20.1 Molecular Biology in Evolution

Molecular biology is fast developing during the last few decades, knowledge of which has transformed every field of biology, including evolutionary biology. Field of molecular biology chiefly deals with—

(i) structures of DNA, RNA and proteins,

(ii) processes like replication, transcription, and

(iii) regulation of levels of gene products.

Information on molecular structures and processes, and the techniques developed by molecular biology have recently been widely used in understanding different aspects of evolution. **Molecular evolution** is, therefore, evolution at the molecular level of DNA and protein sequences. Study of molecular evolution allows us to know how DNA and proteins evolve, which is the molecular form of the question – how genes and organisms evolve?

Field of molecular evolution is multidisciplinary. It uses data and insight from molecular cell biology. Tools of molecular biology have been widely developed and used from the 1970s. Naturally, biologists have shown interest in the study of how biologically important molecules change over time. Now the ability to clone, sequence and hybridize DNA has opened up a new window to the evolutionary biologists who only perceived that gene evolved through mutation, duplication, conversion and transposition. After Darwinism and synthetic theory, evolutionary studies have been flooded by an abundance of parameters through which evolutionary theories can be tested and evolutionary rates can be measured.

Genomes are historical records of evolution and, no doubt, much more powerful than paleontological records. Molecular analyses of genomes throw much light on the dynamics behind evolutionary processes and allow us to reconstruct the chronology of organic changes. The approach not only helps us with to decipher the evolutionary history on this planet, but also provide us with a classification of living organisms according to their true phylogenetic relationships. Relationships previously unimagined between organisms are becoming apparent to us; previously closely related organisms are found to have no actual relationships, rather are distantly related; again, relationship among some organisms are being established, no matter how greatly they differ in phenotypic characteristics. Following the approach of molecular evolution, we are able to construct a **universal tree of life**, which is the root of all systematics.

In the present chapter, we shall restrict our discussion to three aspects—

(i) Principles of molecular evolution studies.

(ii) How molecules with new functions arise?

(iii) How phylogenetic relationships between molecules and organism are determined?

20.2 Methods of Molecular Evolution Studies

Among many methods of molecular biology, study of molecular evolution is primarily based on three methods. They include **polymerase chain reaction (PCR)**, **DNA sequencing**, and techniques for localizing or isolating gene and gene products like RNA and proteins. A brief idea of these techniques is summarized below.

Polymerase chain reaction allows amplification of a selected DNA sequence

When a particular gene is identified, multiple copies of it are required to work with it. One way to do this is to insert that DNA fragment into a bacterium, which by repeated cell division can produce multiple copies. But a more direct approach is polymerase chain reaction or simply **PCR**, which very successfully clone a given DNA sequence in vitro. Kary Mullium received the Nobel Prize in 1993 for inventing PCR technology. PCR technology provides shortcut for many cloning and sequencing applications in molecular biology. This procedure permits us to obtain definitive structural data on genes and DNA sequences, even when very small amount of DNA are available.

DNA sequencing helps us to know the ultimate fine structure map of a gene

The ultimate structure of a gene is its nucleotide-pairs sequence. Each gene has its own specific sequence of nucleotide-pairs that is unique for its function. After 1975, sequencing an entire chromosome became a very popular research and today, sequencing is a routine laboratory procedure. Complete genomes of many viruses and bacteria, yeast, mitochondria, chloroplast, nematode *Caenorhabditis*, plant *Arbidopsis*, and animals like

Drosophila and human have been determined. Many genome projects are in progress.

Researchers have invented two different procedures of DNA sequencing. The first is Maxam and Gilbert procedure and the second is Sanger sequencing procedure. Today all large scale DNA sequencing is done with automated sequencing machine that adopt the Sanger procedure with modification. The automated machine can perform sample loading, electrophoresis, data collection and data analysis simultaneously, which are fully automated. Such a machine can determine more than 100,000 nucleotides of sequence per day but may take 8 years to determine the sequence of human genome containing about 3 billion nucleotide pairs.

Localizing and isolating gene and gene products

RNA transcripts and proteins are generally considered as gene products. Gel electrophoresis provides a powerful tool for separation of different macromolecules like DNA itself and its products. DNA molecules are separated and then identified by **Southern Blot** procedure, which uses agarose gel as molecular sieves. Then DNA fragments are transferred to nitrocellulose membranes, which are then identified by hybridization to labelled DNA. RNA molecules are also separated and analyzed by the same procedure adopted in Southern Blot. But RNA molecules are kept denatured during electrophoresis and the procedure is called **Northern Blot**. This procedure is very helpful in studying gene expression, particularly in finding whether a gene is transcribed in all tissues of an organism or only in certain tissues.

