtering has been used to make alloys and/or oxides, carbides, nitrides of materials. This method is specifically suitable for the preparation of ultrapure and non-agglomerated nanoparticles of metal.

Microwave Plasma Processing:

This technique is similar to the previously discussed CVC method but employs plasma instead of high temperature for decomposition of the metal organic precursors. The method uses microwave plasma in a 50 mm diameter reaction vessel made of quartz placed in a cavity connected to a microwave generator. A precursor such as a chloride compound is introduced into the front end of the reactor. Generally, the microwave cavity is designed as a single mode cavity using the TE10 mode in a WR975 waveguide with a frequency of 0.915 GHz. The major advantage of the plasma assisted pyrolysis in contrast to the thermal activation is the low temperature reaction which reduces the tendency for agglomeration of the primary particles. This is also true in the case of plasma-CVD processes. Additionally, it has been shown that by

introducing another precursor into a second reaction zone of the tubular reactor, e.g. by splitting the microwave guide tubes, the primary particles can be coated with a second phase. For example, it has been demonstrated that ZrO_2 nanoparticles can be coated by Al_2O_3. In this case the inner ZrO_2 core is crystalline, while the Al2O3 coating is amorphous. The reaction sequence can be reversed with the result that an amorphous $Al2O_3$ core is coated with crystalline ZrO_2. While the formation of the primary particles occurs by homogeneous nucleation, it can be easily estimated using gas reaction kinetics that the coating on the primary particles grows heterogeneously and that homogeneous nucleation of nanoparticles originating from the second compound has a very low probability. A schematic representation of the particle growth in plasma's is given below:

Particle precipitation aided CVD:

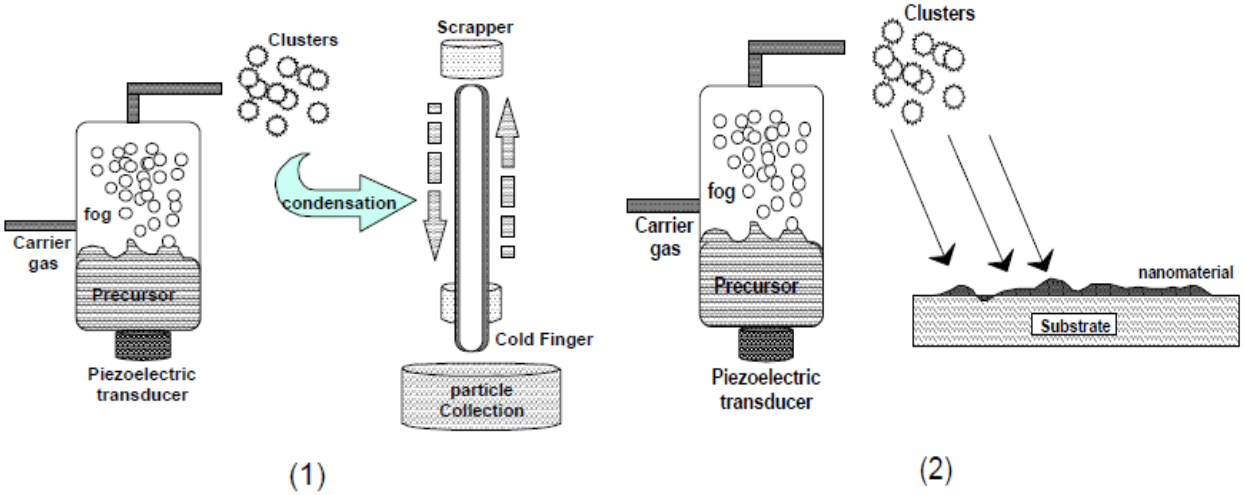


Fig. (11): Schematic representation of (1) nanoparticle, and (2) particulate film formation.

In another variation of this process, colloidal clusters of materials are used to prepare nanoparticles. The CVD reaction conditions are so set that particles form by condensation in the gas phase and collect onto a substrate, which is kept under a different condition that allows heterogeneous nucleation. By this method both nanoparticles and particulate films can be prepared. An example of this method has been used to form nanomaterials eg. SnO_2, by a method called pyrosol deposition process, where clusters of tin hydroxide are transformed into small aerosol droplets, following which they are reacted onto a heated glass substrate.

