Molecular Evolution

Corresponding Author:
Prof. Krupanidhi Sreerama,
Professor of Biosciences and Dean, Department of Biosciences, SSSIHL, Vidyagiri, Prasanthinilayham, 515 134 - India

Submitting Author:
Prof. Krupanidhi Sreerama,
Professor of Biosciences and Dean, Department of Biosciences, SSSIHL, Vidyagiri, Prasanthinilayham, 515 134 - India

Article ID: WMC002049
Article Type: Systematic Review
Article URL: http://www.webmedcentral.com/article_view/2049
Subject Categories: ZOOLOGY
Keywords: Allele Frequency, Gene Families, Gene Duplication, Pseudogenes, Retrogenes, Evolution of Introns, Nucleotide Substitution

How to cite the article: Sreerama K. Molecular Evolution. WebmedCentral ZOOLOGY 2011;2(7):WMC002049

Source(s) of Funding:
Did not receive the support from any funding agencies.

Competing Interests:
The contents of the article do not have any competing interests.
Molecular Evolution

Author(s): Sreerama K

Contents

1. Introduction
2. Classical evolutionary concepts
   2.1 DNA as a crux
   2.2 Theories of Molecular Evolution
   2.3 C-value paradox
3. Gene frequencies in Mendelian Populations
4. Allelic frequency
   4.1 Application of Hardy-Weinberg Law
   4.2 ABO Blood groups
5. Genotype versus ethnicity
6. Evolution of gene families
   6.1 Gene duplication leads to gene families
   6.2 Globin gene family
   6.3 Pseudogenes
   6.4 Retrogenes
   6.5 Gene duplication by exon shuffling
7. The evolution of introns
   7.1 Function of introns
8. Molecular clock hypothesis
   8.1 Examples of nucleotide substitution
   8.2 Evaluation of rate of gene substitution
9. Molecular Trees.

1. Introduction:

The bewildering biodiversity around us is a unique biomass to contemplate upon. Despite the diversity, living beings have many features in common both morphologically and anatomically. At times, behaviour responses of organisms also appear alike. Organisms are so designed as to survive and reproduce to continue their race. All life-forms around us from micro to macro level has the same nucleotide-based hereditary material, the same genetic code, the similar protein-synthesizing machinery and proteins composed of the same 21 amino acids. Using hybridization probes, it has been revealed that the designer genes, such as axis specification genes and genes for pattern formation, among a wide range of organisms are evolutionarily conserved and share extensive sequence homology (Krupanidhi, 2005). Then, how come organisms look different? Why not there is only one life-form? Why there is a diversity and / or similarity among organisms? Is this due to adaptations of organisms to their respective environments? Acquired adaptations and consequent built-in mechanisms brought about morphological and behavioural diversities and they constitute a rich resource and a gold mine for the evolutionary biologists to ponder over. There are a variety of organisms, sharing some features, differing in others, gradation in the extent of vulnerability to diseases, evolving varied life patterns and developing different strategies and pathways for their success. Why did life-processes choose to exist in these diversified forms? The discipline of evolutionary biology enables us to understand the diversified biological entities hosting life. There are two general approaches to study molecular evolution: 1) Use of DNA to study the evolution of organisms (such as population structure, geographic variation and systematics) and 2) Use of different organisms to study the evolution of DNA. The goal is to infer process from pattern. The process of organisal evolution is deduced from patterns of DNA variation. Thus the molecular evolution, an imperceptible solution for biological diversity, is also called as the “Natural History of DNA”.

2. Classical Evolutionary concepts:

The Latin equivalent word of evolution is “evolvere”, which means to “unfold”. Broadly one can consider the word “change” as the typical meaning of evolution. In biological parlance the term evolution was used as “descent with adaptational variation and often with diversification”. Charles Darwin titled his book as “Descent of Man” rather than “Ascent of Man” and elicited the concept of natural selection. The so-called present-day living organisms are descendents of their immediate ancestral stock, carried over conserved features and acquired recent adaptable traits. It is the natural selection that conserves and sustains adaptable features. Hereditary variations among phenotypic expressions of adaptable features are primarily due to mutations, random genetic drift, recombinations, translocations, gene duplications, sorting, unequal crossing over, exon shuffling, etc., and a few of them are elaborated in subsequent sections. Charles Darwin (1858) provided innumerable examples (finches, tortoises, etc.) and evidences for the historical reality of evolution. Darwin hypothesized that the cause of evolution is primarily due to the ongoing forces of natural selection acting upon both physical and genetic traits of vulnerable organisms that are unable to thrive in their contemporary habitats.
Thus, the evolutionary theory is not a mere speculation but a body of hypotheses, which are being supported by experimental evidences. The Synthetic theory of Evolution emerged during 1940s due to the contributions of J B S Haldane, Sewall Wright, Ernst Mayr, Julian Huxley, G G Simpson, T Dobzhansky and several others, remained as the foundation for the contemporary evolutionary theory. The recent mammoth genome projects through comparative genomics have thrown more light on concepts of evolution. The mechanics of evolution is currently explained more through nucleotide substitutions than ever before.