Separation and characterization of proteins involves **polyacrylamide gel electrophoresis** (PAGE). After electrophoresis, separated proteins (or polypeptides) are transferred to nitrocellulose membrane like in Southern Blot technique, and the procedure is called **Western Blot**. Finally, specific protein of interest is identified using antibodies. Western blot procedure is an important tool for separation, identification and characterization of proteins.

20.3 Information obtained from the methods used in molecular evolution studies

The mentioned studies provide us with at least three kinds of information as mentioned below :

I. Estimation of genetic structure of population and species

Small differences in DNA sequences are common in most populations, and they often serve as molecular (genetic) markers. Molecular markers are used to analyze traditional problems in evolution like—

(i) Estimating genetic variations within and among populations.

(ii) Estimating population structure, gene flow and breeding systems.

(iii) Describing genetic differences among species.

(iv) Genealogical analysis of gene tree (or haplotype) to know about histories of population size, gene flow and selection.

(v) Obtaining information on phylogenetic relationships among species and higher taxa.

II. Evidence on the evolution of phenotype

Through same studies, genes are identified, which affect particular morphological, physiological or behavioral characters. These, in turn, give full understanding of the process for phenotypic evolution. In those studies, identification of genes is of utmost importance but is more difficult than estimation of genetic structure of population. The field is gaining more attention and studies in this field include—

(i) Identification of genes underling polygenic variations, and

(ii) Comparative studies of genes, which govern developmental processes.

III. Evolution of genes and genomes

DNA sequences and structure of genomes are distinctive features of organisms, like morphology and behavior. Therefore, these two attributes are studied in their own right. Change in genes and genomes provide insight and information about the mechanism and evolutionary process that have governed the rates and patterns of variations. However, most studies of evolution of genes and genomes made so far have relied on analyzing variations within and among populations and species.

In the research of molecular evolution, a researcher may obtain data himself/herself through

experimental techniques. Otherwise, he/she may analyze databases like GenBank, in which sequence data already published are stored. There are many methods to analyze such data, such as—

1. Phylogenetic analysis is one of the most important approaches, through which the genealogical analysis of sequences, or of the populations or species from where samples are taken, is estimated.
2. Methods of population genetics also analyze sequence data, where observed data are frequently compared with the expected ones derived from mathematical models.

Phylogenetic analysis provides strong evidence for many processes, which are important in the study of molecular evolution. Two of such processes are resurrecting extinct genes and gene transfer.

Resurrection of functional ancestral gene can be possible by phylogenetic analysis. Retrotransposons (or retroposons) are transposable elements (please see sectin 15.12), which encode reverse transcriptase. RNA transcripts of a retroposon are reverse-transcribed into DNA copies, which are inserted into various sites of genome of all organisms. Mammalian genome carries numerous copies of retroposons some of which have active promoters and some have inactive promoters. It is thought that inactive promoters had evolved through various mutations on sequences over the course of about 6 million years. Using phylogenetic analysis of 30 different inactive promoters of mice, Nil Adey and co-authors (1994) have concluded that inactive sequences had evolved from functional ancestral promoters. They also constructed a best estimate of the ancestral sequence.

Gene transfers result in recombination unless there are identical genealogies. If genealogies of different genes differ, recombination and hence gene transfers must be suspected. Exchange of genetic material can be suspected by analysis of gene sequences, which in fact, must be frequent for prokaryotes and eukaryotes. Horizontal gene transfer (section 12.13) is rather common in prokaryotes, as for example, some bacteria (like *Neisseria gonorrheae*, causing Gonorrhea) seems to have acquired resistance to penicillin by horizontal gene transfer from non-pathogenic bacteria. In eukaryotes, evidence of horizontal gene transfer are few. A virogene (gene derived by reverse transcription of a viral gene) in the genome of some cats is very similar in sequence to the same gene in baboons (Fig. 20.2). These two animals are otherwise not very closely related; still this is one documented case of horizontal gene transfer in eukaryotes.

20.4 PRINCIPLES OF MOLECULAR EVOLUTION STUDIES

Variation in DNA sequences were were first known through **Restriction Fragment-length Polymorphism** (RFLP). Later, variations in the form of full nucleotide sequences were also available. Naturally, evolutionary geneticists have focussed on the evolutionary processes that have important roles in sequence evolution. At the same time, the debate whether random genetic drift or natural selection is responsible for variation within population has focussed on DNA sequence variation. Study of DNA sequence variation is primarily based on substitution, that is, partial or complete replacement of a nucleotide or longer sequences by another throughout the population. In this context, we should refer to **mutation**, which is defined as change in single (or more) gene copy that arises through mistakes in DNA replication or repair process. Substitutions are, of course, mutations, which have passed through the filter of selection.