Laser ablation:

Laser ablation has been extensively used for the preparation of nanoparticles and particulate films. In this process a laser beam is used as the primary excitation source of ablation for generating clusters directly from a solid sample in a wide variety of applications. The small dimensions of the particles and the possibility to form thick films make this method quite an efficient tool for the production of ceramic particles and coatings and also an ablation source for analytical applications such as the coupling to induced coupled plasma emission spectrometry, ICP, the formation of the nanoparticles has been explained following a liquefaction process which generates an aerosol, followed by the cooling/solidification of the droplets which results in the formation of fog. The general dynamics of both the aerosol and the fog favors the

aggregation process and micrometer-sized fractal-like particles are formed. The laser spark atomizer can be used to produce highly mesoporous thick films and the porosity can be modified by the carrier gas flow rate. ZrO_2 and SnO_2 nanoparticulate thick films were also synthesized successfully using this process with quite identical microstructure. Synthesis of other materials such as lithium manganate, silicon and carbon has also been carried out by this technique.

Properties of Nanomaterials

Nanomaterials have the structural features in between of those of atoms and the bulk materials. While most microstructured materials have similar properties to the corresponding bulk materials, the properties of materials with nanometer dimensions are significantly different from those of atoms and bulks materials. This is mainly due to the nanometer size of the materials which render them: (i) large fraction of surface atoms; (ii) high surface energy; (iii) spatial confinement; (iv) reduced imperfections, which do not exist in the corresponding bulk materials.

Due to their small dimensions, nanomaterials have extremely large surface area to volume ratio, which makes a large to be the surface or interfacial atoms, resulting in more “surface” dependent material properties. Especially when the sizes of nanomaterials are comparable to length, the entire material will be affected by the surface properties of nanomaterials. This in turn may enhance or modify the properties of the bulk materials. For example, metallic nanoparticles can be used as very active catalysts. Chemical sensors from nanoparticles and nanowires enhanced the sensitivity and sensor selectivity. The nanometer feature sizes of nanomaterials also have spatial confinement effect on the materials, which bring the quantum effects.

The energy band structure and charge carrier density in the materials can be modified quite differently from their bulk and in turn will modify the electronic and optical properties of the materials. For example, lasers and light emitting diodes (LED) from both of the quantum dots and quantum wires are very promising in the future optoelections. High density information storage using quantum dot devices is also a fast developing area. Reduced imperfections are also an important factor in determination of the properties of the nanomaterials. Nanosturctures and Nanomaterials favors of a self-purification process in that the impurities and intrinsic material defects will move to near the surface upon thermal annealing. This increased materials perfection affects the properties of nanomaterials. For example, the chemical stability for certain nanomaterials may be enhanced, the mechanical properties of nanomaterials will be better than the bulk materials. The superior mechanical properties of carbon nanotubes are well known. Due to their nanometer size, nanomaterials are already known to have many novel properties. Many novel applications of the nanomaterials rose from these novel properties have also been proposed.

Optical properties

One of the most fascinating and useful aspects of nanomaterials is their optical properties. Applications based on optical properties of nanomaterials include optical detector, laser, sensor, imaging, phosphor, display, solar cell, photocatalysis, photoelectrochemistry and biomedicine.