DNA as a crux:
Proteins and DNA are macromolecules. They are endowed with the information to dictate biological function. Therefore, it is no wonder that early biologists heaved the burden of heredity also upon them. It was through the findings of Fredrick Griffith and others, DNA came into lime light as a macromolecule of heredity. An individual's nuclear DNA constitutes genomic DNA. It earmarks some of its portions as organized sequences of nucleotides. They are called genes. In eukaryotes, genes are organized into introns and exons, whereas prokaryotes are devoid of introns. The regulatory sequences in genomic DNA specify when and where to transcribe which gene into RNA for protein synthesis. The genetic code is the information system for translating the sequence of RNA into amino acids. Within the triplet code some of the nucleotide positions are silent or synonymous because any nucleotide in that position will do, which reveals that the genetic code is degenerate. The discipline of molecular genetics and population genetics is the outcome of the synonymous and non-synonymous substitutions of nucleotides in the genomic DNA and thus built the magnificent edifice of molecular evolution.

Theories of Molecular Evolution:
There are three theories viz., i) Neutrality: (Neutralism and near-neutralism): (Kimura, 1983): This theory views gene substitution and genetic polymorphism as the two surfaces of the same coin. ii) Selectionism: (Gillespie, 1991) beneficial and balancing selection are considered as the main driving force and finally iii) Mutationism: (Nei and Koehn, 1983) mutational inputs and random genetic drift are thought to play a vital role.

C-value paradox:
One of the most interesting observations to ponder in molecular evolution is the C-value paradox. The C-value of a species is the Characteristic or Constant amount of DNA in a haploid genome. There is a clear increasing trend in biological complexity among the diversified organisms from viruses to humans. However, upon comparison the C-values across the range of organisms reveal that some of the less complex organisms have much more DNA content than the more complex organisms. This presents a paradox. DNA codes for proteins that give us form and function. What is that the lower phyla viz., algal species and amphibians doing with the extra DNA? The possible reasons need to be elucidated. Presumably, most of it may be “non-functional” however; the extra DNA possibly helps in aligning chromosomes during cell divisions and protects the functional genes from obnoxious exposures.

3. Gene frequencies in Mendelian Populations:
Geneticists define populations as groups of individuals that are united by bonds of mating and / or parenthood. Natural populations of sexually reproducing organisms are commonly referred as Mendelian populations. This reflects the fact that genes are passed from one generation to the next according to the Mendelian principles i.e., Genes Flow through generations. If at all, mutations arise within local Mendelian populations, they will spread through an entire species but cannot ordinarily be transferred across species boundaries. The basis of the so-called ‘biological species concept’ is that, groups of Mendelian populations are reproductively isolated from other such groups – i.e., gene flow between populations is intrinsically interrupted. Evolutionary geneticists define evolution as the process that brings about a change in gene frequencies within and between populations in the course of time. This, nevertheless, conveys the very basic notion that all major morphological, physiological, and behavioural changes that characterize different species focus on nucleotide changes in genes present in a population. Populations are thus considered as the smallest evolutionary units because, unlike individuals, populations do not have relatively short life spans and in addition they may genetically evolve over many generations. Individual organisms can live for only a limited period of time, whereas, populations and species persist for many thousands of generations.

The genetic changes that occur in populations ultimately are the basis of evolution. What factors bring about a change in the genetic character of populations over time? What mechanisms led to the formation of new species? What forces bring about differences among individuals of the population in relation to immunological tolerance? The modern synthetic theory of evolution is attempting to answer such queries based on the fact that the genetic
variation is the raw material for evolutionary change. Individuals making up natural populations may differ among themselves and members of other populations in a variety of morphological, physiological, and biochemical traits. Much of the observable variability that exists among individuals, making up natural populations, is reflected in heritable differences or genetic variation. The sum total of genes and their genetic variation present in a population or species is often referred to as the gene pool. Thus, population genetics revolves around gene pool that elucidates the understanding of factors determining how the genetic character of populations is maintained and changed over time.

4. Allelic frequency:

A number of reasons prompt geneticists to unravel genetic variations present in natural populations. For example, hemoglobin S allele causes sickle-cell anaemia in individuals homozygous for the allele, and analyzing gene distribution pattern between populations may provide insight into the origin of the disease. The founder of population genetics, J B S Haldane, observed that the frequency of S allele is the highest in areas of the world with a high incidence of the malarial parasite, Plasmodium falciparum. This correlation led Haldane to hypothesise that the S-form of haemoglobin protein might, in addition to causing anaemia, also be associated with resistance to malarial infection. Thus, this profound survey explained why the frequency of S allele is exceptionally high in contemporary ethnic groups whose ancestors could be traced to the tropical regions of the world plagued by the malarial parasite. Application of Hardy-Weinberg Law:

There are several numerical ways to describe the amount of genetic variation that is detected in natural populations. For example, consider the result of an electrophoretic survey of adult and sickle hemoglobins (Illustration 1-Fig. 1) carried by individuals in the West African population. In this example, blood samples were represented from 500 randomly selected adults. Electrophoretic techniques were used to determine the type of hemoglobins carried by each individual. Because each hemoglobin variant is known to encode by a different allele, these data can be used to determine individual's genotype at the hemoglobin locus. Of the 500 individuals sampled, 336 carried normal hemoglobin protein: that is they were homozygous for normal A allele (AA). A total of 160 individuals carried both normal hemoglobin proteins and its variants, and they were genetically heterozygous (AS) at the hemoglobin locus. Only 4 individuals were homozygous for S allele (SS), and they were severely anaemic. The relative frequency of each hemoglobin genotype in this population is obtained by dividing the number of individuals having each genotype with the total number of individuals sampled. Thus, the frequency of AA homozygous genotype is 0.672 (336/500), AS heterozygote genotype frequency is 0.320 (160/500), and SS homozygote genotype frequency is 0.008 (4/500). Note that the sum of frequencies of each genotype in a population must be equal to 1.00 (0.672 + 0.320 + 0.008 = 1.00). To calculate allelic frequencies at the haemoglobin locus in this example, divide the number of each allele with the total number of alleles in the sample. Being diploid organisms, every human carries two alleles at each locus. Thus, in a sample of 500 individuals, 1000 alleles are represented at the haemoglobin locus. Each individual with AA genotype carries two A alleles, and every heterozygote carries only one A allele. Thus, the frequency of A allele in the sample: \( f(A) = \frac{2(336) + 160}{2(500)} = 0.832 \). The frequency of S allele: \( f(S) = 0.168 \) (The value obtained by subtracting from 1.00 when there are only two alleles). The sum of frequencies of each allele in a population is equal to 1.00 (0.832 + 0.168). The possible reasons for S allele to retain in individuals of West African population could be due to their inhabitation in the tropical region and frequently fighting against malaria.

