# ON THE EVOLUTIONARY SIGNIFICANCE OF INTRONS

SANJAY TYAGI

87P.

Dissertation Submitted to School of Environmental Sciences, Jawaharlal Nehru University, New Delhi, for Partial Fulfilment of the Degree of

MASTER OF PHILOSOPHY

June, 1982.

result of the heavy chain switch, a single idiotypic determinant (antigen binding portion), is associated successively with different effector sites (constant region) to produce first IgM, then IgD and later IgG or IgA. At first, each of those proteins are expressed as membrane bound proteins and then, finally, at terminal differentiation of B cells, as secretory proteins.

Heavy chain switch from IgM to $Ig^G$, IgA or IgE is known to be mediated by translocation of $V_H$ region to a distant $C_H$ region deleting the intermediate DNA from the genome. Trnaslocation at DNA level, however, can not explain the switch from IgM to IgD because both the proteins may be expressed simultaneously in a B cell, while there is only one functional gene for these proteins. This has been thought to be brought about by alternative splicing of a common RNA transcript (21). IgM and IgD constant region genes are adjacent to each other, separated by only a small length of DNA. A transcript is produced with a read-through to the distal IgD end. This RNA transcript has two polyadenylation signals inside the spacer between IgM and IgD and at the end of IgD (see figure 2).

## CERTIFICATE

The research work embodied in the dissertation entitled, "On The Evolutionary significance of Introns", - has been carried out in the School of Environmental Sciences, Jawaharlal Nehru University, New Delhi and has not been submitted so far, in part or full for any degree or diploma of any University.

Sanjay Tyagi
**Student**

J Subbarao
**Supervisor**

C Varshney
**Dean**

**Dated:** 14.6.1982

**School of Environmental Sciences**
**Jawaharlal Nehru University**
**New Delhi - 110 067.**

CONTENTS

## ABSTRACT

The discovery of highly abundant and widespread, noncoding DNA, in eukaryotic genomes poses an intriguing problem for its evolution. Especially, presence of the class of noncoding DNA elements, found inside the genes and termed as introns is most difficult to explain. The prevalence of introns among most of the genes of all the eukaryotes suggests that they might have been present since the origin of eukaryotes. The present understanding of evolution of eukaryotes, stemming from the work of Woese and collaborators, states that, eukaryotes originated together with prokaryotes from a common ancestor almost immediately after the origin of cellular life. It would be of great interest to see whether the transition from precellular evolution to cellular evolution can lead to the evolution of split genes in eukaryotes and continuous ones in prokaryotes. To accomplish this the state of precellular genomes was approximated using the theory of Error Catastroph and Hypercycle put forward by Eigen and Schuster. An hypothesis is developed about the mechanism of this transition which suggests that in one of the

three lines of descent, split gene organization can indeed result, which would become predecessor of eukaryotes. The hypothesis, implies that presence of introns and several other seemingly unrelated features can be linked with the most important attribute of eukaryotes, the nuclear envelope. Once originated in such a way, it is discussed that they would survive and prevail in all the eukaryotes given the splicing machinary and the nuclear envelope are present. The hypothesis further implicate the reason for the absence of introns in prokaryotes.

# 1. INTRODUCTION

"The line of demarcation between eukaryotic and prokaryotic cellular organization is the largest and most profound single evolutionary discontinuity in the contemporary biological world" (1). This is the most fundamental and natural division of the living systems. Eukaryotes differ from prokaryotes in having their genetic material enclosed in the nucleus, in presence of membrane bound, DNA containing subcellular organelles, in details of the molecular biology of protein and DNA synthesis and, probably, in the regulation of gene expression. Recently, the unprecedented developments in the molecular biology of eukaryotes have uncovered one more

basic difference between the two groups (2, 3). Most of the eukaryotic genes are found split, i.e., they possess DNA, which is never translated, inside their coding regions, while the prokaryotic genes are contiguous and are colinear with their protein products. The noncoding intervening sequences, also called introns, are found not only in protein coding genes but in the genes of tRNA, rRNA also. The broken coding portions, called exons may be found separated from each other by thousands of bases pairs. In addition to this mosaic organization of genes, eukaryotic genomes are flooded with other kinds of non-coding DNA, like repeated sequences, satellite DNA, non-functional genes, spacers, etc., causing what is called the C-value paradox*.

Introns, as a rule, are transcribed with the fellow coding portions - the exons, but the former get spliced out and do not appear in the final RNA transcript. If they do not code for any function of the cell, why do they exist?

---

* The lack of correlation between relative genetic complexity and DNA content is referred to as the C-value paradox (4).

Why have they evolved? Because their function is not clear, their evolution is esoteric and the origin, difficult to explain. They might very well have a cryptic and as yet unknown role in the life of a cell. Or, alternatively, they are 'primitive features' and are yet to be eliminated from the eukaryotic cells. This dissertation is intended to discuss these questions.

Introns are so widespread that they constitute one of the most fundamental attributes of the eukaryotes. Can there be any relation, established between the mosaic gene organization and other features of eukaryotes? As discussed later, a natural and direct relationship can be conceived between introns and the nuclear envelope. A hypothesis is developed in order to discern the origin of introns.

## 2. INTRONS IN PERSPECTIVE

No eukaryote lacks introns. Split genes have been found in organisms as diverse as yeast, slime molds, insects, sea urchins, amphibians, mammals (5) and higher plants (6). These are not only protein coding genes which harbor introns but the ribosomal (7) and the transfer RNA (8) genes have also been found to contain them. The mRNA genes, found split, represent a wide spectrum of proteins : specialized proteins of differentiated cells -globins, crystaline and immunoglobulin; hormone inducible proteins-ovalbumin, conalbumin and lysozyme; embryonic- $\alpha$ -fetoprotein; structural- actin and collagen; hormone proteins - insulin and

growth hormone etc. (5). Interferon and histone do not possess any introns.

The number of introns varies with genes, from two ( $\beta$-globin) to 33 (vitolegenin) and 51 (collagen). The lengths of introns also vary from few bases to several kilobases. Sometimes, the coding portion may be just a minor fraction of the total gene : 90 per cent of $\alpha$-fetoprotein and 20 per cent of collagen genes are noncoding, intervening sequences.

The introns can be found at any position of the genes. They can occur between two codons or can split a codon. They can exist in the coding frame or can inhabit the leader sequence. In some genes they are found between the regions representing two active sites of the protein, while in other proteins they sever its active site (9).

In the course of evolution, introns might stick to their positions relative to some residues of the proteins or translocate to some other position in the same gene or leave it altogether. In case of multigene families, the intron stays in the same position among different members of the family, but

occasionally it is found to have wandered a bit in some genes (10).

The process which removes introns from the primary RNA transcript is formally called splicing. Generally same segments would always be recognized as introns. But in several cases alternative splicing patterns are observed : different portions of the primary transcript are identified as introns. This is generally observed in eukaryotic nuclear viruses (11), and immunoglobulin genes (discussed later). Splicing is very accurate, accurate to the base. But sometimes the sequence at the boundaries of introns are repetitive making the splicing point degenerate. For example, it is possible to splice the pre-mRNA of figure 1 at four distinct cutting positions, all giving rise to the same mRNA (2).

## 2.1 Introns do not Code for any Protein

Since introns are spliced from the pre-mRNA before they have a chance of getting translated, it is believed that they do not code for any protein. This, however, does not exclude the possibility of their translation, separately from the fellow exons, in association with some other genetic element.

Table 1: Number of termination codons in some Introns, in all the frames.

| Introns | Number of Termination codons frame | | | Size |
|---|---|---|---|---|
| | 1 | 2 | 3 | |
| Insulin, small intron of RI gene of Rat | 2 | 1 | 1 | 40 x 3 |
| Insulin, small intron of RII gene of Rat | 1 | 0 | 3 | 40 x 3 |
| Insulin, bigger intron of Human gene | 5 | 5 | 6 | 254 x 3 |
| Insulin, bigger intron of Rat gene | 4 | 5 | 4 | 164 x 3 |
| $\beta$-Globin, small intron of Mouse | 7 | 11 | 7 | 168 x 3 |

But no transcript corresponding to any introns is found in the cytoplasm. A number of introns which have been sequenced, harbor several termination codons in almost all the frames (Table 1). Thus, their translation becomes impossible. Any base sequence which codes for proteins does not show very frequent deletions and additions of polynucleotides which are not in multiplies of three, because such deletions and additions change the frame of the mRNA, rendering it functionless. But when some homologous introns of different species were compared, several such deletions/additions were found. A gene evolving under some functional constraint shows a suppressed rate of base substitution because a considerable number of mutations are deleterious and can not be tolerated in the population. On the contrary, introns have been found changing quite freely, suggesting that they do not have any constraint operating on them. In addition to the above, there is one more evidence, developed from the neutrality theory of evolution (12, 13), suggesting that introns do not code for any protein.

At the time of reproduction, new mutations are poured into the gene pool of a population. Initially, when a mutation enters into a gene pool, its frequency is very low. It can have some evolutionary significance only if it increases its frequency steadily and ultimately replaces the other alleles (a mutation at the same position) completely. This is called the 'fixation' of a mutation. According to the Darwinian theory of selection, if the mutation is harmful it would be eliminated, but if it confers some adaptive advantage for its host it would increase its frequency and get fixed, just because it can alter the reproducibility or life span of the individuals harboring it. But most of the mutations are neither deleterious nor useful. Either they are synonymous mutations (not leading to any change in the aminoacid) or occur at some non-strategic place where a change in the aminoacid can be tolerated. Selection theory does not say anything about this class of mutations; at least, their future in the population is not considered very bright. Majority of mutations are of this kind and are called silent or neutral mutations. According to the champions of neutrality theory of molecular

evolution, neutral mutations can also be fixed, even in the absence of any selection pressure, just by 'random drift'.

It can be easily shown that neutral mutations are fixed with the same rate with which they appear in a population. Consider a eukaryotic genome ($\approx 10^9$ bp); the mutation rate is very low ($\approx 10^{-8}$ per base per generation), and every base has the same probability of mutating; thus, it can be assumed that every mutation occurs at a different site. Let N be the average population size and v the mutation rate per genome per unit time; since two gametes make one individual, the total number of mutations introduced into a population in one generation is equal to 2Nv. Some of these mutations would reach ultimate fixation; let u be the probability that a single mutation would reach fixation. Then the rate of fixation

$$k = 2Nvu.$$

At the time of reproduction, a neutral mutation has an equal chance of getting lost or getting duplicated depending upon the fate of the gamete which carries it. If the gamete forms a zygote it would propagate

otherwise it would be lost. Since the mutation is neutral, the probabilities of its being lost or carried over are equal. Thus, the probability of any gene being found in the future generation is the same. Thus, the probability that a particular gene would be fixed,

$$u = \frac{1}{2N}$$

Therefore, k = v, i.e., rate of base fixation is equal to rate of base substitution (13).