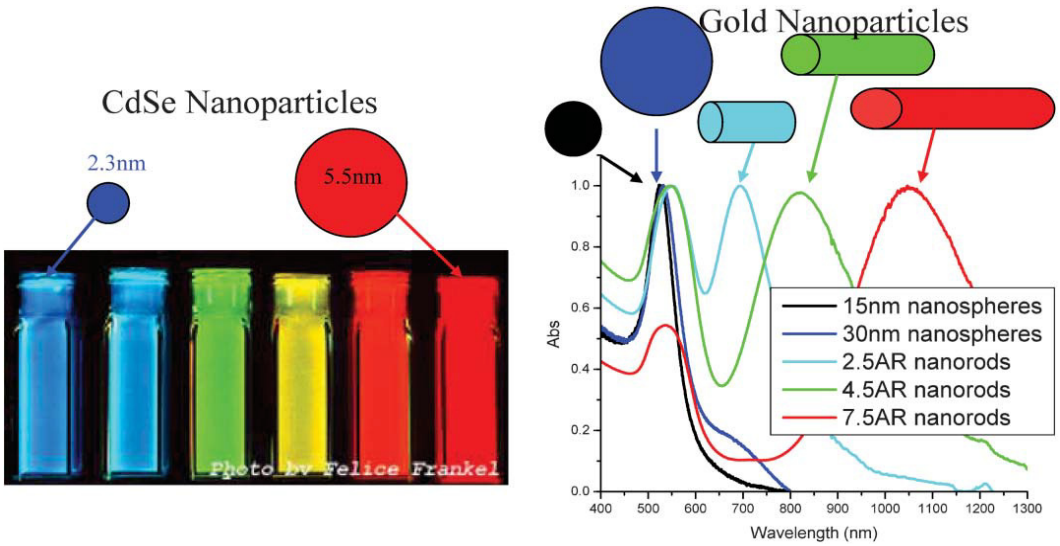


Fig. (12): Fluorescence emission of (CdSe) ZnS quantum dots of various sizes and absorption spectra of various sizes and shapes of gold nanoparticles (Chem. Soc. Rev., 2006, 35, 209–217).

The optical properties of nanomaterials depend on parameters such as feature size, shape, surface characteristics, and other variables including doping and interaction with the surrounding environment or other nanostructures. Likewise, shape can have dramatic influence on optical properties of metal nanostructures. Fig. (12) Exemplifies the difference in the optical properties of metal and semiconductor nanoparticles. With the CdSe semiconductor nanoparticles, a simple change in size alters the optical properties of the nanoparticles. When metal nanoparticles are enlarged, their optical properties change only slightly as observed for the different samples of gold nanospheres. However, when an anisotropy is added to the nanoparticle, such as growth of nanorods, the optical properties of the nanoparticles change dramatically.

Electrical Properties

Electrical Properties of Nanoparticles" discuss about fundamentals of electrical conductivity in nanotubes and nanorods, carbon nanotubes, photoconductivity of nanorods, electrical conductivity of nanocomposites. One interesting method which can be used to demonstrate the steps in conductance is the mechanical thinning of a nanowire and measurement of the electrical current at a constant applied voltage. The important point here is that, with decreasing diameter of the wire, the number of electron wave modes contributing to the electrical conductivity is becoming increasingly smaller by well-defined quantized steps.

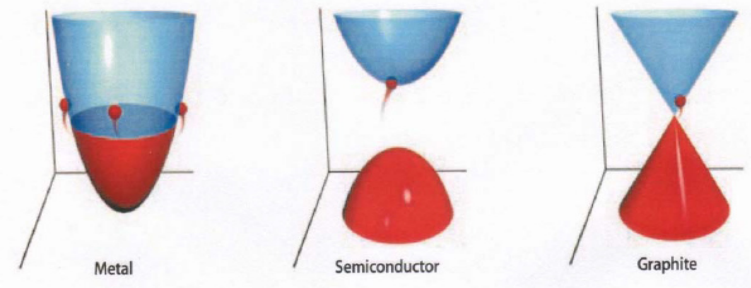


Fig. (13): Electrical behavior of naotubes (P. G. Collins and Ph. Avouris, *Scientific American*, 62, 2000, 283).

In electrically conducting carbon nanotubes, only one electron wave mode is observed which transport the electrical current. As the lengths and orientations of the carbon nanotubes are different, they touch the surface of the mercury at different times, which provides two sets of information: (i) the influence of carbon nanotube length on the resistance; and (ii) the resistances of the different nanotubes. As the nanotubes have different lengths, then with increasing protrusion of the fiber bundle an increasing number of carbon nanotubes will touch the surface of the mercury droplet and contribute to the electrical current transport.