ABO blood groups:

The popular blood group polymorphism is ABO system described by Karl Landsteiner. Genes involved in distinguishing blood groups came into limelight in 1990. A and B groups are co-dominant versions of the same gene. Their products are antigenic. These antigens lie on the surface of human red blood corpuscles. O-group is their recessive form. These genes lie on long arm (p) of chromosome 9. It is a medium sized gene containing 1,062 nucleotides and divided into six short and one long exons. This multiple allelic gene is the recipe for the enzyme, galactosyl transferase. The difference between A and B-alleles is only due to mere seven nucleotides out of 1,062. Of the seven, three are silent (synonymous) i.e., they make no difference in the corresponding amino acids. The remaining four make an issue. The positions of these nucleotides are at 523, 700, 793 and 800. In type A-blood group these nucleotides are C, G, C and G while in type B-blood, they read G, A, A and C. In O-group there is a single deletion at 218th nucleotide, where G is missing. This causes frame shift, rendering the same recessive.
Susceptibility and resistance to diseases among individuals of various ABO blood types is quite fascinating. The relative merits of blood groups A, B, AB and O and their possible selection constitute insightful examples in understanding the distribution of the respective allelic frequencies in ethnic groups. Individuals with O-group are more susceptible to cholera; seem to be more resistant to malaria, less likely to get cancers of various kinds and less susceptible to Syphilis. The most resistant ones for cholera are those with AB-group. In fact, A-group gene is better for cholera resistance than B-gene. Why does natural selection still prefer B-gene? Yet, we have to find a reasonable answer. Presumably, 50% of healthiest children will be the offspring of Aa and Bb parents. This means that 50 % children of these parents may either be with AB or O (ab) types. The higher survival value of O-version of gene is therefore advantageous and thus preferred by natural selection. Therefore, a rough balance is conspicuous between the four variations of ABO multiple alleles in several ethnic groups. For example, roughly 40% of Europeans have type O-blood, 40% of type A-blood, 15% of type B-blood and 5% of type AB-blood.

Case Study: The phenotypic frequencies of blood groups among the student population of an institute (Sri Sathya Sai Institute of Higher Learning, Prasanthinilayam, year 2006) are as follows: O = 45%, A = 24%; B = 26% and AB = 5%. Using Hardy-Weinberg equation (p+q+r =1), the allelic frequencies of p, q and r of A, B and O blood group alleles are found to be 15%, 18% and 67% respectively. The possible explanation for the relatively low percentage of AB phenotype would have been the prevalence of high percentage of recessive allele, namely O-version, which authenticates positive selection in resisting endemic diseases in tropical regions.

5. Genotype versus ethnicity:

There are a number of examples explaining the genetic differences among human races, particularly with reference to the disease tolerance and susceptibility. A few considered in this article are relating to the habitat, lifestyle, endemicity, allele polymorphism and origin of the population. While HLA polymorphism makes an individual or a population unique identity in the biological system as it is being spoken as self- molecule, KIR polymorphism (Rajalingam, et al., 2001) drives human population towards its survival advantage. KIR genes are encoded on chromosome 19q 13.4. These genes are expressed by natural killer cells. Leukocyte receptor complex includes a gene family viz., cell surface receptors made up of Ig-like folds – they are KIRs, HLA, TLRs, etc. There are 16 KIR genes. Of which, a few of their receptors are inhibitory and the remaining are activatory in their function. Haplotypic diversity is noticed among them and it is one of the major contributors to the population diversity. Two haplotypes designated as A and B encompass all KIR genes. KIR - 2DS1, 2DS2, 2DS3, 2DS5, 3DS1 and 2DL5 belong to B-haplotype, whereas A haplotype consists of KIR genes other than afore-mentioned genes. The distribution of A and B haplotypes in the population reflects its biological and health perspectives. Therefore, it is no wonder that KIR gene frequencies are distinctly different among populations. In a study, focusing on Asian population and KIR haplotype distribution, conducted by Vi Chuan Lee, et al., (2008) in Singapore, revealed that the three populations chosen viz., Singapore Chinese, Singapore Malay and Singapore Indians shown to have all four frame- work genes (KIR 3DL2, 3DL3, 3DP1 and 2DL4), whereas a significant variation is noticed among the KIR 2DS2, 2DL2, 2DL5 and 2DS5. It is further reported that the haplotype A is found to be dominant in Singapore Chinese population. Upon comparison, it is revealed that Japanese and Korean populations also follow the pattern reported for Singapore Chinese. Singapore Indians are comparable to North Indian Hindus. While, Singapore Malay follows Thai population in relation to KIR haplotype A and such a survey predicts the uniformity in seasonal immunological sensitivity and accordingly the preventive measures can be taken up in this Asian subcontinent.