20.4.1 Nucleotide Substitution in DNA Sequences

An important outcome of studies in molecular evolution is that, patterns and rate of substitution differ between different parts of same gene. From the knowledge of amino acid sequences in proteins from a variety of organisms, it has become apparent that some amino acid differences are more likely to be observed between two homologous proteins, which have a common ancestor. Moreover, replacing amino acids have similar chemical characters as of those amino acids present in the ancestral protein.

Amino acids that are similar chemically, tend to have similar codons. Thus to convert them, minimum change is required at DNA level. As for example, codon for leucine is CUU, which can be changed to code for isoleucine (AUU) by a single base pair change. But if the codon for leucine is required to convert into a code for asparagines,

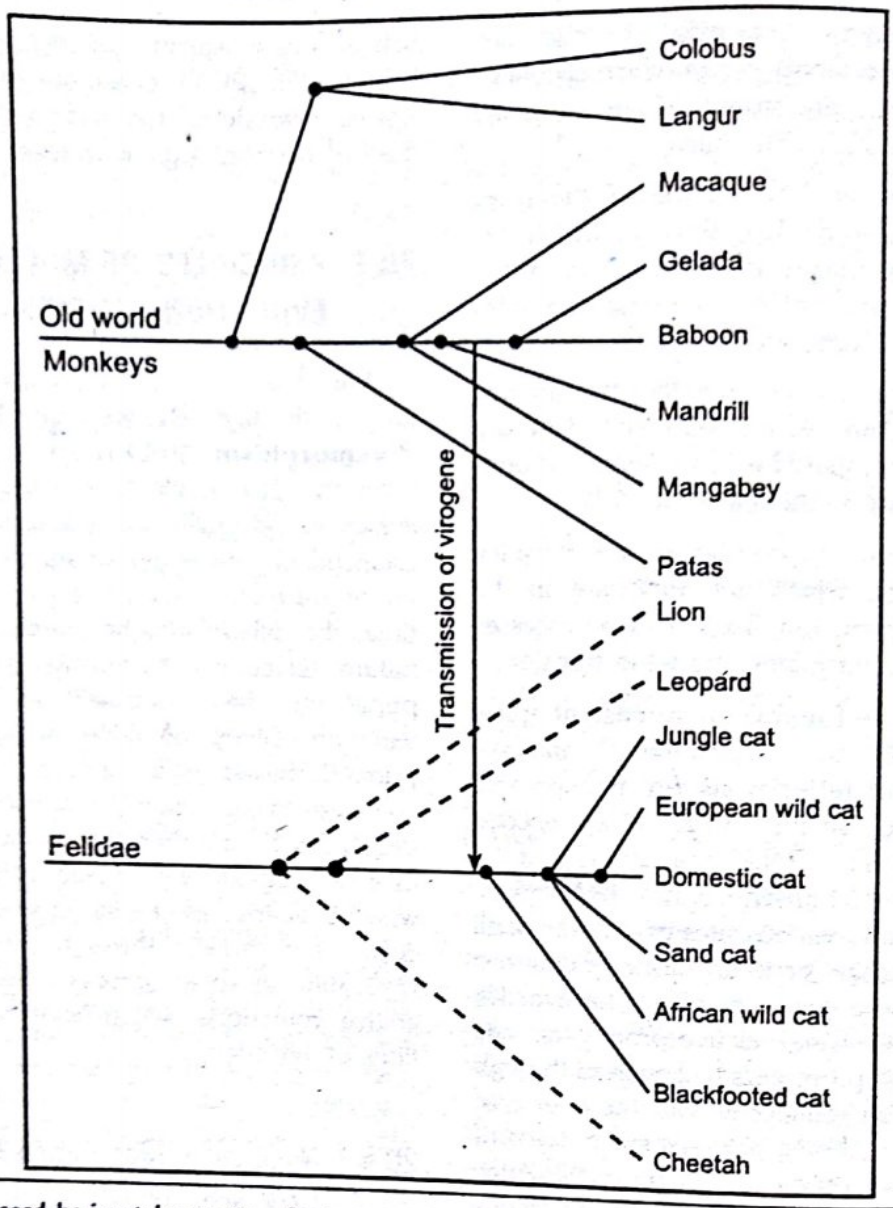

Figure 20.2 : Proposed horizontal gene transfer between Old World monkeys and Felidae. Transfer of virogene probably occurred from primates to cats after divergence of lion, cheetah and leopard. Solid lines lead to species having similar sequence of virogene. Broken lines lead to species lacking the sequence.