Only the fixed mutations have evolutionary importance, and it follows from the above that if some functional constraints operate on a system, its mutation fixation rate would be slowed down, the fraction of the silent sites being smaller. Therefore, non-functional DNA should change faster than the functional one. Within a gene also, less constrained portions should change faster. The third base of every codon has a greater degree of freedom than the other two because of the degeneracy of the genetic code: it can change without leading to any change in the corresponding amino acid. When a number of homologous genes of related species were compared, it was observed that the third position is more variable than the first and second of each codon (14). However, if two rRNA, tRNA or non-coding

Table 2: Difference in the change of bases in Evolution at First, Second and Third Position of the codons

| Gene | Comparison between | Number of Differences at position | | |
|---|---|---|---|---|
| | | I | II | III |
| Insulin, Coding region | Rat I & Rat II genes | 4 | 2 | 14 |
| Insulin, Coding region | Rat I & Human genes | 10 | 11 | 41 |
| Insulin, Small intron | Rat I & Rat II genes | 3 | 6 | 12 |
| Insulin, bigger intron | Human & Rat II genes | 113 | 113 | 115 |
| β-globin, smaller intron | Mouse & Rabbit genes | 16 | 18 | 12 |
| β-globin, bigger intron | Mouse & Rabbit genes | 90 | 95 | 96 |

DNA, like flanking regions of the genes, are compared no such pattern should be expected. Similarly, in introns also this pattern should not be observed if the introns do not have a coding function. When some homologous introns were compared, all the three positions showed similar variability (Table 2). This proves that the nuclear introns do not code for any protein.

## 2.2 Functions of the Introns

Although it is not necessary for every bit of DNA of a genome to possess a function for the cell, as there can be means by which certain DNA elements can ensure their survival in the genome without coding for any function (15, 16), but the search for functions for new genetic elements, whose function is not apparent, can not be abandoned. In case of introns also, some mechanisms and processes can be conceived which can help in their preservation and perpetuation in a genome during evolution (discussed later), but there are very important functions introns can perform in the life and evolution of a cell. Most of the functions ascribed to introns are of 'evolutionary adaptations' type,

i.e., they facilitate genetic rearrangements thereby increasing evolutionary versatility (2,3,9,17). Other direct phenotypic functions have also been attributed to them, but they are in the preliminary stage of verification.

2.2.1 Phenotypic

The split gene organization can make it possible to produce variants of a single protein by differential splicing of the primary transcript. Eukaryotic DNA viruses customarily use these methods to produce a large number of proteins from their petite genomes (10, 11). They use the splicing as one of the means of regulation of gene expression (18). One primary transcript is spliced in a number of ways, using different segments as exons such that many kinds of mRNA are produced. In vertebrates, the best examples of differential splicing are found in the expression of the antibody genes (19).

During the B lymphocyte ontogeny, which produces antibodies, some sequential changes take place in the immunoglobulin (Ig) molecules. Heavy chain switch and transition from the membrane bound to the secretory Ig are among these changes. As a

result of the heavy chain switch, a single idiotypic determinant (antigen binding portion), is associated successively with different effector sites (constant region) to produce first IgM, then IgD and later IgG or IgA. At first, each of those proteins are expressed as membrane bound proteins and then, finally, at terminal differentiation of B cells, as secretory proteins.

Heavy chain switch from IgM to $Ig^G$, IgA or IgE is known to be mediated by translocation of $V_H$ region to a distant $C_H$ region deleting the intermediate DNA from the genome. Trnaslocation at DNA level, however, can not explain the switch from IgM to IgD because both the proteins may be expressed simultaneously in a B cell, while there is only one functional gene for these proteins. This has been thought to be brought about by alternative splicing of a common RNA transcript (21). IgM and IgD constant region genes are adjacent to each other, separated by only a small length of DNA. A transcript is produced with a read-through to the distal IgD end. This RNA transcript has two polyadenylation signals inside the spacer between IgM and IgD and at the end of IgD (see figure 2).

If the inner poly-A signal is recognized, IgM mRNA is produced; if terminal poly-A signal is recognized a composite RNA is made from which Cμ and spacer can be spliced out using the 'splicer sites' (indicated in figure 2, see the next section) and a contiguous IgD mRNA is generated (21).

Although the transition form membrane bound (M) to secretory type immunoglobulin does not have any direct bearing upon the above mechanism, it can also be accomplished by alternative splicing of a common primary transcript (22, 23). The structure which makes a protein membrane bound is called the signal peptide. It is a highly hydrophobic peptide of transmembrane nature, found on the C terminal of some membrane bound proteins. Curiously enough, signal peptide in IgM is coded by a separate exon. A putative poly-A site and splice sites are found in the intron between Cμ4 exon and the signal peptide exon. When the cell wants to make secretory IgM, the poly-A site recognization is inhibited, now splice site can work and can join the signal exon with the main body of the mRNA, which can code for a membrane bound protein (24) (see figure 3).

Alternative splicing has been found to generate protein polymorphism (production of two sequences of the same protein in the same cell), in ovomucoid proteins of chicken (25).

In addition to providing the opportunity for differential splicing for various purposes, intron might harbor many types of regulatory signals (as already discussed about poly-A signal), for the expression of gene. It has been shown in few cases by construction of chimaeric genes by manually deleting the introns and their splicing junctions, that they can not make any stable RNA if the intron is absent from the gene (26).

### 2.2.2 Evolutionary

The observation that generally the sequences of introns are not very important (because they keep changing in the course of evolution), but lengths are, since there is no trend showing decrease in their length, leads to a very attractive speculation: the presence of introns can facilitate genetic rearrangements which would increase evolutionary versatility (17). The presence of extra length of DNA inside a genome or a gene would increase

the rate of recombination, since the frequency of recombination is directly proportional to the physical distance between the two portions of the same gene. Increase in recombination frequency is desired because, eukaryotes lack some of the means by which lower organisms can shuffle their genes (like horizontal transfer of information etc.)(27). This speculation can very well be tested by comparing variability of various genes with reference to the number and lengths of introns they possess. It follows from this hypothesis that slowly evolving genes would have smaller and lesser number of introns. The observation that histone genes, which do not have any introns (28) and show extremely slow rate of evolution (10), is consistent with this hypothesis. However, nothing can be said about the converse because of the lack of information.

A very elegant idea was put forwarded by Gilbert about the evolutionary significance of introns (9, 15), which seems to be consistent with a number of experiments. According to this thesis, some proteins are composed of several structural or functional units, called domains, which are, more

or less, continuous stretches of aminoacid residues. When structural, a domain can be regarded as a part of a protein molecule which forms a compact, continuous globular region, separated from the rest of the molecule. If the domain is a functional one, it would represent a collection of residues forming an active site or an effector function involved in a particular function of the molecule. 'Introns are there to delineate structural or functional domains from each other', the hypothesis suggests. Each domain should be coded by one separate exon. This kind of organization can give rise to new functions, just by recombining parts of already existing ones (9).

The organization of IgG heavy chain constant ($C_H$) region genes gives a strong support to this idea. $C_H$ can be divided into three well defined domains, $C_{H1}$, $C_{H2}$ and $C_{H3}$, each made-up of about 110 aminoacids. Each domain has different effector functions. $C_{H1}$ contributes to Fab portion of an antibody with light chain constant region. $C_{H2}$ works in complement fixation. $C_{H3}$ is involved in cell surface interactions. The fourth one, a smaller

domain in the hinge region, passes information from Fab to Fc, in complement fixation (see figure 4). It is shown that there are four exons in this gene, and sequences of intron-exon junctions reveal that each domain of the protein is represented by an exon in the gene (29).

It has been shown that domain reagrangement has indeed been taken place during evolution of immunoglobulins (30).

It is possible to correlate exons with structural and functional domains in hemoglobin proteins also (31). The domain pattern in hemoglobin is not obvious. The three dimentional structure does not show very distinct domains. However, diagonal distance maps drawn from its three dimentional structure, show, that, three subdomains or compact structures can be conceived to exist in the molecule (32). When included in the diagonal plot, exon boundaries neatly divide these subdomains, with the exception of the large central exon which show two subdomains (globin has three exons). It suggests, if the hypothesis is assumed correct, that this exon might be a fused product of two exons with the

elemination of the intervening sequence. Remarkably enough, in leghemoglobin gene, an analogous-plant protein, this exon is split into two, the extra intron being at the same place as predicted by subdomain analysis (33). Exon pattern can be correlated with functional domains also (34). Most of the residues in contact with heme and $\alpha_1$ peptide - $\beta_2$ peptide contact correspond to the central exon and $\alpha_1$ peptide - $\beta_1$ peptide contact residues to third exon (34). When the peptide corresponding to the central exon was excised from the rest of the protein and tested for its activity, it was found to bind with the heme tightly and specifically (35). All these observations indicate that exons correspond to structural and functional domains of hemoglobin proteins.

Hen lysozyme, whose exon pattern is known, shows considerable homology with some portions of $T_4$ lysozyme. Although large parts of $T_4$ lysozyme do not have any correspondence in the hen, and similarly, some portions of hen lysozyme do not have any parallel in $T_4$, but a 75 aminoacid stretch is common in both of them and this is where the

active sites of the molecules lie. This stretch is coded by two central exons of the hen lysozyme gene (36). This,observation might appear circumstantial, but if considered in detail and hen lysozyme is compared with other lysozymes, it would support Gilbert's hypothesis.

In case of alcohol dehydrogenase enzyme, two domains can be defined - one binds to the coenzyme while other is responsible for substrate specificity. One of the introns delineate these two sites while other one interrupts one site (35). A similar situation is found in the chicken ovomucoid gene, where three introns delineate three functional domains and one signal peptide, while another set of introns split these domains (25).

If these ideas are correct, then introns can play a very important role in evolution of higher organisms. For the evolution of new functions it was believed that after duplication of a gene, one of the daughter products becomes non-functional and diverges freely from its sister gene. Ultimately, it is possible that it attains a sequence which is capable of performing some new functions. Now, more efficient means are available with the eukaryotes

for this. Take one function from one gene, another from somewhere else, recombine them and get a new function. Introns are there to facilitate this process. For example, for the evolution of hemoglobin - like function, it would be convenient to take a heme binding domain, from a more primitive protein like cytochrome and recombine it with an exon which could give rise to a rudimentary hemoglobin molecule. This type of evolution may be a very significant feature of eukaryotes : the '<u>en bloc</u>' recombination'.

