SUMMARY

An extensive study on gene sequences and protein structures belonging to

- the trypsin family of serine proteases
- the EF-hand calcium binding proteins
- actins
- immunoglobulin superfamily
- dehydrogenases and catalase

were undertaken with a view to understand their split gene organisation. It was of particular interest to study the possible origin of intervening sequences and the important roles they could play in evolution of these gene families.

- Studies on dehydrogenases and catalase revealed that introns correspond closely with the borders of major structure/function domains and that their positions have been maintained between species. This fact, along with the observed correlation between intron positions and the secondary structural elements suggests that introns could not have been randomly inserted into their genes.

- Analysis of the trypsin family of serine proteases has shown that the functional domains bearing the conserved catalytic residues exhibit a great deal of similarity in the protein structure and could be regarded as folded motifs (It is noteworthy that they do not exhibit the features of domains as observed for dehydrogenases). The fact that introns are found to border the functional domains bearing the catalytic residues
validates Gilbert's idea that exons code for meaningful units of protein structure. Occurrence of introns in positions corresponding to functionally important residues suggests the functional significance of introns.

It has been established that the genes coding for collagens and immunoglobulins evolved by gene duplication event and that introns have been found to separate the repeat units observed for the duplicated segments/domains. An analysis of the gene sequences of class I MHC molecule and gamma crystallin reveal similar lines of evidences. In class I MHC, intron positions are highly conserved in the homologous α1 and α2 domains of the polymorphic heavy chain, which are known to have evolved by gene duplication. It is also interesting to observe a very strong conservation in intron positions in the structurally related motifs of γ-crystallin. This further strengthens the idea that gene duplication could have been an important event in their evolution.

Analysis on the EF-hand type of calcium binding proteins such as calmodulin, troponin C and parvalbumin reveal that introns do not occur at the borders of the EF-hand domains, but are predominantly found to interrupt the loop regions coding for calcium binding residues. In other words, introns are found to separate functionally important regions that are known to bind to calcium ions and thus validates Gilbert's idea that introns are units of protein function.
The distribution of introns in highly conserved regions of these proteins is suggestive of a common ancestral gene that could code for the helix-loop-helix calcium binding domain and also supports the idea that introns might not have been inserted randomly into a pre-existing gene. In particular, studies on the intron/exon organisation of various calmodulins have shown that the number of introns found in lower eukaryotes are equal to that of higher eukaryotes and further, there exists strong correlation at least for a few of their intron positions. This offers a strong support to the idea that introns could have been more ancestral in origin and the absence of a few intron positions could be accounted for intron loss in a bid to streamline their genome.

A compilation of gene sequence data for the ubiquitous actins in a wide range of organisms has revealed interesting results. Analysis on their gene structures from various organisms has revealed that there may be at least 25 intron positions distributed at different locations in the coding regions, and several of these intron positions are maintained strictly between the species. The fact that intron positions are conserved even between lower eukaryotes, alga and higher organisms argues strongly once again for an ancestral origin of introns.

Cumulative analysis on the features of introns such as phase, number and conservation of splice junction sequences for the various gene families considered in the present analysis support the 'introns-early' view.