(AAU)—a dissimilar amino acid, two base pair change are needed. Because amino acid and nucleotide changes are rare, those involving minimum change are more likely to occur. Again, action of natural selection has caused proteins to have amino acid sequences, which are optimally suited for their role. A drastic change in primary structure of protein may have a deleterious effect on its function, which should not escape the scrutiny of natural selection.

To analyze changes at both amino acid and nucleotide levels between two or more gene sequences, the first task is to make an alignment of all sequences. Alignments are made with the help of computer programs from which evolution of sequences can be predicted. A best possible alignment can be made that reflects the true ancestral relationship of a protein or gene. As for example, alignments of two sequences can be interpreted as follows—

Alignment 1 ⟶ - A T A C C G -

Alignment 2 ⟶ - A T – C A G -

(i) Four of six nucleotides at the position 1, 2, 4 and 6 have not changed from their common ancestor.

(ii) A substitution has occurred at position 5.

(iii) An insertion (for alignment 1) or a deletion (for alignment 2) has occurred at position 3.

Long periods of evolution altered sequences drastically leaving little common between them, which make alignment difficult. Therefore, studies are based on approximately true alignments or **optimal alignments**. In optimal alignments, gaps are inserted to maximize similarity between sequences. But some gaps, which may be insertion in one sequence or deletion in another are difficult to fill up, and are called **indels**. Optimal alignment through computer programming seeks maximum number of matching amino acids or nucleotides between sequences and minimum number of indels.

Measuring rate of nucleotide substitution

In all molecular evolution analysis, a most important parameter is to measure the number of substitution that occurred in two sequences since they last shared a common ancestor. If the number of substitution per site is K and divergence time is T, then the rate of substitution (r) is easily calculated as $r = K/2T$.

Substitutions are likely to occur simultaneously and independently in two sequences, hence the T is doubled in the equation. To estimate substitution rate, data must be obtained at least from two species. Then comparing substitution rates within and between genes, we can have an idea about mechanisms involved in molecular change. If evolutionary rates of several species are similar, substitution rates can help to predict the date of evolutionary event, though there is no physical event.

20.4.2 Types and Rates of Substitutions within Gene, i.e. Intraspecific Variations

In general, substitutions are categorized into two - **transition** and **transversion**. Substitution of one purine by another or one pyrimidine by another in known as transition. Again, substitution of one purine by another pyrimidine or vice versa is called transversion (detail in section 15.6). In recent years it has been found that DNA sequences (both nuclear and mitochondrial) of several species transitions outnumbered transversions. Thus transitions occur at a much faster rate and accumulate quicker than transversions.

Studies of molecular evolution have shown that different parts of genes evolve at different rates. A typical eukaryotic gene has several components. Some nucleotides specify the amino acid sequences of the product protein and are called **coding sequences**. **Non-coding sequences** do not code for amino acids and they include introns, leader and trailer regions, and flanking regions. Additionally there are **pseudogenes**. We shall consider substitution at each of these regions.

Substitution at coding sequences – synonymous and non-synonymous substitutions

A substitution of coding sequence at nucleotide level that does not change the resulting amino acid sequence of a protein is called **synonymous substitution**. Contrarily, a substitution that results in a change in amino acid sequence of a protein is known as **non-synonymous substitution**. Therefore, synonymous substitution is called **silent** while non-synonymous substitution is called **replacement**. Synonymous and non-synonymous substitutions are likely to occur with equal frequencies. But in fact, rate of synonymous substitutions is five times greater than the observed rate of non-synonymous substitutions. Because non-synonymous substitutions change the resultant protein, they are detrimental to fitness and are eliminated by natural selection. Synonymous substitutions are much less detrimental and are tolerated by selection.

Substitutions at non-coding sequences

Non-coding sequences include introns, leader and trailer regions, which are transcribed, and flanking regions, which are not transcribed. **Introns** are intervening DNA sequences in eukaryotes that must be excised from the primary gene transcript to convert it into mature RNA transcript. Usually sequences in introns are not used to translate into amino acid sequences and therefore, they are spliced out properly. But some introns occasionally code for proteins in some tissues, and not for others, due to alternate splicing. Then substitutions in introns may not be undetected by natural selection. Naturally, rate of substitutions in introns is high but not as high as in synonymous changes.