One more evolutionary function has been attributed to introns : increasing the rate of divergence among some members of multigene families. It has been observed in the case of the tandem multigene families that different members of the same family maintain an extremely high degree of homology with each other. This homology is far more than what is expected from the normal rate of mutation (37). This has been thought of as occurring because of 'concerted evolution' between various members of a family, i.e., they keep correcting each other's sequence by matching them. This involves mechanisms like gene conversion and unequal

crossingover, both of which are types of homologous recombination (37). Some members of these families need to diverge faster, and get rid of the discipline of the family. Introns can help them in this pursuit. The variable length of intron would decrease the homology between two members thus, suppressing homologous recombination. Thus, in this way introns may help the outgoing members of multigene families in their departure.

It is clear that presence of introns gives eukaryotes a profound evolutionary advantage over prokaryotes. Evolutionary versatility of eukaryotes would be enhanced because of them. This view of evolutionary adaptation has been criticized on the ground that this is not an adaptation in the Darwinian sense and does not confer any advantage to the organism in its life time or to its population. But if the existence of these elements is granted by, say, self-preserving and perpetuating mechanisms, the descendents of the population possessing introns would gain more diversity and hence, more chances of survival in natural catastrophics. So even without helping in Darwinian evolution, introns

can support an immense amount of diversity in eukaryotes.

## 2.3 Splicing

During transcription, a continuous RNA is made from the gene. Both intronic and exonic DNA are represented in this. However, before translation introns are tailored off and exons are joined precisely. This process is formally called splicing and always takes place inside the nucleus. Splicing has to be extremely accurate, because the mistake of even one base would change the whole frame and render a mRNA functionless. However, in some cases, at the junctions of introns, redundency is observed which can allow the cut to be made at more than one place, thereby making the requirement of accuracy less stringent. Splicing is not mRNA specific, i.e., different type of mRNAs do not have different mechanisms for their splicing. But existing evidences are sufficient to say that tRNAs are spliced differently from mRNAs.

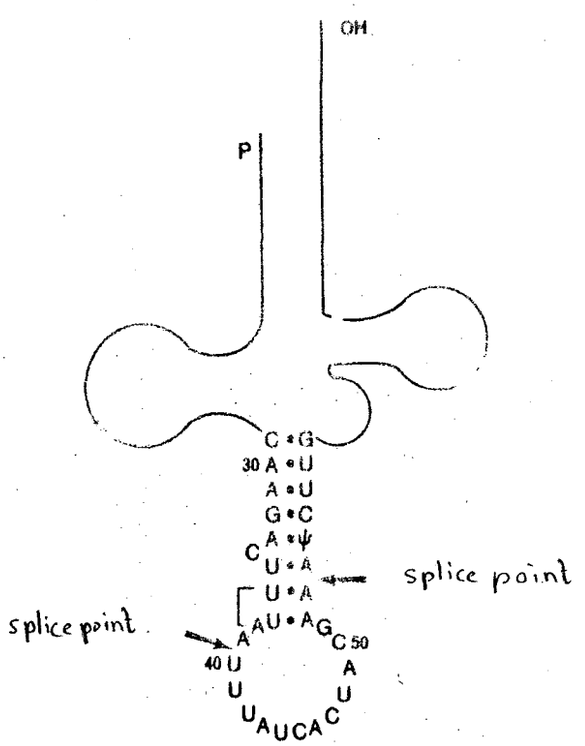



figure 5 presence of intron does not alter the overall configuration of a tRNA molecule

### 2.3.1 Splicing of pre-tRNA

Yeasts, Xenopus, Dictiostelium, Drosophila and silk-moth all have been found to contain introns in their tRNA genes. The intron location is conserved. In most of the cases, the intron is found one base in the 3' direction from the anticodon (see figure 5). Intron size may vary from 13 to 50 base. Presence of intron does not alter its overall secondary structure, only the anticodon loop is base-paired with a complimentry region of intron. Splicing mechanims of tRNA is well characterized (8): splicing apparatus is located in the nucleoplasm (38), nicking and ligation functions can be uncoupled, splicing activity requires ATP and Mg, RNA species corresponding to cut fragments of tRNA and intron can be located (39, 40).

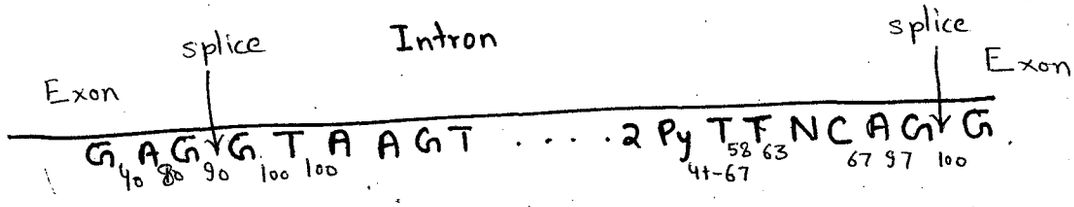
The mechanism of tRNA splicing is fundamentally different from that of mRNA splicing, a yeast mutant is known which is defective only in tRNA splicing, while mRNAs are spliced normally (41), indicating there are two-separate mechanisms. The boundaries of introns do not show any resemblence with consensus sequence of mRNA intron junctions, neither do they

observe AT, GT rule (see the next section). A tRNA precursor mutated at the splice site is spliced accurately (42), while this is not possible with mRNA (43). Probably, the tRNA splicing machinary recognizes the overall secondary structure of the tRNA, cutting of anything that bulges out from anywhere. The sequence of intron is not important, because an insertion of polynucleotide as big as the Lac operator does not hamper the splicing (44). Fortunately, the Lac operator sequence forms a perfect hairpin and thus does not interfere in the overall configuration of tRNA. Splicing machinary is conserved over a wide range of organisms; tRNA from yeast and humans can be spliced in Xenopus and yeast can be spliced in humans (45).

### 2.3.2 Splicing of pre-mRNA

Messenger RNA splicing is more complex than the tRNA splicing. The secondary structure criteria can not be used here, since mRNAs are so diverse and introns are so big and numerous, that every pre-mRNA would show a different base pairing pattern. Instead of secondary structure, the boundaries of introns at the exon junctions are more important here.

All the introns sequenced have GT at their 5' end and and AG at 3' end (46). The sequence inside is not always the same but some amount of aggrement is observed. The consensus sequence is :




The subscript of each nucleotide represent per cent occurrence of the most common base (47, 48, 49). The sequence at the 5' end of the intron is called, 5' splicing sequence (5'SS), and at the 3' end, 3' splicing sequence (3'SS) (45). Splicing sequences are so crucial that a point mutation G → C at 5'SS can stop the splicing completely (50).

Small nuclear RNAs (SnRNA), generally found in most of the eukaryotic nuclei, have been identified as the splicer RNA. As first suggested by Morry and Holliday (51), one of the SnRNA, the U1 RNA contains a contiguous sequence, complimentary to 5'SS and 3'SS in such a way, that intron which is flanked by

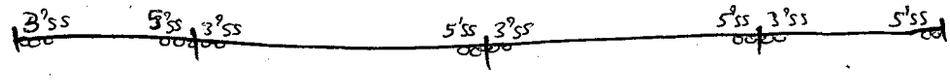
them is looped out and the exon ends come in juxtaposition. In this condition, the endonuclease and ligase can do the splicing (see figure 6).

The SnRNAs which exist in the nucleus as ribonucleoprotein particles (RNP), are found to co-sediment with the large hnRNAs (usually considered as pre-mRNAs),. In a mutant where U1 RNA has lost the sequence complimentary to 5'SS and 3'SS, this co-sedimentation is not observed (48). This suggests that SnRNAs may indeed be the splicer RNAs. Splicer RNA sequences are conserved over a wide range of eukaryotes (48, 52), and 5'SS and 3'SS of mRNAs are also conserved (as discussed before). This suggests the presence of a general mechanism in all the organisms. A chicken ovalbumin gene is correctly spliced and expressed in frog oocytes (53). A chimaeric gene, constructed from a SV40 exon and 5'SS and mouse β-globin exon and 3'SS can get correctly spliced in Xenopus oocyte (54).

Both the splicing sequences are very small and it is very likely that they would be found elsewhere also. On random assignments, 5'SS sequence would occur at every 2000 bases and 3'SS

at every 500 bases. They have indeed been found inside some introns and inside some exons too. A 3'SS is found 6 bases inside an exon in ovomucoid gene, and the cells use this SS occassionally and produce a mRNA, 6 nucleotide short in length; a corresponding protein showing a deletion of two amino acids is also found in the cell (25). The supersplit gene, collagen, containing 51 introns possesses, a number of putative splicing sequences inside an intron. These also have been found to be employed in splicing, since RNAs of intermediate lengths have been found. An interesting thing about these internal splicing sequences is that they are found in pairs, in such a way that whichever sequences are chosen first, for splicing, the resultant transcript also possess splicing sequences, which can be recognized again in a second splicing event; the final splicing would remove the whole intron (55).

I



If only the SS are important in splicing, the possibility of two splicing sites of two introns getting recognized and skipping the exon flanked by them cannot be ruled out. This can be avoided if splicing is processive, i.e., starts from the one end of the script and moves on to the other, splicing the introns sequentially.

Splicing apparatus can be conceived as a ribosome like entity, containing an RNA and some protein molecules. The RNA here would work as a ruler, matching the splicing sites and proteins helping in this, and nicking and ligation. Probably, it would be located at the nuclear membrane. Pre-mRNA would enter from one side and tunnel through the apparatus, to the cytoplasm. This speculation does not have any basis except the observation that no mature transcript is found inside the nucleus.

## 3. EMERGENCE OF INTRONS

A complete explanation for a phenomenon as common and widespread as introns should, not only involve a mechanism for their origin but also a mechanism for their successful maintainence and propagation during evolution. The machinary for splicing appears to be universal; yeast tRNA precursor can be spliced in man (48) and chicken's ovalbumin gene can be expressed in frog (53). This suggests that the ancestors of eukaryotes also possessed introns and had developed mechanisms for their splicing. The fact that prokaryotic genes and some eukaryotic ones do not contain introns, does not necessarily indicate that split

genes arose from contiguous ancestral genes. In fact, theories have been put forward suggesting that split gene organization is a primitive feature and prokaryotes represent comparatively more evolved organisms, which have got rid of their introns (3, 56). But there is no reason for such an a priori assumption. An hypothesis is developed, systematically, in this chapter, in order to explain the evolution of introns, which addresses itself to the question of the origin and perpetuation of introns separately.