Human leukocyte antigen (HLA) plays a pivotal role in immunological presentation of peptide fragments of pathogens to immune competent cells. Furthermore, HLA polymorphism is so wide that each individual or population possesses a unique set. HLA haplotypes constitute the fingerprint of an individual/population. HLA complex is located on chromosome 6 p21.31 and a few of HLA haplotype combinations are being associated with specific autoimmune disorders (e.g. HLA-B8/HLA-DR3 haplotype is associated with Graves’ disease, HLA-B27 is associated with Ankylosing spondylitis, HLA-DR4 with Rheumatoid arthritis) in several ethnological groups. In one specific example, HLA class-1-A and B distribution and their haplotype frequencies in the region of Tamil Nadu of South India revealed the presence of 4 major groups among its inhabitants which differ in their origin and hence suggested to consider the caste system for disease prognosis (Pitchappan, 2001). In another...
observation, it is noticed that there is a similarity between Mediterranean and European populations in relation to HLA allele frequencies. HLA polymorphism also distinguishes a population from its neighbors viz. Turkish population is genetically more similar to its geographic neighbors than its historical neighbors in Central Asia.

Prostate specific antigen (PSA) is a biomarker for detecting the stage in prostate cancer. However, the mean values of PSA markedly vary between countries. The highest range of PSA values (4.0 ng. to 10.0 ng. per ml serum) are reported in India whereas the least values (1.28 ng to 4.0 ng per ml serum) are reported for American males. Hence, there is a variation in prostate cancer disease prognosis among patients of these two countries.

Wilson’s disease is hepatolenticular degeneration due to the defect in copper transporter protein viz. ATP7B, whose prevalence in Japan is higher. The lysosomal storage disease is known as Gaucher disease and it causes early childhood death. Its highest incidence is common among Ashkenazi Jewish population of South Africa due to an inherited deficiency in the enzyme, glucocerebrosidase. This population is being identified by having four specific marker alleles viz N370S, L444P, 84GG and IVS2 responsible for this inherited disease. Yet another health defect viz. Mediterranean anemia, exhibits different gradations in its severity, is collectively named under Thalassemia. This anemia is mostly distributed in South East Asia, Middle East, China and in the regions of Mediterranean. Thalassemia is due to mutations in a-globin and b-globin genes and accordingly the disease condition is named as a-Thalassemia and b-Thalassemia. There are four a-globin gene copies whose defects lead to a severe anemic situation, whereas a mild anemia is due to point mutations replacing G to A nucleotide in the first intervening sequences of b-globin gene.

6. Evolution of Gene Families:

The evolutionary process includes innumerable number of mechanisms that eventually shape genetic character of populations and species eventually. Comparative studies of DNA structure and function in various genomes have provided considerable insight into the kinds of molecular diversity. This must have accompanied by the emergence of the vast array of phenotypic diversity. A few evolutionary changes occurred at the molecular level seem to have been adaptively neutral, whereas others must have played important roles in the acquisition of new and significant molecular structures and functions. Genomic sequences include billions of repeating nucleotides viz., A, G, C and T which determine shapes of diversified organisms. This discipline of molecular evolution examines the consequences of these changes on morphology, physiology, behaviour, adaptability and other aspects of organic evolution. The molecular phylogeny of a few genes viz., haemoglobin, insulin, b-globin, haemocyanin, KIRs (Killer immunoglobulin-like receptors), HLA (Human leukocyte antigen) and cytochrome C, whose structural and functional complexities have been elucidated, offered as typical examples to understand the mechanism of gene evolution. Eukaryotic genomes contain related groups of genes. These related groups of genes, consisting of two or more exons with similar or identical DNA sequences — are called gene families.

A few examples: genes encoding interleukins, contractile proteins, dehydrogenase enzymes, b-globin (Illustration 1-Fig.2 and Illustration 2-Fig.3), HLA antigens, KIRs, rRNA or histone proteins, possibly have descended by duplication and divergence from common ancestral genes. The DNA sequence similarity within a family can range from identical, or nearly identical, to quite different. Most members of a gene family are clustered in close chromosomal proximity to one another; however, some are located on different chromosomes that prompted molecular evolutionist to contemplate upon.

These dispersed gene family members presumed to have been translocated to their different locations. (E.g. MHC, HLA and α- and β-globin genes). Generally members of a gene family have the same or related functions. For example, mammalian haemoglobin gene family encodes proteins whose physiological function is to transport oxygen. However, even when members of a gene family have the same basic function, they are not always expressed at the same time during development (Illustration 1-Fig.2).

Different members may be expressed at different stages of life and/or in tissues reflecting the fact that evolutionary divergence has occurred at the level of gene regulation. Some members of the mammalian haemoglobin gene family are expressed in adults, whereas others are expressed only at the fetal stage of development. Gene duplication leads to Gene families:

Gene families are believed to arise from a succession of gene duplications. Unequal crossing over during meiosis is one of the primary mechanisms of gene duplication. This process can generate deletions as well as duplications. For e.g. thalassemia is a disease resulting from an inability to make functional a and b globin genes either due to unequal crossing over or
point mutation. Unequal crossing over is not an uncommon phenomenon. It plays a significant role in gene duplication and thus in the evolution of multigene families.