Leader region is located at the 5′ end of mRNA molecule, from the 5′ terminus to the translation initiation codon. **Trailer region** is the

part of mRNA molecule from translation termination codon to the end of the 3′ terminus of mRNA. Both these regions are transcribed but not translated, and they provide important signals for processing and translation of coding mRNA. Substitutions in the DNA sequences for these regions are limited, much less than in synonymous changes but, of course, higher than in non-synonymous changes.

Flanking regions are found in functional genes both at 3′ and 5′ ends. Sequences in the 3′ flanking regions have no known effect on amino acid sequences in proteins. Any substitution in this region is tolerated by natural selection. Consequently the rate of substitution at this region is as high as in synonymous changes. 5′ flanking regions are also not transcribed or translated but they are important for gene expression, because they contain the promoter and other regulatory elements. A small difference in consensus sequences (such as TATA box) may have harmful effect on fitness of the organism. Naturally, substitution rate in these regions are kept low by natural selection.

Substitutions at pseudogenes

Pseudogenes are stable but inactive component of a genome resembling a gene and apparently evolved from active gene through mutations. Pseudogenes do not code for any protein. Because any change in them does not affect the fitness of the organism, natural selection does not bother about their change. Therefore, highest rate of substitution is observed in pseudogenes, more than that of synonymous changes and about 5 times higher than non-synonymous changes.

Relative rates of evolutionary changes in DNA sequences of mammalian genes based on substitution rate are given in Table 20.1. From the above discussion and Table 20.1, we can make generalized conclusion. A sequence with more functional quality has slower rate of evolution.

Many genome projects are producing large amount of information on nucleotide sequences. But about 96% of nucleotides are functionally not important within a complex organism. So, it is often difficult to make sure which portion of a genome is associated with definite functions. When a similar sequence is discovered in the genomes of two distantly related species, it may be used to suggest functional importance. A pair-wise comparison of the sequences on the basis of evolutionary rate and functional significance is done and this method is known as **comparative genome analysis**. The method is rapidly taking the position of time-consuming research, in which every position of a gene is evaluated.

Table 20.1 : Relative rates of evolutionary changes in DNA sequences of mammalian genes

Sequence	Nucleotide substitution rate Per site per year ($\times 10^{-9}$)
Functional gene	
Coding sequences	
Synonymous	4.65
Non-synonymous	0.88
Non-coding sequences	
Introns	3.70
Leader	1.74
Trailer	1.88
3′ flanking	4.46
5′ flanking	2.36
Pseudogenes	4.85

20.4.3 Variations in Rates of substitution of different Genes, i.e. Interspecific Variations

Rates of substitutions not only vary within different sequences of a gene, there is striking difference in the rate of substitution between genes. If stochastic factors (that is, differences due to sampling error or genetic drift) are excluded, then the differences must be due to any one or combination of two following factors—

(i) difference in mutation frequency, and

(ii) extent of action of natural selection on that locus.

Statistical analysis helps to distinguish whether the difference is adaptive or due to random change in nucleotide sequence. The patterns of within-species polymorphism and between-species divergence at the synonymous and non-synonymous sequences of genes are compared. If the ratio of non-synonymous to synonymous substitutions within species is same as between species, then

Table 20.2 : Relative rates of evolutionary changes in DNA sequences of different mammalian genes

Gene	Synonymous substitution per site per year x 10^{-9}	Non-synonymous substitution per site per year x 10^{-9}
Histone H_4	1.43	0.004
Insulin	5.41	0.16
Prolactin	5.59	1.29
α– Globulin	3.94	0.56
β– Globulin	2.96	0.87
Albumin	6.72	0.92
MHC	2.40	5.1

substitutions should be neutral. If the ratios are not same, then natural selection must be involved. Synonymous and non-synonymous substitution rates observed in different classes of mammalian genes are shown in Table 20.2. Though some regions of genomes undergo more random changes than others, synonymous substitution rates vary little. Certainly, non-synonymous substitution rates vary roughly 1000 folds (from 0.004 to 5.1). This variation of substitution rates between genes must be due to different intensity of natural selection at each locus.

We can consider two examples of substitution rates, one with the slowest and another with the highest rate in Table 20.2. **Histones** are essential binding proteins present in all eukaryotes. Almost each amino acid in histone H_4 interacts with specific chemical residue of DNA. Any change in amino acid sequence of histone H_4 decreases its ability to bind with DNA, therefore, reduces fitness value. In fact, it is possible to replace Yeast histone H_4 by human homology, though each one has undergone independent evolution for hundreds of million years.