## 3.1 Origin of Introns

Universality of introns in eukaryotes suggests that they arose with the emergence of the eukaryotes. Considering the currently accepted view of origin of eukaryotes and prokaryotes (57), and a widely appreciated mechanism of precellular evolution (58), a hypothesis is made to explain the origin of introns. Following are the postulates of the thesis, each of which is discussed in detail in later sections or chapters :

1. Prokaryotes and eukaryotes share a common ancestor, the 'progenote state', rather than eukaryotes having a prokaryote type predecessor.

2. The low fidelity of early replicases would limit the length of the primeaval replicons. Thus, the genomes of the progenotes existed as small, distinct fragments of nucleic acid.

3. 'The fragments' were circular nucleic acid molecules, containing one or few genes. The distinct circles were linked with each other only by a 'functional' link.

4. When the fidelity of replication increased after a period of evolution, longer sequences were permitted and integrations of the fragments were allowed. Rudimentary 'illegitimate recombinations' and corssing overs would mediate integration of all the fragments into one big 'chromosome'.

5. Integrations were probably random and were target unspecific at this early stage of evolution and would invariably result

in splitting of genes with a minor fraction of integrations between the genes.

6. Transition from the RNA genome to the DNA genome took place at this stage and two (probably three) (59) lines of ascent diverged, depending upon their strategy to overcome the problem of the broken genes. The organisms which were to become prokaryotes, selected genomes with intact genes (integration between genes). The individuals which could use integrates with broken genes and splice their broken messages with the help of the copies of unsplit RNA fragments, developed a nuclear envelope to separate the transcription compartment from the translation compartment and became 'urkaryotes' the predecessors of eukaryotes.

7. Development of nuclear envelope, which was a necessity for an efficient use of broken genes, allowed, further prevalence and expansion of the split genes and the

splicing machinery which in turn helped eukaryotes to acquire the immense amount of adaptability.

## 3.2 The Fidelity of Replication Limits the Length of the Message

The second statement of the hypothesis follows from the behaviour of ensemble of self-replicating informational molecules. It can be derived quite naturally that the number of molecular symbols of a self-reproducible unit is restricted, the limit being inversely proportional to the average error rate of the symbol reproduction (58, 60, 61). An hypothetical evolutionary reactor is used in order to show this.

Consider the reactor in figure (7). There are n species of polynucleotide $I_1$, $I_2$,... ...$I_n$, in the solution, each defined by its length and its sequence. Polynucleotides replicate using energy-rich materials as substrates and themselves as templates. Total concentration of nucleic acid is kept at a constant level by a flux $\phi$ . Energy-rich materials are constantly added to and energy poor materials removed from

the reactor.

The replication which may be catalysed by an enzyme, replicase, is associated with a finite error rate. The probability that a nucleotide is replicated correctly is $q$ (assuming that it is the same for all four of them). If the $i$th species has $V_i$ nucleotides, then the probability is

$$Q_i = q^{V_i}$$

(assuming that the errors are independent of each other), which means that fraction of correct polynucleotides produced by $i$th species is $Q_i$.

The rate of change of $i$th species is of the form :

$$\frac{dx_i}{dt} = A_i Q_i x_i - D_i x_i + \Sigma_{k \neq i} w_{ik} x_k - \phi_i \qquad [1]$$

where - $i$ is the running index, 1,2,3,....., n, attributed to all distinguishable sequences; $x_i$ is the concentration of $i$th species; $Q_i$ is the fraction of correct sequences produced by $i$; $A_i$ is the adaptability of $i$th sequence to the replicase and depends upon the concentration of energy rich materials too; $D_i$ is the spontaneous

decay rate of $i$; $w_{ik}$ is the rate of reverse mutation from $k$ to $i$; $\phi_i$ is the flux of $i$ out of the reactor.

If $\phi_t$ is the total flow out of the reactor, the flow of $i$ would be

$$\phi_i = \phi_t \frac{x_i}{\sum_k x_k} \qquad [2]$$

In the reactor, $\phi_t$ is adjusted in such a way that

$$\sum_k x_k = \text{constant.}$$

In order to satisfy this constraint, $\phi_t$ has to be steadily regulated to compensate for the overall excess production.

$$\phi_t = \sum A_k x_k - \sum D_k x_k = \sum (A_k - D_k) x_k$$

$$= E_k x_k$$

$E_k$ can be considered as the excess productivity. From equation [2],

$$\phi_i = \frac{\Sigma E_k x_k}{\Sigma_k x_k} x_i$$

On the substitution of $\phi_i$ equation [1] becomes

$$\dot{x}_i = \left[ (A_i Q_i - D_i) - \frac{\Sigma E_k x_k}{\Sigma_k x_k} \right] x_i + \Sigma_{k \neq i} w_{ik} x_k$$

$$\dot{x}_i = (W_{ii} - \bar{E}(t)) x_i + \Sigma_{k \neq i} w_{ik} x_k \qquad [3]$$

where, $W_{ii}$ is called the selective value of $i$, $\bar{E}(t)$ is the average excess productivity.

If equation [3] is represented in matrix form $W_{ii}$ and $w_{ik}$ can be summed into one matrix

$$W = \begin{bmatrix} W_{11} & w_{21} & \cdots & w_{n1} \\ w_{12} & W_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ w_{1n} & \cdots & & W_{nn} \end{bmatrix}$$

and

$$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

therefore,

$$\dot{X} = WX - \bar{E}(t)\,X \qquad [4]$$

Let $Y = KX$

$K^{-1}Y = K^{-1}KX$

$\Rightarrow \quad X = (K^{-1})\,Y$

where, $K$ is a nxn matrix

$$y_1 = \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1n}x_1$$

$$y_2 = \alpha_{21}x_1 + \alpha_{22}x_2 + \cdots + \alpha_{2n}x_n$$

$$\vdots$$

$$y_n = \alpha_{n1}x_1 + \alpha_{n2}x_2 + \cdots + \alpha_{nn}x_n$$

$$Y = \begin{bmatrix} \alpha_{11}\,\alpha_{12} & \cdots & \alpha_{1n} \\ \vdots & & \\ \alpha_{n1}\,\alpha_{n2} & \cdots & \alpha_{nn} \end{bmatrix} X$$

From equation [4],

$$(K^{-1})\,\dot{Y} = W K^{-1} Y - \bar{E}(t)\,K^{-1} Y$$

$$\dot{Y} = KWK^{-1} Y - \bar{E}(t)\,Y$$

$$KWK^{-1} = W_d$$

Let $K$ be the matrix that diagonalises $W$. Let the eigenvalues of $W$ be $\lambda_1, \lambda_2, \ldots \lambda_n$.

$$\dot{y}_i = (\lambda_i - \bar{E}(t))\, y_i \qquad [5]$$

$\bar{E}(t)$ remains invariant after the transformation, and its value becomes -

$$\bar{E}(t) = \frac{\Sigma_k \lambda_k y_k}{\Sigma_k y_k}$$

because, $\Sigma \dot{y}_k = 0$, as the condition of constant total concentration dictates.

Equation [5] tells how the reactor would behave in time. Each species and a distribution of closely relative species, derived from it by erroneous copying which are similar in sequence and adaptibility to the enzyme would act like one type, called a 'quasispecies.' A quasispecies whose

$\lambda$ is below $\bar{E}(t)$ would decrease its rate of production and ultimately die out. The one with $\lambda$ higher than $\bar{E}(t)$ would grow but simultaneously help the $\bar{E}(t)$ to increase. $\bar{E}(t)$ would steadily increase and the number of quasispecies with higher than it would decrease. Ultimately this would result in the selection of only one quasi-species which would always keep its $\lambda$ higher than $\bar{E}(t)$.

$$\bar{E}(t) \longrightarrow \lambda_{max}. \qquad [6]$$

To calculate the value of $\lambda_{max}$ in terms of known parameters consider the matrix $W$. Its structure is such that the diagonal $W_{ii}$ elements do have some finite value while the non-diagonal elements have magnitudes which diminish as one goes away from the diagonal. This is because the probability that distant relatives give rise to by reverse mutation is very low. Only those elements would contribute to $\lambda$ which are in the immediate viscinity of the diagonal. Eigenvalues of such

a matrix by second order perturbation theory can be given as -

$$\lambda_{max} \approx W_{mm} + \sum_{k \neq m} \frac{w_{mk} + w_{km}}{W_{mm} - W_{kk}}$$

Assuming that the second term is very small as compared to $W_{mm}$ (the assumption is valid, as shown by the study performed by Eigen and Schuster (58)), and can be numerically neglected,

$$\lambda_{max} < W_{mm}$$

From [6],

$$\bar{E}_{k \neq m} < W_{mm}$$

from [3]

$$\bar{E}_{k \neq m} < A_m Q_m - D_m$$

$$Q_m > \frac{\bar{E}_{k \neq m} + D_m}{A_m}$$

$$Q_m > \sigma^{-1}$$

$$Q_{max} > Q_{min} = \sigma^{-1}$$

$$q^{V_{max}} = \sigma^{-1}$$

$$V_{max} = \frac{\ln \sigma_m}{1 - q} \qquad [7]$$

Equation [7] is an important generalization. "The number of molecular symbols of a self-reproducible unit is restricted; the limit being inversely proportional to the average error rate per symbol : $\frac{1}{1-q}$ (58). This theory is consistent with some experimental observations. The RNA viruses, prokaryotes and eukaryotes possess replicases (DNA polymerase and repair/recombination functions), of different fidelities. Their genome sizes never exceed the threshold predicted using the fidelity of replication. In case of the $Q\beta$-phage genome, the size is exactly the same as predicted, because of the optimization involved in the phage cycle, demands more and more genetic material while accuracy of the replicase limits it. In the evolution experiments with replication system, where in vitro self-replication is performed on the lines of the reactor just discussed, any sequence leads to the selection of a quasispecies which is always the same for the replicase used (62).

Fidelity of the earliest replicases must have been very low because of obvious reasons. This would limit the length of the distinct sequences. Any self-sustaining, self-replicating system has to

have a minimum of few genes, i.e., for translation and replication functions. This would demand existence of longer messages or many, small, distinct messages. As we have already seen competition can allow only one sequence to exist.