Mechanical Properties

"Mechanical Properties of Nanoparticles" deals with bulk metallic and ceramic materials, influence of porosity, influence of grain size, superplasticity, filled polymer composites, particle-filled polymers, polymer-based nanocomposites filled with platelets, carbon nanotube-based composites. The discussion of mechanical properties of nanomaterials is, in to some extent, only of quite basic interest, the reason being that it is problematic to produce macroscopic bodies with a high density and a grain size in the range of less than 100 nm. However, two materials, neither of which is produced by pressing and sintering, have attracted much greater interest as they will undoubtedly achieve industrial importance.

These materials are polymers which contain nanoparticles or nanotubes to improve their mechanical behaviors, and severely plastic-deformed metals, which exhibit astonishing properties. However, because of their larger grain size, the latter are generally not accepted as nanomaterials. Experimental studies on the mechanical properties of bulk nanomaterials are generally impaired by major experimental problems in producing specimens with exactly defined grain sizes and porosities. Therefore, model calculations and molecular dynamic studies are of major importance for an understanding of the mechanical properties of these materials.

Filling polymers with nanoparticles or nanorods and nanotubes, respectively, leads to significant improvements in their mechanical properties. Such improvements depend heavily on the type of the filler and the way in which the filling is conducted. The latter point is of special importance, as any specific advantages of a nanoparticulate filler may be lost if the filler forms aggregates, thereby mimicking the large particles. Particulate-filled polymer-based nanocomposites exhibit a broad range of failure strengths and strains. This depends on the shape of the filler, particles or platelets, and on the degree of agglomeration. In this class of material, polymers filled with silicate platelets exhibit the best mechanical properties and are of the greatest economic relevance. The larger the particles of the filler or agglomerates, the poorer are the properties obtained. Although, potentially, the best composites are those filled with nanofibers or nanotubes, experience teaches that sometimes such composites have the least ductility. On the other hand, by using carbon nanotubes it is possible to produce composite fibers with extremely high strength and strain at rupture. Among the most exciting nanocomposites are the polymer-ceramic nanocomposites, where the ceramic phase is platelet-shaped. This type of composite is preferred in nature, and is found in the structure of bones, where it consists of crystallized mineral platelets of a few nanometers thickness that are bound together with collagen as the matrix. Composites consisting of a polymer matrix and defoliated phyllosilicates exhibit excellent mechanical and thermal properties.

Magnetic properties

Bulk gold and Pt are non-magnetic, but at the nano size they are magnetic. Surface atoms are not only different to bulk atoms, but they can also be modified by interaction with other chemical species, that is, by capping the nanoparticles. This phenomenon opens the possibility to modify the physical properties of the nanoparticles by capping them with appropriate molecules. Actually, it should be possible that non-ferromagnetic bulk materials exhibit ferromagnetic-like behavior when prepared in nano range. One can obtain magnetic nanoparticles of Pd, Pt and the surprising case of Au (that is diamagnetic in bulk) from non-magnetic bulk materials. In the case of Pt and Pd, the ferromagnetism arises from the structural changes associated with size effects.