Globin gene family:

The mammalian hemoglobin (hb) gene family consists of eight active globin genes encoding hb proteins. On the basis of relative degree of sequence homology, they can be subdivided into two clusters, the α-globin cluster and β-globin cluster. The α-globin consists of ζ-genes (early embryonic) and two nearly α 1 and α 2 in fetus and adult life. θ gene is recently discovered. β-globin cluster (Illustration 1-Fig.2 and Illustration 2-Fig.3) consists of: embryonic θ-globin gene and two nearly identical fetal globin genes (Gy and Ay) and a minor adult δ in addition to the major adult β-globin gene. Although all members of the mammalian globin gene family display a significant degree of sequence homology, some are nearly identical, whereas others are more divergent. From these analyses, it is estimated that an incipient duplication of ancestral globin gene must have occurred about 800 mya (millions of years ago), establishing a precursor of the α and β genes-clusters. This ancestral gene must have been related to the myoglobin gene, or leg hb of plants. Leg hb has one intron compared to the myoglobin gene; otherwise, it is nearly identical to myoglobin gene. Each cluster must have undergone duplications, transpositions and mutations over the last 400-500 million years (Illustration 2-Fig.3).

Pseudo genes:

In a number of instances, the non-functional gene family members have been found. They lack the necessary promoter sequences and also one or more of introns that are characteristic of active members of the gene family. These non-functional gene family members are called pseudogenes. Some members of these gene families do not transcribe. Even if they produce transcripts that are not properly processed and translated into functional proteins (For ex: Ψ member of β-globin gene cluster). They are comparable to fossils and thus through light on the evolutionary history of the extant genes. Hence, pseudogenes are the disabled duplicates of working genes. Pseudogenes can be born in either of the two ways. During cell division, an extra copy of a gene can be inserted into a new location. Alternatively, a new version of a gene can also be created through reverse transcription. Pseudogenes made from mRNAs lack introns and are described as processed pseudogenes (Mark Gerstein and Deyou Zheng,2006). E.g. (1) A single ribosomal protein gene known as RPL21 has spawned more than 140 pseudogene copies. (2) Olfactory epithelium is studded with invisible odorant receptor molecules that detect smells. A family of more than 1000 genes encoding olfactory receptors in mammals has been identified. Large-scale pseudogenisation is most often seen among genes of olfactory receptor family. An organism’s pseudogenes reflect the species-specific adaptations and their evolutionary history.

Retrogens:

Replicate genes generate through reverse transcription of partially or completely processed progenitor gene transcripts. In most instances, however, retrogenes would not be expected to flank downstream of promoter sequences and thus would not be transcribed unless inserted by chance downstream of an active promoter. Consistent with this expectation, most retrogenes do in fact lack promoters, and as a consequence they are transcriptionally inactive. These inactive members of gene families have been called retropseudogenes. They are the members of many gene families, including mammalian globin gene family. In a vast majority of cases, retrogenes do not retain their functionality due to the following three reasons: i) the process of reverse transcription is not always accurate. Hence, many differences (mutations) accumulate between the RNA template and the consequent cDNA, ii) As there is no involvement of RNA polymerase III, retrogene usually does not contain the necessary regulatory elements and hence liable to be inactive and (iii) The transposition of retrogenes may be at a different genomic location that may not be adequate for its proper promotion and expression. A few gene duplications are adaptively significant. At times, natural selection favors situations where, more of duplicated gene products is better and gene duplication would likely to be adaptively beneficial. For example: Eukaryotic chromosomes contain reasonably a good number of DNA-binding histone proteins. It is thus not surprising that multiple copies of histone-encoding genes among eukaryotes would have been the result of duplication. Multiple copies of inhibitory KIR genes expressed in human NK cells minimize the host vulnerability to diseases. Inhibitory KIR gene duplication is thus a typical example of immunological adaptation as the same favors the survival advantage of its host. Pathetically, the lack of inhibitory KIR genes does not promote the survival advantage of an individual as he is vulnerable to microbial infections.

Gene Duplication by Exon Shuffling:

Retrotransposition: This is a process by which a natural cDNA copy is made from RNA transcript and
then sporadically inserted into a new chromosomal location. This molecular tool is a very powerful agent in sculpting human genome. 40% human genome consists of retrotansposition-derived repeats and a small percentage of these repeats are actively transposing. Thus, the transposition mechanism has made an important contribution to gene evolution by making copies of exons and shuffling them from one genome location to another, and further facilitating a few genes to be duplicated. LINE1 (Long interspersed nuclear elements) belong to the class of retrotansposons and can transpose robotically like jumping genes of maize. Under experimental conditions, LINE1 elements insert into the intron and they make franking exon to transpose into another gene complex. LINE1 machinery has a weak specificity for its own 3’ end (Illustration 2-Fig.4). This makes it to overlook its own regulatory signals at its 3’ end. Soon after a LINE1 element inserts into an intron of a gene, transcription of LINE1 repeat often by pass its own weak poly A signal and extends to terminate transcription at the flanking host exon 3’ poly A signal. Thus, it can make a copy of a host exon, which can be stitched into another gene during subsequent retrotansposition event (LINE1 mediated 3’ transduction). This mechanism is the basis of exon shuffling, one of the important events in gene evolution (Illustration 2-Fig.4). Thus, molecular evolutionists believe that gene duplication is a critically important process in the evolution of gene functions.