3.3 Small Sequences of Distinct Messages can Co-exist If Linked Through a Hypercycle :

We are faced with a dilemma that the length of the genes/genome is restricted because of the accuracy of the replicase, and the accuracy of replicase can not be increased without increasing the length of the gene coding for it. The problem however, can be overcome by introducing some cooperation between distinct quasispecies. It can be shown that only a second order, autocatalytic and cyclic cooperation can bring stability and coexistence of two distinct nucleic acid sequences (58). On introduction of a second order, cyclic cooperation term into [3],

$$\dot{x}_i = \left(W_{ii} - \bar{E}(t)\right) x_i + \sum_{k \neq i} W_{ik} x_k + Z_{i\,i-1} x_i x_{i-1}$$

$i = 1,2,3,\ldots,N.$
$N + 1 = 1.$

This type of second order cyclic cooperation has been termed as 'Hypercycle' (58, 60, 63-65). Hypercycles consisting of several members have been studied analytically and it has been shown that hypercyclic organization can ensure stability of larger quantity of information, using a limited accuracy of replicase (58).

What can be the physical source of cooperation between two nucleic acid sequences? For the purpose of illustration, let us consider a two member hypercycle. $I_1$ and $I_2$ are the two nucleic acid species which contain as much information as permitted by the threshold (figure 8). $I_1$, on translation, produces an enzyme $E_1$ which replicates $I_2$; $I_2$ has a folding pattern which makes it an ideal substrate for $E_1$. $I_2$ can produce replicase $E_2$ which helps in the replication of $I_1$. $E_1$ and $E_2$ are the agents of cooperation and work either by increasing the rate of replication or by reducing the accessibility of their target nucleic acid towards hydrolytic cleavage. This is the kind of organization most primitive kind of hypercycles would have (63).

In an advanced stage of precellular evolution when replicases had gained a fidelity of the order of $10^{-3}$ or $10^{-4}$, the individual sequences could be as big as two or three rudimentary genes. These sequences would probably be circular molecules of RNA. Here, both the points deserve some elaboration because of their pivotal importance in the hypothesis. Molecular evolutionists favour RNA over DNA for premordial genetic material on account of two arguments. Firstly, oxy-sugars are more readily formed than deoxy-sugars in the simulated prebiotic synthesis experiments and secondly, RNAs can become more diverse targets of selection at the genotypic level also. Whether a nucleic acid sequence is selected or not in evolution would depend upon the proteins it makes, but if it is RNA it can be selected or rejected by the virtue of its folding pattern also, which determines its suitability as template for the replicating enzyme. Folding pattern of a RNA species, also determines its degradation rate : more the number of bases paired less liable it is to the hydrolytic attacks. This is crucial for its survival in a population of self-replicating molecules.

The assertion that these fragments were circular is made again for two reasons. The circular molecules are more stable and can withstand rigours of hydrolysis. There is no free end. To remove a base two bonds have to be broken and two stacking interactions have to be interrupted; while in the linear DNA, gradual chewing at the end is very easy. Since the same replicase has to replicate both + and - strands, both of them should have a folding pattern recognizable by the enzyme. Symmetric arrangement of the recognization sites is responsible for this in the RNA phage, like Qβ . It is easier to have symmetric organization in a circular molecules than in linear ones.

## 3.4 Increase in the Fidelity of Replication Allow Integration of Fragments into One Genome

As the hypercyclic precellular systems evolved, the fidelity of replication improved. This could have taken place by sorting out the best sequence for the basic replicase and by introducing one or more proofreading functions into the enzyme. Increase in the fidelity would permit existence of bigger sequences. The structural

integration of individual fragments would become permissible. This can very well take place by fusion of, otherwise hypercyclically linked circular fragments. The fusions would be probably random, and mediated by rudimentary, 'illegetimate' and non-homologous recombination events and by unequal crossing-overs. As discussed below, this is the time when the transition from a single stranded genome to a double stranded genome would take place, and thus, above processes would be possible.

If they are random, most of the integrations would take place inside the genes with a minor fraction going between genes. This would result in gaint chromosomes, with a number of genes broken because of the presence of other genes in them. A small fraction of integrates would, however, have all the genes intact. This would correspond to the fusions taken place only between the genes.

Integration of small fragments into one gaint chromosome would be of immediate advantage for the system. In a hypercyclic organization, the segregation of genetic material at the time of division of the cell would be very inefficient one. When all the members of the hypercycle are multiplied, they

would devide, and very seldom both the daughter cells would get a full complement of the genetic material. Most of the time one daughter cell would receive all the members of the hypercycle, and other one would lack some members and hence would soon disintegrate. The integration into a single replication unit would solve this problem where all the members of the set are structurally linked.

This is the time when transition from RNA to DNA genome would take place. The argument which suggests this is as follows : The selection operates at both phenotypic and genotypic levels in the hypercyclic state. An RNA species is selected both because of its translation products and its target function. The former being an indirect 'effect' of the sequence of the gene and the latter as a direct manifestation of its sequence. The target function means how efficient a substrate the RNA is as the template for the replicase produced by the previous member of the hypercycle. This imposes a restriction of dual functions on the evolution of the sequence: it has to work as a target of a replicase and has to produce a replicase. A mutation which can be

advantageous for the translation product would not be allowed just because it makes the RNA a less efficient target for the replicase. The single stranded nature of RNA generates myriads of folding patterns. However, when the fragmented genome is structurally integrated into a gaint chromosome, the hypercyclic couplings become dispensible and hence single strandedness is not required anymore. The system would readily turn to double stranded genome and shed off the restriction of the 'genotypic selection'. It is known that DNA is a better form for the double strands than RNA, so a transition from RNA to DNA would be favoured. But initially, the system must have managed with RNA double strands only and later gradually shifted over to DNA.

The double stranded genome would be favoured because it can help in increasing the fidelity of the replication process. Some of the correction mechanisms like post-replication repair, etc., are possible only with the double strands. In such processes if a mismatch is encountered by the repair enzymes, the erraneous nucleotide is removed with the probability of 0.5. However, if a repaired strand is subjected to homologous recombination,

the chance of the erraneous nucleotide surviving in the progeny is reduced to only 25 per cent (because there also the correction is done with 50 per cent surity, see reference 58).

The prebiotic translation mechanism would also favour the differentiation of the genetic materials into the hereditary and the messenger molecules, the later being responsible for phenotypic expression only. The models of prebiotic translation (66), proposing the concept of the static template surfaces for translation (66, 67), are particularly relevant in this context. According to this model a messenger molecule is permanently occupied by the the adapters (proto-tRNAs), such that it is never free for replication. For replication, a different molecule of the same 'quasispecies' is chosen. It means that two different molecules of the same sequence perform the functions of the heredity and the phenotypic expression from the very beginning. It would be highly advantageous for the system if the molecules are adapted to their respective functions, without changing their sequences. Hereditary material must be very stable and it should be possible to achieve high fidelity with this; the

double stranded DNA is a powerful candidate for this. The messenger function can be performed only by a single stranded nucleic acid: it facilitates decoding and high stability whereas, high fidelity is not at all crucial for it. Thus, RNA and DNA differentiation should take place at this time.

DNA is for more stable molecule than RNA, firstly because it is double stranded (very few groups are exposed for hydrolysis) and secondly the internucleotidal bonds in DNA cannot be broken easily in alkaline medium, while mild NaOH can break down, RNA into monomers.

## 3.5 The Divergence of Prokaryotes and Eukaryotes

The postulated hypercyclic organization represent the 'progenote state' (see the next chapter), from which the prokaryotes and eukaryates have developed. After the spree of integration, two types of chromosomes would result; with split genes and with intact genes. Initially, these chromosomes would coexist with their fragmented counterparts (all of them do not get integrated), the single stranded RNAs. The individuals (cells) which had chosen the intact chromosomes became the

ancestor of prokaryotes. Others which had chosen their genomes from the immense variety and quantity of chromosomes with split genes became the ancestors of the eukaryotes. Initially, the second line of ascent was extremely difficult to start, but once it took off it would enjoy the boundless amount of adaptability offered by eukaryotic organization. The initiation is difficult because it requires the emergence of splicing as a precondition. A tentative mechanism however, can be suggested taking the advantage of the possibility, that unsplit individual RNA fragments also co-existed with the gaint chromosomes of split genes. The messenger transcript produced after the transcription of split gene would anneal with unsplit RNA fragment in the regions corresponding to the coding portion. The extra-portion would loop out and could be cut and the ends be ligated (figure 9).

This is a scheme different from the contemporary splicing apparatus, which in case of tRNA utilizes overall secondary structure of the pre-tRNA and in case of mRNA, the boundaries of the introns. This

mechanism has obvious advantage over the one postulated for the premordial splicing, which would have given way to the contemporary one quite promptly. The contemporary mechanism set the coding portion free from the constraint of the splicing function, while the primordial mechanism demands that coding sequence must not be too different from the splicer RNA.

The onset of splicing has to be associated with the localization of the genome in a separate comparatment. This is necessary because in an one compartment system (like contemporary bacteria) translation and transcription take place simultaneously. The intronic DNA would be translated together with the coding one before it has a chance of getting spliced. With the lineage leading to bacteria, however, there is no problem. The problem of compartmentation is discussed in the next chapter in greater detail.

## 3.6 Perpetuation of Introns

Having described the mechanism of origin of introns, one has to consider their maintenance too. Cells certainly have some mechanisms for the removal

of its extra DNA, for example, some times in somatic cell differentiation, some specific segments of DNA are deleted from the genome. No trend has, however, been observed, in the direction of decreasing the amount of non-coding intervening DNA. Once the splicing machinery is established and a nuclear envelope is formed, introduction of new introns would be tolerated. Introns must be continually added to the system. The total DNA of a cell should represent a dynamical equilibrium state of deletion and addition process. Occassionally DNA elements can propagate in a genome even if they do not confer any phenotypic advantage to the system (12). Some of such elements (Selfish elements) do this either by hitch-hicking to some crucial gene or by possessing a property of self-propogation (like transposition) inside a genome. If an element confer some selective advantage of 'evolutionary adaptation' type and its deletion mechanism is not very prompt it would spread in the population (15, 16). Introns probably represent this class of DNA, they are tightly bound to the gene they inhibit, they confer evolutionary advantages to the whole genome and to the gene (see

Section 2.2.2).

Two types of mechanisms can be conceived for the evolutionary propagation of introns, which are consistent with some observations.

## 3.7 Amplification of Already Existing Introns

Primordial cells must have had exceedingly low number of genes, as compared to the modern cells. The additional genes which are seen today have been derived from the archaic stock. And if the hypothesis that predecessors of the eukaryotes had their genes split is correct, than the genes which are derived from them should also be split.