Surely, amino acid substitutions are deleterious, but natural selection favors tremendous variability of amino acid sequences for some genes within population. **Major histocompatibility complex** (MHC) is a large multigene family of mammals whose protein products are essential for immune system and provides ability to receive foreign antigens.

About 90% of human beings receive different sets of MHC genes from their parents. Such high level of diversity in these regions is favored by natural selection because diverse immune system is much better against any single virus than identical immune system. As a result, rate of non-synonymous substitution is very high within MHC and even more than double the rate of synonymous substitutions.

20.4.4 Interpretation of variations in DNA sequences – role of natural selection

Motoo Kimura (1991) proposed that much of the pattern of evolutionary changes in molecules could be explained by a combination of random mutations and random chance fixation of alleles. This model popularly known as **neutral theory of molecular evolution** was applied to both intra-specific and inter-specific variations of DNA sequences. Initially, many available data seemed to fit good statistically to the predictions of this model. The model acknowledges the presence of extensive genetic variation but proposes that this variation is neutral with regard to natural selection (see section 15.17). Therefore, the patterns of genetic variations we see in a natural population are shaped by random processes like mutation and genetic drift, not by natural selection. But recent data on patterns of molecular variations are found to depart significantly from those predicted by neutral theory. Evidence are increasing that selection of molecular variants may be fairly common.

Natural selection can modify neutral patterns of variations in three ways—

1. **Positive directional selection** fixes a sequence that includes an advantageous mutation and also reduces variations at closely linked sites. We assume—

   (i) occurrence of an advantageous mutation for which neutral variation exists in population;

   (ii) there is no recombination nearby of advantageous mutation, and

   (iii) the mutation is fixed by selection

   Then it is expected that all gene copies in the population will be descended from the single copy, which has the advantageous mutation. The neutral variant sites linked to this mutation will also be fixed. Thus all neutral variations in the gene are eliminated, and the phenomenon is called **selective sweep**. Then, new neutral mutation will occur in the copies of the advantageous gene. Selective sweep is similar to bottleneck effect in population because it reduces variations and increases relatedness among gene copies. But the difference between the two is that, bottleneck affects the entire genome, not just some DNA sequences around an advantageous mutation.

2. **Balancing selection** maintains variant sequences in a population and acts in opposite direction to positive directional selection. We assume that rate of recombination is very low near a nucleotide at which two variants are maintained at polymorphic state by selection. Then all gene copies in the population are descended from two ancestral copies. Each of these genes was the progenitor of a lineage of genes that have accumulated neutral mutations near the selected site. Thus sequences subjected to balancing selection will show more variations near the selected sites than a sequence having only neutral variations.

3. **Purifying selection** eliminates or reduces the frequency of deleterious mutations in a population. Because this kind of selection works against deleterious mutation, neutral polymorphism at closely linked sites is also reduced. When a copy of deleterious mutation is eliminated from population, selectively neutral mutations linked to it are also eliminated, the effect is known as **background selection**. Thus for a DNA sequence with low rate of recombination, population size is reduced to proportions of gametes that are free from deleterious mutations. This also affects the heterozygosity for neutral mutation, which will reduce if mutation rate is high, mutation is strongly deleterious and rate of recombination is very low.

Selection may act on silent substitution

It has been already discussed that synonymous substitution in a coding sequence does not produce any change in amino acid sequence of proteins; they are called **silent substitutions**. Therefore, silent substitutions are expected to have higher rate according to rate of substitution/functional significance relationship. But Table 20.1 shows that relative rate of change in synonymous substitution is slightly lower than that of pseudogenes. This observation suggests that synonymous substitutions are not entirely neutral from the influence of natural selection. Evidence that selection do act on silent mutations is increasing since the discovery of a phenomenon – **codon usage bias** (Grantham *et al*, 1980).

The hypothesis is strengthened by the finding that synonymous codons are not used equally throughout the coding sequences of many organisms. As for example, redundancy of genetic code allows leucine to be coded by six different codons – UUA, UUG, CUU, CUC, CUA and CUG. It has been found that 60% of leucine codon used by bacteria is CUG but 80% in yeast are UUG. Since alternate synonymous codons specify same amino acid, they are likely to be used equally. But differential use of codon (that is bias for codon) suggests that selection must be favoring some synonymous codon over others.