7. The evolution of introns:

Introns are of active interest to molecular evolutionists. At the time of origin of life, there must have been several trails in a variety of molecules in the preparation of a core molecule for the genesis of life. Of the various macromolecules, the molecular organization of pyrimidines and purines and their contribution in synthesizing nucleic acids must have been an incipient attempt and a preferred event among the molecular competitions in the early Paleozoic era. Between proteins and two nucleic acids, the DNA being thermo-stable, double helical and self-replicative in structure, it must have had a rewarding advantage to be a core molecule for the genesis and perpetuation of life. Moreover, DNA is a polymer. It can extend through polymerase chain reaction to accommodate billions of nucleotides and covalently binds to histone proteins for its integration to form a nuclear genome (chromatin). Thus, the genomic DNA must have evolved with an innumerable number of bases, which constitutes the “Molecular World”, the manipulation and maneuvering of which over the time must have led to the existence of the present day biodiversity. Thus, the primordial nucleotide sequence generated would have been the incipient long stretches of sequences, similar to introns, which over the time must have intermittently developed the functional sequences called exons whose transcripts are interpreted by ribosomes. As a result, eukaryotic genes are typically discontinuous in structure, consisting of alternate stretches of coding regions – called exons and non-coding regions – called introns. All prokaryotic genes that have been studied so far lack introns. Two possible hypotheses for intron and exon temporal relations are: The primordial protein coding genes arose initially as interrupted structures. The primordial protein-coding genes were initially uninterrupted sequences of DNA (as observed in prokaryotes) into which introns were subsequently inserted.

Which comes first - either intron or gene? Some introns may be older than genes in which they are contained. A Nobel Laureate, Walter Gilbert postulated that genes in eukaryotic cells arose as collections of exons brought together by recombination within intron sequences. According to Walter Gilbert’s hypothesis, introns provide large stretches of DNA in which recombination events may occur while preserving a gene’s reading frame and thus its encoded function. Gilbert’s hypothesis was supported by two important findings. Firstly, in a number of genes, introns have been found to separate regions of DNA sequences and / or structural domains. A domain is a region of a complex protein sequences that can be associated with a particular structure or function. Secondly, in many cases two closely related genes contain variable numbers of introns. The common ancestral genes have been found to contain at least as many introns (and in the same position) as the descendent gene. Thus, the presence of introns is a very old state of affairs and intron loss may be a common evolutionary phenomenon.

Function of introns:

Introns have the ability to cut themselves out of RNA transcripts either by processing unique autocatalytic activities or by employing specific endonucleases. This ability of introns to control their own excision from nascent transcripts has promoted the hypothesis that at least some introns may have evolved from semi autonomous entities similar to what today we refer to as transposable elements. Further evidence in support of this hypothesis is that, some transposable elements in higher eukaryotes can function as introns. Maize transposable elements carry consensus splice
sequences at their borders that allow transposable element sequences to be spliced out at the RNA processing stage. These findings supported the hypothesis that at least some recently acquired introns might have evolved from transposable elements.

8. Molecular clock hypothesis:

Genes are continuously evolving through various recombinations of nucleotide bases on the primordial long stretches of DNA. Recombinations and substitutions of nucleotides are ongoing processes. All genes (exons) must have undergone a change in their DNA sequence over evolutionary time. The rate at which the nucleotide substitution leads to a new gene function can be evaluated by comparing the sequence changes that exist between closely and more distantly related members of a gene family. Using these findings, one can presumably estimate the evolutionary time that separates contemporary species from their common ancestors.

Examples of Nucleotide substitutions:
The first example considered here is the Insulin gene family. In this, changes in nucleotide sequence occurred over evolutionary time. In a protein molecule, not all regions or domains are having equal functional significance. The amino acids critical to the function of a protein do not undergo changes and the same is conserved by natural selection. Conversely, substitutions at positions encoding amino acids non-essential to protein function can withstand adversities of natural selection. As a result, these positions may change relatively rapidly over evolutionary time. Preproinsulin gene consists of three regions: A, B and C (Illustration 3-Fig.5). A and B regions become the two chains of the functional insulin protein, and C region serves as an ephemeral structural role and contributes disulfide bonds between the A and B-chains (Illustration 3-Fig.5). The C-peptide is enzymatically detached from the mature insulin protein. Because C-peptide is playing an ephemeral structural role, its amino acid substitutions may not lead to functional irregularity. As an outcome, the amino acid substitution is higher in C-peptide than A- and B-peptides over evolutionary time (Illustration 3-Fig.6). Proteins that have rigid amino acid specifications for its structural integrity and function would be expected to evolve at a lower rate rather than proteins that have less rigid requirements. The second example considered here is cytochrome C (Illustration 4-Fig.7). It is one of the electron carriers in the electron transport chain and is an essential protein with a very specific and uncompromised function. The internal hydrophobic amino acid groups have been highly conserved in cytochrome C protein. This molecule interacts at its surface with cytochrome oxidase and cytochrome reductase, both of which are large macromolecular complexes. To accommodate them, the surface (hydrophilic) of cytochrome C molecule has relatively strict structural requirements, especially with regard to the surface distribution of charged groups (Illustration 4-Fig.7). Thus, in essence, the overall rates of substitution are proportionally greater for haemoglobin and fibrinopeptide than for cytochrome C (Illustration 4-Fig.8). The possible explanation is that surface requirements of proteins that do not bind to other molecules (e.g., fibrinopeptide) or proteins that bind relatively small molecules such as oxygen (e.g. haemoglobin) would be expected to have proportionally less inflexible requirements for the conservation of surface structures. Thirdly, the classic example is hemagglutinin gene from flu virus, Influenza: This virus was isolated at different times for over a period of 20 years and maintained in the laboratory. Further, utilizing molecular tools of DNA sequencing analysis, it was calculated the number of nucleotide differences among the chronological sample isolates. Later, an attempt was made to estimate their evolutionary time. The data indicated that nucleotide substitutions have accumulated in this gene at a steady rate. The haemagglutinin gene thus reported to function as a molecular clock.