Conventionally it was believed that new genes arise after duplication of an old one followed by a drift in its sequence by mutation and recombination. Ultimately, by chance it attains a sequence which on translation can make a useful protein. This process is highly implausible, since the number of sequences, a normal gene length can attain, approaches astronomical figures, and the meaningful sequences are a very minute fraction of that number. Resultantly on the evolutionary time scale a population will never be able to test sufficient number

of them. The more favoured idea of the *en bloc* recombination of already existing functions is used to generate new functions (as discussed in Section 2.2.2). This method of production of new genes would inevitably give rise to the new introns. Following processes can generate and at the same time lead to the origin of the new introns.

1) When genes duplicate by usual methods like, unequal crossingover, etc., it is highly improbable that duplication is exact, i.e., no extra DNA gets duplicated between them. Some spacer would inevitably appear. If, later in evolution, the spacer, together with the termination signal of the first gene attains a sequence of splicer sites and the joint product of the two genes makes a meaningful protein, the composite gene would become functional and the spacer would be established as an intron (figure 10).

The structures of immunoglobulin genes suggest that this might have been the first step in their evolution (29).

2) Insertion of an exon with its flanking regions (parts of introns), inside an intron might prove

beneficial for the construction of a new gene. This process results in the formation of a new intron (figure 11).

Hemoglobin gene could very well have arisen by such a procedure, the exon corresponding to heme binding portion being recruited from a gene like cytochrome (see section 2.2.2).

3) An exon within a gene, on duplication with its flanking regions would generate a new gene, with an additional intron and exon. One possible mechanism may be the unequal crossingover mediated through respective elements present inside the introns (68)(see figure 12).

This duplication can take place by crossing-over between the two exons also, if any homology exist between them (figure 13).

Several genes could have evolved by this mechanism. Collagen gene represents an interesting case strongly suggesting the occurence of such unequal recombination events in its evolutionary history. The gene has a total of 51 introns. The Middle 14 exons, correspond to a region of protein where three $\alpha$-helices supercoil to form a triple

helix (69)(see Lehinger, Biochemistry for the Structure of the Protein). This region contains repeats of aminoacid triplet Gly-X-Y, where X and Y are often proline and hydroxyproline. Each exon in this region is a multiple of 9 bp, which code for the repeating triplet. Out of 14 exons, two contain 45 bp, seven contain 54 bp, three 99 bp, and two 108 bp. Large and small introns are alternated in a pattern : 99 - 45 - 99-54- 108 - 54 - 99. Aforesaid type of duplication must have taken place in this gene, since unequal crossingover between different exons would be a very likely event (because of their homology). If recombination between two exons at non-identical positions takes place inside the 9 bp repeat, it would result in disruption of the helical structure of the collagen, and thus be deleterious. Therefore, exon size would always be kept as multipleof 9 bp (69).

A set of exons can get duplicated together. This has probably taken place in $\alpha$-Fetoprotein and albumin genes, where a set of 4 exons have triplicated jointly (70). Structure of conalbumin gene is also suggestive of a similar duplication : it can arise

by a duplication of an ancestral gene already split in seven or eight pieces (71). The Ovomucoid gene may also be a product of an ancient intragenic duplication (25).

## 3.8 Insertion of New Introns

Not in all the cases is it possible to explain the existence of introns by duplication or continuation of already existing ones. In these cases when homologous genes are compared it is found that introns positions are not conserved or sometimes an intron is missing or has been added at a new position. For example, in actin gene family, introns interrupt various genes at different positions and genes from different organisms also show sometimes different locations of introns (10). There is no obvious domain - exon correspondence in this protein. This can only be explained by random deletion or insertion of a few introns. In other genes also, where introns are randomly arranged and change their locations, it may be necessary to postulate the insertion or deletion of introns.

This kind of insertion can easily take place through a number of illegetimate recombination

methods, mediated by transposable elements, DNA viruses or retroviruses, etc., (72, 73). Since many copies of the genes exist, insertion or deletion in one gene would not be lethal for the individual, immediately.

# 4. THE CONSEQUENCE OF NUCLEAR ENVELOPE

## 4.1 The Three Line Descent

Conventionally it was believed that eukaryotes arose from an ancestor, similar to the contemporary prokaryotes, quite recently in the early precambrian epoch. The assumption was made basically because of the difference of the level of complexities of the two classes. Since the eukaryotes have more complex cellular organisation and genome, they must arise from the prokaryotes which are simpler (74, 27, 75). This view of early cellular evolution is undergoing extensive revisions (75, 76, 59, 77-79). Woese, Fox and

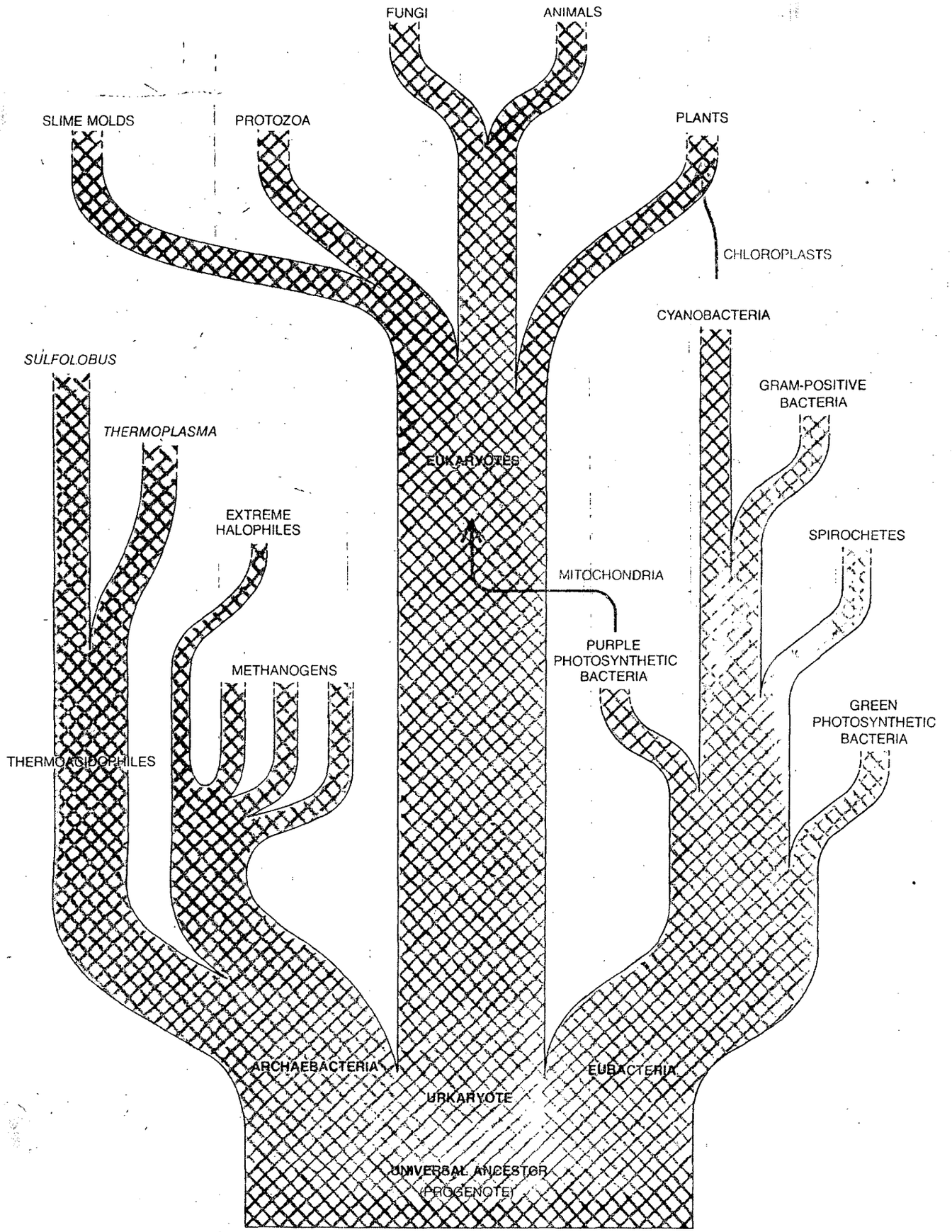

FUNGI
ANIMALS
SLIME MOLDS
PROTOZOA
PLANTS
CHLOROPLASTS
CYANOBACTERIA
SULFOLOBUS
GRAM-POSITIVE BACTERIA
THERMOPLASMA
EUKARYOTES
EXTREME HALOPHILES
SPIROCHETES
MITOCHONDRIA
PURPLE PHOTOSYNTHETIC BACTERIA
METHANOGENS
GREEN PHOTOSYNTHETIC BACTERIA
THERMOACIDOPHILES
ARCHAEBACTERIA
EUBACTERIA
URKARYOTE
UNIVERSAL ANCESTOR (PROGENOTE)

Figure 14

others have propounded a theory, the three line descent theory, that both eukaryotes and prokaryotes, arose together, with a third group called 'Archbacteria', from a common ancestor very early in the history of life (59, 77, 78). The ancestral form which was far simpler than any of the prokaryote living today has been called as the 'progenote state'. The three groups which represent the most fundamental trichotomy of life are - the archbacteria including methanogenes and halobacteria, the eubacteria, involving common bacteria and cynophytes and the eukaryotes. In the beginning, eukaryotes had only the nucleus as the attribute of being eukaryotic. This form has been termed as 'urkaryote'. The cytoplasmic orgenelle owe their origin to the lineage of eubacteria, some members of which established symbiotic relationship with the nuclear host and ultimately became an unseparable component of the eukaryotic cell (figure 14).

The three line descent theory, unlike the conventional view is supported by some experimental evidences. The primary structure of ribosomal RNAs from a large number of organisms and organelles have

been characterized and compared. The comparisons were done with the oligonucleotide sequences, obtained from the T1 endonuclease digestation of rRNA, rather than with the complete sequences. The comparisons reveal that the sequences cluster into three conspicuously distinct groups. While the members of the same group are extremely similar to each other, the different groups show such a vast difference that they are called, the three primary kingdoms. They correspond to archbacteria, eubacteria and eukaryotes. This observation has been explained by the hypothesis -

> "The three lineages diverged at a presumably very early stage when genetic organization and machinary for gene expression were undergoing rapid Darwinian evolution in the direction of increased effeciency and accuracy. The profound differences between these lineages are best interpreted as independently achieved solutions to the problems of how genes should best be organized and expressed".(75).

The time of divergence can be assumed to be at least 3 billion years or more because that

is when earliest blue green algae are found (80). In addition to these genealogical evidences, difference at the phenotypic level between archaebacteria, eubacteria and eukaryotes are compelling enough to make us postulate the three line descent hypothesis. There are strong reactions to the theory, but most of them are semantic in nature (Reference 81 is a more serious criticism).