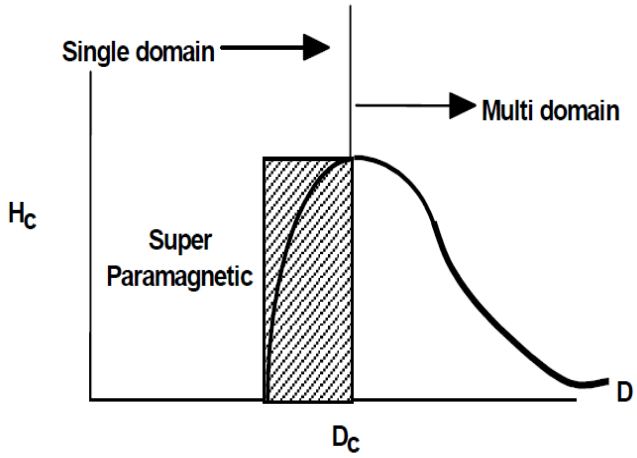


Fig. (14): Magnetic properties of nanostrucutred materials

However, gold nanoparticles become ferromagnetic when they are capped with appropriate molecules: the charge localized at the particle surface gives rise to ferromagnetic-like behavior. Surface and the core of Au nanoparticles with 2 nm in diameter show ferromagnetic and paramagnetic character, respectively. The large spin-orbit coupling of these noble metals can yield to a large anisotropy and therefore exhibit high ordering temperatures. More surprisingly, permanent magnetism was observed up to room temperature for thiol-capped Au nanoparticles. For nanoparticles with sizes below 2 nm the localized carriers are in the 5d band. Bulk Au has an extremely low density of states and becomes diamagnetic, as is also the case for bare Au nanoparticles. This observation suggested that modification of the d band structure by chemical bonding can induce ferromagnetic like character in metallic clusters.

Selected Application of nanomaterials

Nanomaterials having wide range of applications in the field of electronics, fuel cells, batteries, agriculture, food industry, and medicines, etc... It is evident that nanomaterials split their conventional counterparts because of their superior chemical, physical, and mechanical properties and of their exceptional formability.

Fuel cells:

A fuel cell is an electrochemical energy conversion device that converts the chemical energy from fuel (on the anode side) and oxidant (on the cathode side) directly into electricity. The heart

of fuel cell is the electrodes. The performance of a fuel cell electrode can be optimized in two ways; by improving the physical structure and by using more active electro catalyst. A good structure of electrode must provide ample surface area, provide maximum contact of catalyst, reactant gas and electrolyte, facilitate gas transport and provide good electronic conductance. In this fashion the structure should be able to minimize losses.

Carbon nanotubes - Microbial fuel cell

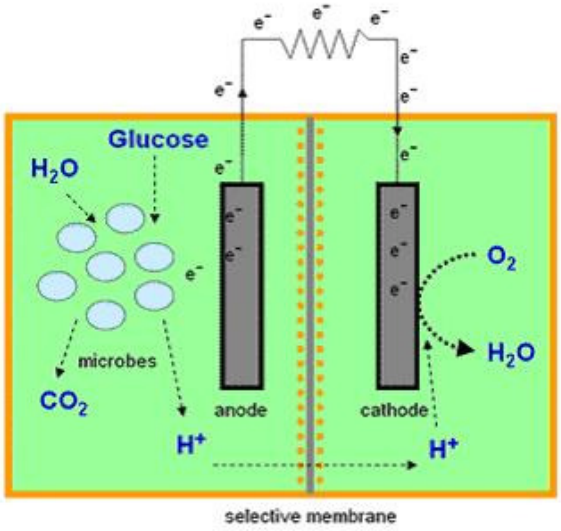


Fig. (15): Schematic representation of microbial fuel cell

Microbial fuel cell is a device in which bacteria consume water-soluble waste such as sugar, starch and alcohols and produces electricity plus clean water. This technology will make it possible to generate electricity while treating domestic or industrial wastewater. Microbial fuel cell can turn different carbohydrates and complex substrates present in wastewaters into a source of electricity. The efficient electron transfer between the microorganism and the anode of the microbial fuel cell plays a major role in the performance of the fuel cell. The organic molecules present in the wastewater posses a certain amount of chemical energy, which is released when converting them to simpler molecules like CO_2. The microbial fuel cell is thus a device that converts the chemical energy present in water-soluble waste into electrical energy by the catalytic reaction of microorganisms.

Carbon nanotubes (CNTs) have chemical stability, good mechanical properties and high surface area, making them ideal for the design of sensors and provide very high surface area due to its structural network. Since carbon nanotubes are also suitable supports for cell growth, electrodes of microbial fuel cells can be built using of CNT. Due to three-dimensional architectures and enlarged electrode surface area for the entry of growth medium, bacteria can grow and proliferate and get immobilized. Multi walled CNT scaffolds could offer self-supported structure with large surface area through which hydrogen producing bacteria (e.g., E. coli) can eventually grow and proliferate. Also CNTs and MWCNTs have been reported to be biocompatible for different eukaryotic cells. The efficient proliferation of hydrogen producing bacteria throughout an electron conducting scaffold of CNT can form the basis for the potential application as electrodes in MFCs leading to efficient performance.