Evaluation of rate of gene substitution: (Fixation Probability (P) of an Allele and Rate of Gene Mutation (K))
The neutral nucleotide substitutions preferred by natural selection either replace or retain the substituted version of a gene and its functions (Illustration 5-Fig.9). Deleterious substitutions also appear in a gene and cause mutations in a population. This leads to the fixing of a mutant allele either with loss or gain of function. These mutant recessive alleles persist in the population at a very low frequency with recessive traits. On the other hand, there are incidences where a mutant allele completely replaces wild type allele in a population, and the same represents gene substitution and such endurable substitutions contribute for molecular evolution. Thus, a new allele starts entering in a population. How much time does it require for a new allele to fix in a population? The ‘fixation time’ explains the mean time required for a new allele to establish by withstanding adversities of environment. In biological evolution the phenomenon of replacing inadvertently one allele by another is not uncommon.
In the evolutionary process a few genes become extinct, life sustaining genes known as house-keeping genes are relatively uniformly distributed among diversified groups and a few species-specific novel genes are supposed to be new arrivals. Thus, it is of interest to contemplate on the rate of gene substitution, i.e., the number of new alleles adaptively fixed per unit time. An allele, getting fixed in a population, depends upon three factors, viz. Frequency (q), selective advantage/disadvantage (s) and effective population size (Ne). For example, the relative fitness of three genotypes AA, AB and BB are 1, 1+s and 1+2s respectively. The probability of fixation of ‘B’ allele is: 

\[ P \sim q \]  

(Equa.1)

Thus, for a neutral allele, when absolute value of ‘s’ is small, the fixation probability (P) equals to its frequency (q) in the population. For example, a neutral allele with a frequency of 35% will become fixed in 35% of the cases and will be missing in the remaining 65%.

A new mutant arising as a single copy in a diploid population of size ‘N’ has an initial frequency of 1/2N. The probability of fixation of a new neutral allele is therefore:

\[ P \sim \frac{1}{2N} \]  

(Equa.2)

For positive selective advantageous values of ‘s’ and large values of ‘N’ (population size), the probability of fixation of a new allele ‘P’ is approximately twice its selective advantage assuming Ne = N:

\[ P \sim 2s \]  

(Equa.3)

For example, if a new co-dominant mutation (with s=0.01) arises in a population, the probability of its eventual fixation is:

\[ P \sim 2 \times 0.01 \times 100 \sim 2\% \]

A new gene launches in a population because of several reasons namely accumulation of mutations, duplications, shuffling, unequal cross over, etc. Of course, the last three processes are taken care of during meiosis. However, the rate of gene substitutions largely depends upon the number of mutants reaching fixation per unit time. For example, if neutral mutations occur at a rate of ‘u’ per gene per generation, then their number in a diploid population of size ‘N’ is 2Nu per generation. Since the probability of fixation of each of these mutations is ‘1/2N’ (see equation 2), and if ‘K’ is the rate of substitution then:

\[ K = 2Nu.P \]

(1/2N) \[ K = u \]  

(Equa.4)

Rate of Gene Substitution = Rate of neutral mutations per gene per generation.

For more advantageous neutral mutations (s), the rate of gene substitution (K) can be obtained by using Equa.3, which gives probability of fixation of each of this mutation as 2s.

\[ K = 2Nu.P = 2Nu.2s \]  

(P=2s from Equa.3) \[ K = 4Nus \]  

(Equa.5)

Thus, the rate of substitution of an advantageous neutral allele in genetic selection depends on the population size (N), selective advantage (s) and rate of mutation (u).

9. Molecular Trees:

Evolution involves the transformation and splitting of lineages. The genetic differences among species are used to reconstruct their evolutionary history or tree. The assembled data on the amino acid sequence of cytochrome C in a variety of organisms facilitate to construct molecular tree. It is observed that the amino acid sequence of cytochrome C undergoes very slow change during evolution. Their sequences in cytochrome C between human and chimpanzees are identical; there is a difference of one amino acid between humans and rhesus monkeys denoting their divergence from a common ancestral species approximately 20 million years ago. The number of amino acid differences between cytochrome C in humans and a variety of other organisms elucidates that man is closely related to other mammals, and further, cytochrome C in man differs by 13, 36 and 56 amino acids from that of dogs, moths and yeast respectively. More than one nucleotide change may be required to change a given amino acid. Thus, the minimal mutational distance between genes of any two species is established by totaling the number of changes in nucleotide substitutions. The resultant amino acids due to new nucleotide substitutions replace the existing amino acids in the corresponding position in a protein. Molecular phylogeneticists always focus on nucleic acids and/or proteins. They compare them among individuals or species in order to establish evolutionary relationships between organisms or populations. This of course, complements classical phylogenetic approaches. Sequence homology is used to derive quantitative scores describing the extent of relationship between sequences. The similarity approach seeks to maximize the number of matched nucleotides. The distance approach aims at minimizing the number of mismatches. Once the sequence has been aligned, evolutionary trees can be constructed. They are most commonly represented as dendrogram, that depict combinations of lines (branches) and nodes as primary, secondary and tertiary clusters. DNA and protein sequences of extant organisms, which are under comparison, are clustered as pairs of branches and thereby connected to interior nodes (intersections
between the branches) representing ancestral forms for those two clusters comprising of two or more organisms. A rooted tree infers the existence of a common ancestor and indicates the direction of the evolutionary process. An unrooted tree does not infer a common ancestor and shows only the evolutionary relationships among the organisms under study.