## 4.2 The Hypothesis is Consistent with the Three Line Descent Theory

One of the most immediate implications of the hypothesis presented in the previous chapter, for the origin of introns is that the eukaryotes and prokaryotes arose simultaneously from a common ancestor. As is clear from the previous section, the three line descent theory says exactly the same. At the time of transition from 'hypercyclically' organized, fragmented genome (RNA) to an integrated genome (DNA), the lineages of prokaryotes and eukaryotes took off. The lineage of the ancestors of eukaryotes (the urkaryotes) had introns, splicing machinary and the nucleus, while the prokaryotic lineage had a bare chromosome with unsplit genes.

If true, this hypothesis throws some light on the mystical progenote state. The early self-replicating systems with hypercyclic organization can be considered as the progenote state. This state would be characterized by, fragmented but hypercyclically linked genomes, single stranded (circular) RNA as the genetic material, a rudimentary cell membrane and anaerobic heterotrophism. It is considered anaerobic because, no oxygen was present in the atmosphere and heterotroph because it could utilize energy rich monomers abiotically synthesized by the action of UV, etc.

The most fundamental and universal attributes of contemporary living beings, like the genetic code and translation apparatus, etc., must have evolved to at least their rudimentary form in the progenote state. This is expected because of the universality of these features in all the three groups. The rest of the features which differ within the groups have been acquired later by the individual lineages.

As far as the origin of archbacteria is concerned, it can easily be speculated that they

belong to the same stock which selected unsplit, integrated chromosomes but because they had chosen the anaerobic habitates they diverged from the other prokaryotes. Both urkaryotic and prokaryotic organisms must have contributed to the transition of the reducing atmosphere into an oxidising one. The gradual transition would diminish the regions of low oxygenecity which ultimately would get restricted to places like marshes, etc. Thus, the sphere of archbacteria would become limited to such habitates. Their independent evolution would shape their biochemistry into its present make-up.

Should the organisms representing the progenote state, be found today? It is very unlikely. It is unlikely, since hypercyclic organization can not compete with structurally linked genomes, and would attenuate to oblivion (58).

## 4.3 The Nuclear Envelope

According to the hypothesis, developed in the previous chapter, the survival of gaint chromosomes with split genes requires coevolution of an envelope around it. It is necessary to separate the transcription compartment from the translation

compartment. If the same compartment is used for both the processes, translation would start while transcription is still going on, and the intron would be translated even before it had a chance to get spliced. Bizarre translation products would be formed. Similar thing would happen to ribosomal and transfer RNA also, for example, if a rRNA gets assembled into a ribosome, with its intron in, the resultant ribosome would be functionless. A solution of this problem is : separate the compartment of translation from the compartment of transcription, which can easily be achieved by constructing an envelope around the transcription site. To generalize : "The compartment of function must be separated from the compartment of formation".

One of the implications of this idea is that evolution would not favour introns in one compartment systems. "Absence of compartmentation mean absence of intron." This has indeed been observed : no prokaryotes, no chloroplast and no mitochondria have introns in their genes. (The example of mitochondrial cytochrome C introns, does not violate this rule as discussed in section 4.5). There are

some RNAs, which do not code for any protein or cytoplasmic RNA, but instead work inside the nucleus only. This class of RNA which never have to go out of the nucleus are called small nuclear RNA (Sn RNA; involved in splicing as indicated in Section 2.3.2). According to the principle stated above their genes should not possess introns. Fortunately, a gene for one of the Sn RNA, the UI RNA have been characterized and have been shown to harbor no introns (83).

Once the splicing machinary is established and the envelope is formed, presence of introns would be consolidated. Not only the already split genes would remain split, splitting of the new genes would also be favoured (by the processess described in section 3.7 and 3.8), because the mosaic organization of genomes confer several evolutionary advantages (discussed in the section 2.2.2). Presence of nuclear envelope around the genetic material stipulates conditions, very different from the prokaryotic organization. For example, horizontal transfer of information cannot take place,

regulation of gene expression is altered, mode of the mRNA utilization is different and monocistronic transcriptional units are necessary.

In prokaryotes a cell can receive and utilize, DNA from cells of the same species or a relatedspecies, in its life time, by the processes of transduction, conjugation and transformation. This process, called horizontal transfer of information (27), is possible here because DNA has to cross only one barrier that is the cell membrane to reach to the genome. The horizontal transfer is not utilized naturally, in eukaryotes because of the second barrier, offered by the nuclear envelope and the hostile cytoplasm. This blockage to a very important source of variability is compensated in eukaryotes by mechanisms like <u>en bloc</u> recombination assisted by introns as discussed earlier.

Existence of the nuclear envelope necessiate the use of mRNA in a different way. The prokaryotic mRNA is highly unstable; sometimes, even before the completion of transcription it starts disintegrating. This liability is essential because it makes the transcriptional regulation very effective. If

the half-life of mRNA is very small, the rate of protein synthesis can be changed by merely altering the rate of transcription (84). However, such an unstable mRNA cannot work in eukaryotes primarily because, the transcription and translation is uncoupled, by the nuclear envelope. The eukaryotic mRNAs are highly stable (because of the cap and the poly-A tail), and hence the regulation at translation becomes significant. Besides this, the eukaryotic condition provides two additional levels of regulation - the splicing and transport of mRNA across the nuclear membrane.

The monocistronic nature of the messages in eukaryotes may also be a consequence of nuclear envelope. In the prokaryotes where polycistronic messeges are found, the ribosomes draggle behind the polymerase and the ribosome nearest to the polymerase molecule is almost attached to it. Between each gene in a polycistronic mRNA, are weak transcription termination signals. These signals are skipped if the ribosomes keep following the polymerase. If, because of some reason ribosomes are unable to follow the polymerase, the transcription stops at the first termination signal encountered

(85). The reason for this is not known, but the fact that 3 out of 4 polypeptides of 'QB RNA-phage replicase' belong to host translation system (EFTu, Ts and S1) and one of the activators of normal RNA polymerase the $\psi$ factor is EF-Tu of the translation system, indicate a tight coupling between translation and transcription. From these observations, if we assume that the coupling between the ribosomes and RNA polymerase is essential, for an unhindered transcription of polycistronic messengers, the nuclear envelope would not allow the existence of polycistronic operons in eukaryotes. The envelope intersects the transcription from the translation.

## 4.4 The Extranuclear Compartments

At the time of transition from hypercyclic, progenate state to cellular organization, different fates awaited the membrane systems of the two major lineages. Prokaryotes developed a wall outside to their cell membrane while membranes of urkaryotes gained the properties of exocytosis/endocytosis and extensive invagination. These are the properties which can be instrumental in the development of the nuclear envelope (and mitochondria/chloroplast as

well)(87). The nuclear envelope which is a double membrane, has been considered as, coalesence of invaginating cell membrane. These features of the membrane would allow the urkaryotes to develop the mechanism of engulfing particles of supramolecular sizes. The internal membrane system would enable them to increase the total volume of the cell, while the prokaryotic cells would be captivated in the boundaries set by their cell wall (Prokaryotes would have to live upon only the diffusable food).

The self-replicating systems of the progenote state, presumably depended upon the supply of energy rich monomers synthesized abiotically in the environment by the action of UV etc. Ultimately, this had to be replaced by endogenous production, by using the solar energy or by oxidising the reduced materials obtained from pinocytosis or diffusion assimilation. This would be accomplished by the development of the electron transport chains, the vectorial transport of proton across the cell membrane and the ATP synthesis apparatus.

Contemporary cells produce their ATP by two processes, the substrate level phosphorylation and

the electron transport system, the later constitute the major portion of cell's energy production. The mechanism of ATP synthesis through electron transport system is explained by widely accepted chemiosomotic theory (88, 89). According to this theory, when electrons flow, down the electron transport chain, located in the cell membrane of bacteria or inner membrane of mitochondria, the protons are translocated across the membrane from the inside to the outside medium. This vectorial translocation of hydrogen ions leads to establishment of an electrochemical gradient across the membrane. The electrochemical gradient, then is used to drive the ATP synthesis (88, 89).

As the size of cells increased in the urkaryotes, the number of electron transport chains would become insufficient to establish gradiants strong enough to drive the ATP synthesis. The number of electron transport chains can increase only to a limited extent, the limit being set by the surface area of the membrane. The volume of a cell would increase with the third power of its radius, while the surface area, with only the second power. Therefore,

it will not be possible to accommodate sufficient number of electron transport chains in the cell membrane (90).

A solution of this problem is to make separate compartments for energy transducing systems. As is clear from the requirement of proton gradient for ATP synthesis, the compartment must be a close one. The nature of the functions these compartments have to perform, is such that they should contain their own DNA and protein synthetic machinary inside themselves. This requirement is necessary because some proteins must be inserted into the membrane only from inside, to achieve the required asymmetry of membrane for vectorial functions. The evolution of the chloroplast and mitochondria was an outcome of this requirement. There are two contrasting views about the origin of the two organells - the autogenous and the symbiotic hypothesis. According to the autogenous evolution theory (27, 87, 91, 92), invagination of cell membrane around a plasmid or some other DNA element, formed the first mitochondria. But the endosymbiotic theory says that, the prokaryotic cell which had acquired the ability of pinocytosis,

engulfed some respiring bacteria, which instead of becoming the food of the cell became a symbiont (79, 27, 75). This idea owes its origin to the immense amount of similarities between the bacteria and mitochondria/chloroplast: similar electron transport system, size and organization of genome, and protein synthetic machinary, etc. When the chloroplast rRNA sequences were compared with rest of the organisms, they fell into a group which is otherwise made up of prokaryotes only (77). Similar analysis has put mitochondrial rRNA also into the group of bacteria (93). These observations point towards the hypothesis that at least the genetic material of chloroplast and mitochondria is derived from the bacterial pool. This, however, does not constitute an evidence in favour of the orthodox endoymbiotic theory, since the membrane system, still can very well be supplied by the recipient.

Although neither of the hypotheses can be accepted conclusively at this stage, few trends can be conceived about the early evolution of mitochondria and chloroplast. Genomes of the symbiotic bacteria or of autogenously developed protomitochondria would lose genes whose functions can be allowed to express

by the nuclear genome. For example, the genes whose products can be inserted from outside into the membrane. This trend would be favoured because it reduces the autonomy of the organelle and allow the host to have complete control over the duplication of mitochondria/chloroplast. Besides this, some changes would be necessary in the biosynthetic and metabolic machinary of the organelle, to avoid muddling of cytoplasmic functions with organelle's functions. There would be serious problems, if cytoplasmic mRNAs can get expressed in mitochondria or mitochondrial ones in the cytoplasm. Probably, this is what led the mitochondria to develop a modified genetic code (94).