We know that some synonymous codons pair with different tRNA, though they carry same amino acid. Thus a mutation in a synonymous codon does not change amino acid sequence but may change the tRNA with which it pairs during translation. Studies on tRNA within cells reveal that amount of different tRNA accepting the same amino acid (isoacceptor tRNA) differs. The tRNA, which pairs with most frequently used codon, is most abundant. Selection, therefore, may favor one synonymous codon over another because tRNA for the first codon is available most. Consequently, translation of that codon is more efficient. Alternatively, energy required for bonding between codon and anticodon of synonymous codon may differ slightly because

different bases are being involved. In pairing, these extremely subtle differences in translation efficiency and bonding energy may result in differences in fitness, when subjected to natural selection over the course of thousand or millions of generations.

We have two derivations from the codon usage bias theory—

(a) Selection for favored codon should be quite weak, as evident from the slight difference between the rate of change in synonymous coding sequences and pseudogenes. Then selection (probably purifying) against synonymous mutations can be effective if the population size is large and organisms have short generation time. In smaller populations, such mutations will be neutral; so little codon usage bias will evolve. This is supported by the finding that codon usage bias is pronounced in bacteria and yeast, less pronounced in *Drosophila* and weak or absent in mammals.

(b) If codon usage bias is a result of purifying selection, then stronger the bias, lower the number of neutral mutations. Thus neutral theory predicts that rates of evolutionary change by random fixation of synonymous substitution are low in genes with high codon usage bias. The prediction is consistent with the result of comparing the sequences of 23 genes in closely related bacteria – *Escherichia coli* and *Salmonella typhimurium* (Sharp and Li, 1987)

20.5 THE MOLECULAR CLOCK

Rates of nucleotide and amino acid replacement between nuclear genes differ strikingly but are primarily the result of different selection forces on each individual protein. However, loci with similar constraints have quite uniform rates of molecular evolution over long periods of evolutionary time. Emile Zukerkards and Linus Pauling performed some earliest comparative studies of protein sequences in 1960s and suggested that substitution rates within homologous proteins were almost constant over tens of million years. In fact, they suggested that accumulation of amino acid changes occurred in such a steady rate that it can be compared with the ticking of a clock – called **molecular clock** for evolution. Molecular clock may run at different rates in different proteins but appear to be well correlated for two homologous proteins. The hypothesis stimulated intense interest in evolutionary study of biological molecules. A steady rate of change between two sequences should help in two ways—

(i) in determining the phylogenetic relationship between species, and

(ii) in determining the time of divergence of the two species, almost like geological dating using radioisotopes.

A strong support for the molecular clock is the constant number of differences in amino acid sequence for the same hemoglobin chain derived from different vertebrates. When compared with shark hemoglobin, other vertebrates differ from it by fairly constant number of amino acid changes – carp 85, salamander 84, chicken 83, mouse 79 and human 79. Despite considerable morphological changes in these different lineages over a 400 million year period, constant rates of mutation may have been occurring for at least some proteins.

In spite of its tremendous promise, molecular clock hypothesis has been controversial. A study of amino acid sequences in seven different proteins has been done in 18 vertebrate taxa. The result shows that the rate at which all proteins have changed together varies significantly among different lines of descent. It indicates that molecular changes are not uniform.

To avoid controversy, scientists devised a simple method to estimate the overall substitution rate in different lineages, from which relative rate of substitution in each lineage is determined. Relative rate tests performed on homologous genes show that substitution rates in rats and mice are almost same; rates in rats is twice the rate of primates; rates in human is half of that of old world monkeys. It is clear that molecular clock varies among taxonomic groups, and such departures from constant clock rate pose a problem for the usage of molecular divergence.

Possible explanations are forwarded to account for the differences in evolutionary rates estimated by relative rate test. For example, monkeys have shorter generation time than humans and rats have even less generation time. Germ-line DNA replication occurs once per generation, naturally rates of substitution should be higher with shorter generation time. Difference of rates between two