Summary:

The vast biodiversity around us is the ultimate product of organic evolution. Phenotypic diversity and genotypic diversity in a population are the result of evolution. There are morphological, behavioural and physiological variants within the phenotypic diversity. The genotype diversity involves the variation in the genetic map, cytogenetic map and physical map of chromosomes. The genetic variation is the raw material for the evolution to occur. The alteration in a gene or genetic variation is the primary cause for organic evolution. The gene is a hereditary unit. It is also a functional unit of DNA. The underlying force for evolution is concealed in genes. Genes undergo mutations. The allelic frequencies of haemoglobin variants such as hbs and ABO blood groups and their balanced polymorphism are the fascinating examples to ascertain the survival of genes. Single or many nucleotide substitutions either by point or frame shift substitutions cause genes to mutate. HLA and KIR haplotypes are unique to the respective individual / population as shown in a few representatives in the present monograph viz., Asian, South Indian, Mediterranean, Turkish and European populations and hence their resistance to infectious and autoimmune diseases vary. Eukaryotic genes are conveniently grouped under varied gene families, each of which is represented by a good number of genes sharing two or more exons of similar or identical sequences (ex. Hb, muscle proteins, etc.). Unequal crossing over is a phenomenon during spermatogenesis and oogenesis. They are the reduction divisions occurring in meiosis. At times, during reduction division due to unequal crossing over, genes may either duplicate or miss in gametes. Yet another process called, retrotransposition, utilizes reverse transcriptase enzyme of infected retroviruses, makes another copy of a gene by assembling a natural cDNA derived through mRNA as a template. The exon shuffling by the so-called jumping machinery such as transposons and small interfering nuclear elements make a portion of a gene called an exon to shuffle. This process makes the gene either to lose a portion of its frame or gain with unwanted frame. Therefore, the gene is the target for all molecular mechanics. The gene may either be positively retained with endurable substitutions or replaced with deleterious substitutions. Consequently, the launching of a new version of a gene takes place in a population subjecting for selection that may determine the fate of a gene either to be advantageous or disadvantageous or neutral. Thus, all events related to sequences of genes constitute the background story in molecular evolution. Ultimately, the hallmark of organic evolution goes with diversified phenotypes and genotypes.

Acknowledgements:

S. Krupanidhi acknowledges with thanks all scientists whom he listed under reference section for having made use of their material and figures as the source for the preparation of the present article.

References:

Illustrations

Illustration 1

Fig. 1: Isoelectric focusing of haemoglobin showing the difference between adult and sickle cell haemoglobins mobility. AA: Adult Haemoglobin, AS: Heterozygote Adult Haemoglobin, SS: Sickle Cell Haemoglobin. Fig.2. Ontogeny of Globular Proteins.
Illustration 2

Fig. 3: Evolution of haemoglobin and myoglobin. Source: muhlenberg.com
Fig. 4: Exon shuffling mediated by transposable elements.
1, 2, 3, I and II are exons. AAAAAA=mRNA
Illustration 3

Fig. 5: Insulin becomes functional by removing C-peptide. The Blue color indicates cleaving site. Source: drknp.com Fig. 6: Histograms showing nucleotide substitution. Y-Axis denotes % nucleotide substitution. X-Axis denotes preproinsulin peptides.

Illustration 4

Fig. 7: Cartoon showing tolerable surfaces for amino acid substitutions in cytochrome C molecule. Fig. 8: Proportion of amino acid substitution over evolutionary time. 1. Cytochrome C, 2. Haemoglobin, 3. Fibrinopeptide.
Illustration 5

Fig. 9: Histograms showing nucleotide substitutions (NS) in various regions of the active centers of DNA. Y-axis = % NS/site/109Years. 1) Pseudogene  2) 3' flanking zone  3) Intron  4) Coding zone (silent sites)  5) Coding zone (Re-placement site)  6) 5' Flanking zone
Disclaimer

This article has been downloaded from WebmedCentral. With our unique author driven post publication peer review, contents posted on this web portal do not undergo any prepublication peer or editorial review. It is completely the responsibility of the authors to ensure not only scientific and ethical standards of the manuscript but also its grammatical accuracy. Authors must ensure that they obtain all the necessary permissions before submitting any information that requires obtaining a consent or approval from a third party. Authors should also ensure not to submit any information which they do not have the copyright of or of which they have transferred the copyrights to a third party.

Contents on WebmedCentral are purely for biomedical researchers and scientists. They are not meant to cater to the needs of an individual patient. The web portal or any content(s) therein is neither designed to support, nor replace, the relationship that exists between a patient/site visitor and his/her physician. Your use of the WebmedCentral site and its contents is entirely at your own risk. We do not take any responsibility for any harm that you may suffer or inflict on a third person by following the contents of this website.