## 4.5 Introns of the Maveric Mitochondria

According to the 'thesis of compartments' no introns should be found in bacteria or mitochondria. That has indeed generally been observed but several genes in lower eukaryotes have been found to contain several introns. Apocytochrome gene of yeast (95, 96) and of Aspergillus (97) contain five to seven and one intron(s) respectively. These introns, considering their splicing mechanism, do not violate the

basic premises of the hypothesis proposed in the earlier sections. Organization of those genes is such that final product can be made only after the removal of the introns. Removal of the introns employs a machinary radically different from the nuclear splicing. The preceeding intron makes a protein in collaboration of an exon, the maturase, which removes all the introns from the RNA transcript. The final translation can take place only in the processed mRNA. Translation is separated from transcription not through a spatial membrane but by a functional barrier. Moreover, these introns are not obligatory for the survival of the mitochondria, mutants are known where all the introns are missing and apocytochrome protein is made normally (96).

## 4.6 An Intron is not for Ever

Most of the introns are as stable as exons in evolution, but occasionally some of them are found leaving their genes. The mechanisms which describe these deletions does not challenge the general mechanism proposed for the origin of introns.

One of the $\alpha$-like globin pseudogene has lost all its introns (98). The deletion is so

accurate that not a single base is deleted from or added into the reading frame. Similar phenomenon has been observed for a pseudogene of human immunoglobulin (99). From the organization and sequence of these pseudogenes, it appears that they are reverse transcription products of a fully processed mRNA, which have been conveyed back to the chromosome by some retrovirus like element (100). There are few observations suggestive of this : pseudogene has been translocated to a chromosome different from its family location; it contains repeat sequences like the retrovirus terminals; a stretch of poly-A is found at the end of the gene (99, 100). It would be quite an aesthetically appealing hypothesis that a retrovirus kidnapped the mRNA and inserted its DNA copy at a random site.

## Key to Figures

| | |
|---|---|
| AAAA | Polyadenylation Signals |
| C | Constant region of antibody molecules |
| E | Exon |
| G | Gene |
| I | Intron (Except figure 7,8) |
| → R | Repeat Sequence Inside the Introns |
| 3'SS, 5'SS | Splicer Sequences (see Section 2.3.2) |
| V | Variable Region of Antibody Molecules. |

## REFERENCES


1. Stainer, R.Y., Adelberge, E.A. and Ingrahm, J.L. (1976), The Microbial World, Prentice Hall, Englewood Cliffs, N.J.

2. Crick, F.H.C. (1979), Science, 204, 264.

3. Darnell, J.E. (1979), Science, 202, 1257.

4. Lewin, B. (1976), Gene Expression-2, John Wiley.

5. Breathnach, R. and Chambon, P. (1981), Ann. Rev. Biochem., 50, 349.

6. Sun, S.M. et al. (1981), Nature, 289, 37.

7. Wild, M.A. and Sammer, R. (1980), Nature, 283, 693.

8. Ogden, R.C. et al. (1981), Trend. in Biochem. Science, 6, 154.

9. Bleak, G.C.F. (1978), Nature, 273, 267.

10. Fyrberg, E.A. (1981), Cell, 24, 107.

11. Ziff, E.B. (1980), Nature, 287, 491.

12. Kimura, M. (1979), Sci. Ame., 241 (5), 94.

13. King, J.L. and Jukes, T.H. (1969), Science, 164, 788.

14. Miyata, T. et al. (1980), Proc. Nat. Acad. Sci., 77, 7328.

15. Doolite, W.F. and Supienza, C. (1980), Nature, 284, 604.

16. Orgel, I.E. and Crick, F. (1980), Nature, 284, 604.

17. Gilbert, W. (1978), Nature, 271, 501.

18. Wilson, C.M. (1981), Nature, 290, 113.

19. Robertson, M. and Hobart, M. (1981), Nature, 290, 543.

20. Molgard, H. (1980), Nature, 286, 657.

21. Maki, R. et al. (1981), Cell, 24, 353.

22. Alt, W.F. et al. (1981), Cell, 20, 293.

23. Rogers, J. et al. (1980), Cell, 20, 303.

24. Early, P. et al. (1980), Cell, 20, 313.

25. Stein, J.P., et al. (1980), Cell, 21, 681.

26. Hamer, H.D. and Leder, P. (1980), Cell, 18, 1299.

27. Broda, E. (1978), Evolution of Bioenorgetic Processes, Pergamon Press.

28. Keeds, L.H. (1979), Ann. Rev. Biochem., 48, 838.

29. Sakano, H. et al. (1979), Nature, 277, 627.

30. Barker, W.C., et al. (1980), J.Mol. Evol., 15, 113.

31. Blake, C.C.F. (1981), Nature, 291, 616.

32. Mitco, Go (1981), Nature, 291, 90.

33. Jensen, E.O. et al. (1981), Nature, 291, 677.

34. Eaton, W.A. (1981), Nature, 284, 183.

35. Craik (1980), Proc. Nat. Acad. Sci., 77, 1384.

36. Benyajati, C. et al. (1981), Proc. Nat. Acad. Sci., 78, 2717.

37. Ohata, T. (1981), Evolution and variation of Multigene Families, Springer Verlag.

38. DeRobertis, E.M. et al.,(1981), Cell, 23, 89.

39. Peebles, C.L. et al. (1979), Cell, 18, 87.

40. Knapp, G. et al. (1979), Cell, 18, 37.

41. Hopper, A.K. and Schiltz, L.D. (1980), Cell, 19, 741.

42. Colby, D. et al. (1981), Proc. Nat. Acad. Sci., 78, 415.

43. Montel, C. et al. (1982), Nature, 295, 380.

44. Johnson, J.D. et al. (1980), Proc. Nat. Acad. Sci., 77, 2564.

45. Standbring, D.N. et al. (1981), Proc. Nat. Acad. Sci., 78, 5963.

46. Chambon et al. (1978), Proc. Nat. Acad. Sci. 75, 4853.

47. Lewin, B. (1981), Cell, 22, 324.

48. Lemer et al. (1980), Proc. Natr Acad. Sci., 283, 220.

49. Rogers et al. (1980), Proc. Nat. Acad. Sci., 77, 1877.

50. Sharp, P.A. (1981), Cell, 23, 643.

51. Murry, V. and Holliday, R. (1979), FEBS Letts., 106, 5.

52. Branlant, C. et al. (1980), Nucleic Acid Rec., 8, 4143.

53. Wickens, M.P. et al. (1980), Nature, 285, 628.

54. Chung and Sharp, P.A. (1981), Nature, 289, 378.

55. Avvedimento, V.E. et al. (1980), Cell, 21, 689.

56. Doolite, W.F. (1978), Nature, 272, 581.

57. Doolite, W.F. (1980), Trend in Biochem. Sci. 146.

58. Eigen, M. and Schuster, P. (1979), The Hypercycle, Springer Verlag, New York.

59. Woese, C. (1981), Sci. Amer., 244 (6), 98.

60. Schuster, P. (1981), in Biochemical Evolution, Ed. by Gutfreundh, Cambridge University Press.

61. Eigen, M. (1971), Naturwissenschdfen, 58.

62. Kuppers, B. (1979), Naturwissenschften, 66, 228.

63. Smith, M.J. (1979), Nature, 280, 445.

64. Eigen, M. at al. (1981), Sci. Ame., 244 (4), 88.

65. Eigen, M. et al. (1980), J. Theor. Biol., 85, 407.

66. Tyagi, S. (1981), Origins of Life, 11, 343.

67. Kuhn, H., and Waser, J. (1981), Angew. Chem. Int. Ed. Engl., 20, 500.

68. Jeffrey, A.J. and Harris, S. (1982), Nature, 296, 9.

69. Wozney, J. et al. (1981), Nature, 294, 129.

70. Eiferman, A. et al. (1981), Nature, 294, 713.

71. Cochet, M. et al. (1979), Nature, 282, 567.

72. Campbell, A. (1980), Cold Spring Symp. Quant. Biol., 45, 1.

73. Shapiro, J.A. (1977), Trends in Biochem. Sci., 176.

74. Margulis, L. (1970), Origin of Eukaryotic Cells, New Haven and London University Press.

75. Doolite, W.F. (1980), Trends in Biochem. Sci. 196.

76. Wilkinsons, (1978), Nature, 271, 707.

77. Woese, C.R. and Fox, (1977), Proc. Nat. Acad. Sci., 74, 5088.

78. Woese, C.R. (1977), J. Mol. Evol., 10, 1.

79. Woese, C.R. and Gupta, R. (1981), Nature, 289, 95.

80. Schopf, J.W. (1978), Sci. Amer., 239 (3), 84.

81. Van Volen, M.L. and Maiorana, V.C. (1980), Nature, 287, 248.

82. Woese, C.R. et al. (1970), J. Mol. Evol., 11, 245.

83. Rooy, D.R. et al. (1981), Cell, 23, 671.

84. Volkenstein, M.V. (1977), Molecular Biophysics, Academic Press.

85. Shanker, A. Personal communication.

86. Travers, B.A. et al. (1970), Nature, 228, 748.

87. Cavelier Smith,(1976), Nature, 256, 465.

88. Mitchell, P. (1979), Science, 206, 1148.

89. Hinkle, P.C. (1978), Sci. Ame., 238, 104.

90. Hall, J.B. (1973), J. Theor. Biol., 38, 413.

91. Ralf and Mokuter, (1972), Science, 177, 575.

92. Unzzel and Spolskey (1974), Amer. Sci. 62, 334.

93. Kuntzel, H. Hoctiel, M.G. (1981), Nature, 293, 751.

94. Lagerkuist, U. (1981), Cell, 23, 305.

95. Borst, P. and Grivel, L.A. (1981), Nature, 289, 439.

96. Lazowaska, J. et al. (1980), Cell, 22, 333.

97. Waring, R.B. et al. (1981), Cell, 27, 4.

98. Vanin, E.F. et al. (1980), Nature, 286, 222.

99. Hollrs, G.F. et al. (1982), Nature, 296, 321.

100. Plavel, R.A. (1982), Nature, 295, 370.

Thanks,

Most of the ideas presented in this dissertation have sprouted from the discussions with Dr, J. Subba Rao, Prof Hiren Das, Shreeniwas Kumar, Meetha Medhora, Nishikant Singh, Indranil Dasgupta, Suhas Pardnis, and C.P. Geevan. The financial assistance from R.D. Birla smarak Kosh is acknowledged. I am greatful to Gajanan for carefully typing the manuscript.

Sanjay Iyengar