Catalysis

Higher surface area available with the nanomaterial counterparts, nano-catalysts tend to have exceptional surface activity. For example, reaction rate at nano-aluminum can go so high, that it is utilized as a solid-fuel in rocket propulsion, whereas the bulk aluminum is widely used in utensils. Nano-aluminum becomes highly reactive and supplies the required thrust to send off pay loads in space. Similarly, catalysts assisting or retarding the reaction rates are dependent on the surface activity, and can very well be utilized in manipulating the rate-controlling step.

Phosphors for High-Definition TV

The resolution of a television, or a monitor, depends greatly on the size of the pixel. These pixels are essentially made of materials called "phosphors," which glow when struck by a stream of electrons inside the cathode ray tube (CRT). The resolution improves with a reduction in the size of the pixel, or the phosphors. Nanocrystalline zinc selenide, zinc sulfide, cadmium sulfide, and lead telluride synthesized by the sol-gel techniques are candidates for improving the resolution of monitors. The use of nanophosphors is envisioned to reduce the cost of these displays so as to render high-definition televisions (HDTVs) and personal computers affordable to be purchase.

Next-Generation Computer Chips

The microelectronics industry has been emphasizing miniaturization, whereby the circuits, such as transistors, resistors, and capacitors, are reduced in size. By achieving a significant reduction in their size, the microprocessors, which contain these components, can run much faster, thereby enabling computations at far greater speeds. However, there are several technological impediments to these advancements, including lack of the ultrafine precursors to manufacture these components; poor dissipation of tremendous amount of heat generated by these microprocessors due to faster speeds; short mean time to failures (poor reliability), etc. Nanomaterials help the industry break these barriers down by providing the manufacturers with nanocrystalline starting materials, ultra-high purity materials, materials with better thermal conductivity, and longer-lasting, durable interconnections (connections between various components in the microprocessors).

For example: Nanowires for junctionless transistors:

Transistors are made so tiny to reduce the size of sub assemblies of electronic systems and make smaller and smaller devices, but it is difficult to create high-quality junctions. In particular, it is very difficult to change the doping concentration of a material over distances shorter than about 10 nm. Researchers have succeeded in making the junctionless transistor having nearly ideal electrical properties. It could potentially operate faster and use less power than any conventional transistor on the market today. The device consists of a silicon nanowire in which current flow is perfectly controlled by a silicon gate that is separated from the nanowire by a thin insulating layer. The entire silicon nanowire is heavily n-doped, making it an excellent conductor. However, the gate is p-doped and its presence has the effect of depleting the number of electrons in the region of the nanowire under the gate. The device also has near-ideal electrical properties

and behaves like the most perfect of transistors without suffering from current leakage like conventional devices and operates faster and using less energy.

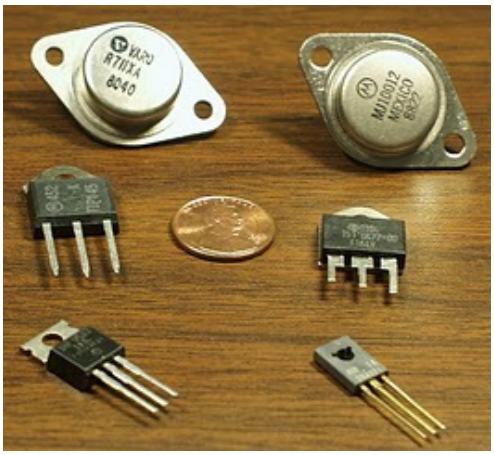


Fig. (16): Silicon nanowires in junctionless transistors

Elimination of Pollutants

Nanomaterials possess extremely large grain boundaries relative to their grain size. Hence, they are very active in terms of their chemical, physical, and mechanical properties. Due to their enhanced chemical activity, nanomaterials can be used as catalysts to react with such noxious and toxic gases as carbon monoxide and nitrogen oxide in automobile catalytic converters and power generation equipment to prevent environmental pollution arising from burning gasoline and coal.