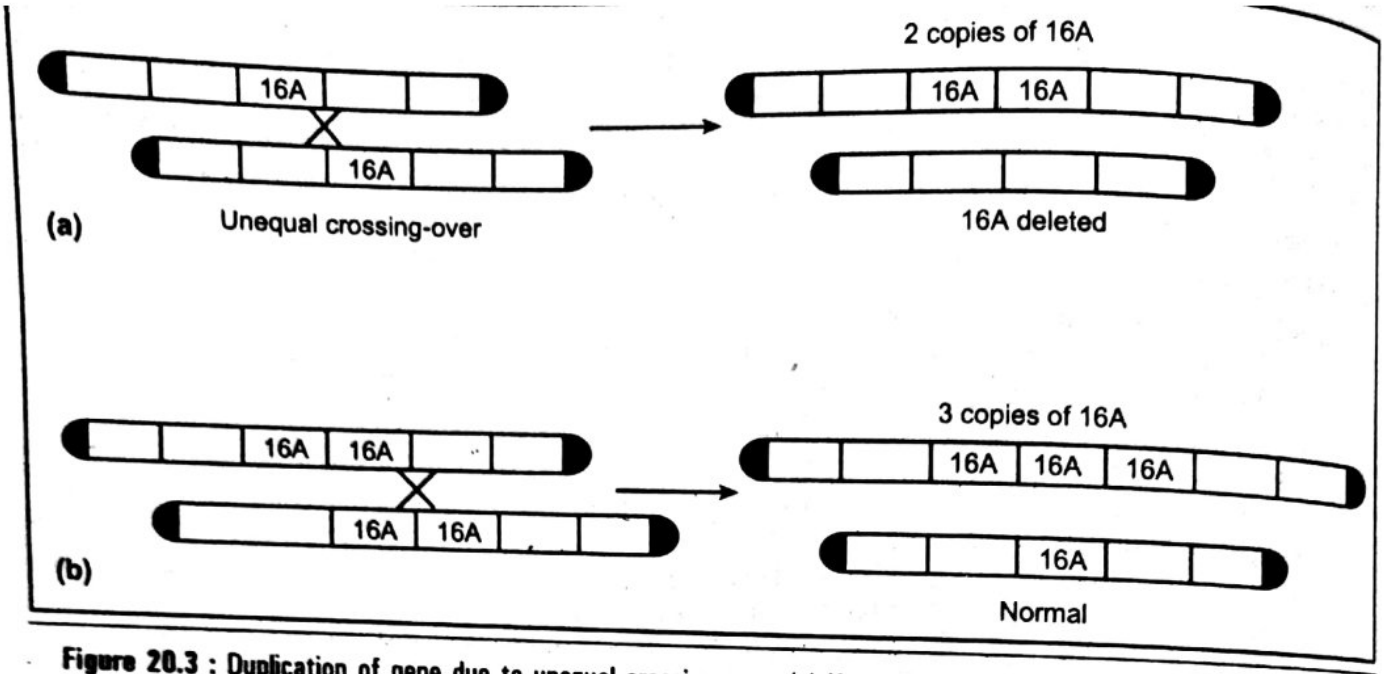

Figure 20.3 : Duplication of gene due to unequal crossing over. (a) Unequal crossing over of the X chromosome of *Drosophila* results in two copies of 16A segment in one chromosome and none in another. Homozygosity for duplicated 16A results in Bar eye mutation. (b) Another unequal crossing over brings three copies of 16A segment in one X chromosome and a single copy in another. Homozygosity for three 16A segments results in double-Bar mutant flies.

lineages since their divergence from common ancestor may also result due to average repair efficiency, average exposure to mutagens and opportunity to adapt to new ecological niches.

20.6 ORIGIN OF NEW GENE FUNCTIONS

A very basic question in the study of molecular evolution is how genes with new functions arise. Haldane (1932) first suggested that new genes arose from redundant copies of already existing genes through mutation. Haldane's suggestion is still applicable even though several other means of origin of new functions are now known.

Multigene families are multiple copies of some ancestral genes

Genomes of eukaryotes often carry tandemly arranged multiple copies of genes, all having identical or very similar DNA sequences. These sets of related genes have evolved from some ancestral genes through the process of duplication, and is known as **multigene families**.

Gene duplication provides raw material for evolution of new genes

Gene duplication often arises as a result of misalignment of sequences during crossing over, known as **unequal crossing over**. Unequal crossing over may result when homologous chromosomes pair inaccurately, perhaps because similar DNA sequences occur in neighboring regions of chromosomes (Fig. 20.3). Consequently, one chromatid receives lowered and other receives greater number of copies of genes. Once greater number of copies arise in one chromatid, unequal exchange becomes more likely, because a copy on one chromosome can pair with any of the copies on another. Repeated sequences within gene can also arise by unequal crossing over.

The globin gene family that encodes the protein part of hemoglobin in our blood is a classic example of multigene family through gene duplication. Globin genes are also found in other animals and globin-like genes are found in plants, indicating that this is a very ancient gene family. Almost all functional globin genes in animal species have same general structure, consisting of three exons and two intervening introns.

The similarity in structure and sequence of all globin genes suggests that they all arose from an ancestral globin gene. History of vertebrate globins has been traced (Goodman, 1982) and found that duplication of an ancestral globin gene in an