Sun-screen lotion

Prolonged UV exposure causes skin-burns and cancer. Sun-screen lotions containing nano-TiO_2 provide enhanced sun protection factor (SPF) while eliminating stickiness. The added advantage of nano skin blocks (ZnO and TiO_2) arises as they protect the skin by sitting onto it rather than penetrating into the skin. Thus they block UV radiation effectively for prolonged duration. Additionally, they are transparent, thus retain natural skin color while working better than conventional skin-lotions.

Sensors

Sensors rely on the highly active surface to initiate a response with minute change in the concentration of the species to be detected. Engineered monolayers (few Angstroms thick) on the sensor surface are exposed to the environment and the peculiar functionality (such as change in potential as the CO/anthrax level is detected) is utilized in sensing.

Disadvantages of Nanomaterials

- Instability of the particles - Retaining the active metal nanoparticles is highly challenging, as the kinetics associated with nanomaterials is rapid. In order to retain

nanosize of particles, they are encapsulated in some other matrix. Nanomaterials are thermodynamically metastable and lie in the region of high-energy local-minima. Hence they are prone to attack and undergo transformation. These include poor corrosion resistance, high solubility, and phase change of nanomaterials. This leads to deterioration in properties and retaining the structure becomes challenging.

Fine metal particles act as strong explosives owing to their high surface area coming in direct contact with oxygen. Their exothermic combustion can easily cause explosion.

- Impurity - Because nanoparticles are highly reactive, they inherently interact with impurities as well. In addition, encapsulation of nanoparticles becomes necessary when they are synthesized in a solution (chemical route). The stabilization of nanoparticles occurs because of a non-reactive species engulfing the reactive nano-entities. Thereby, these secondary impurities become a part of the synthesized nanoparticles, and synthesis of pure nanoparticles becomes highly difficult. Formation of oxides, nitrides, etc can also get aggravated from the impure environment/ surrounding while synthesizing nanoparticles. Hence retaining high purity in nanoparticles can become a challenge hard to overcome.

- Biologically harmful - Nanomaterials are usually considered harmful as they become transparent to the cell-dermis. Toxicity of nanomaterials also appears predominant owing to their high surface area and enhanced surface activity. Nanomaterials have shown to cause irritation, and have indicated to be carcinogenic. If inhaled, their low mass entraps them inside lungs, and in no way they can be expelled out of body. Their interaction with liver/blood could also prove to be harmful (though this aspect is still being debated on).

  Difficulty in synthesis, isolation and application - It is extremely hard to retain the size of nanoparticles once they are synthesized in a solution. Hence, the nanomaterials have to be encapsulated in a bigger and stable molecule/material.

  Hence free nanoparticles are hard to be utilized in isolation, and they have to be interacted for intended use via secondary means of exposure. Grain growth is inherently present in nanomateirals during their processing. The finer grains tend to merge and become bigger and stable grains at high temperatures and times of processing.

- Recycling and disposal - There are no hard-and-fast safe disposal policies evolved for nanomaterials. Issues of their toxicity are still under question, and results of exposure experiments are not available. Hence the uncertainty associated with affects of nanomaterials is yet to be assessed in order to develop their disposal policies.

References

1. Nanomaterials – B. Viswanathan, published by Narosa Publishing House

2. Optical properties and spectroscopy of nanomaterials - Jin Zhng Zhang, published by World Scientific Publishing Co. Pte. Ltd.

3. Anisotropic nanomaterials: structure, growth, assembly, and functions, P. R. Sajanlal, T. S. Sreeprasad, A. K. Samal and T. Pradeep, NANO REVIEWS, vol 2, (2011).

4. Physical Properties of Nanomaterials, Juh Tzeng Lue, *Encyclopedia of Nanoscience and Nanotechnology,* Volume X: Pages (1–46).

5. Nanomaterials – An introduction to synthesis, properties and application, Environmental Engineering and Management Journal, 2008, Vol. 7, No.6, 865-870.