What is a Gene?

2. A Question with Variable Answers

S C Lakhotia

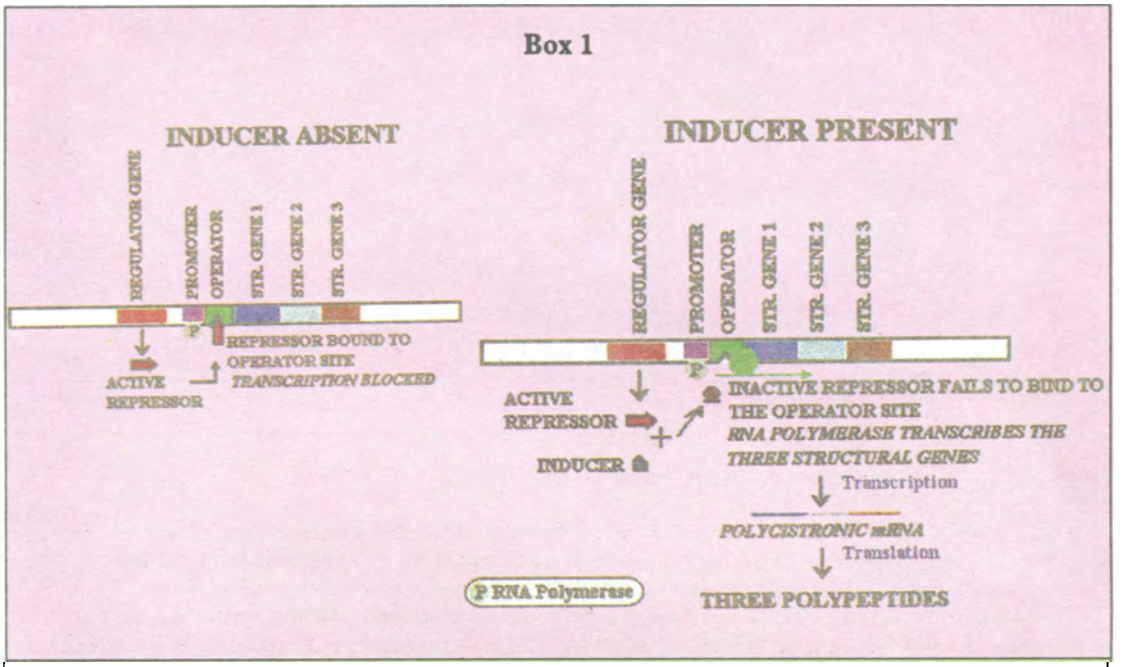
Subhash C Lakhotia teaches at the Department of Zoology, Banaras Hindu University and is particularly interested in Cytogenetics and Molecular Genetics. He has extensively used the fruit fly, *Drosophila*, to study organization of chromosomes and expression of genes in eukaryotes. His current research interests concern the heat shock or other stress induced gene activity and its biological significance. His other interests include reading, photography and listening to classical music.

The first part of this general article appeared in April 1997.

The first part of this article traced the evolution of the concept of a gene from Mendel's times to the middle of this century: starting from the imaginary *factors* of Mendel, the *genes* were shown, in the first few decades of this century, to be physical entities many of which were linked in a linear order on a single chromosome. Each of these were believed to be indivisible units of function, mutation and recombination. Subsequent studies in microorganisms as well as higher organisms revealed the *gene* to be divisible in all its properties. The *classical* concept of indivisible gene thus gave way to something with a more complex organization. The advent of molecular biology allowed more detailed studies on the organization of the gene and the way it functions. In keeping with the remarkable diversity and complexity of biological systems, the *gene* has also turned out to be equally diverse and fascinating.

Genes Control Phenotype through RNA and Protein

While the physical organization of genes was being understood, progress was also made in understanding the relationship between deoxyribonucleic acid (DNA), the chemical on one hand, and the phenotype, the biological manifestation of the genetic information, on the other. Although the involvement of the ribonucleic acid (RNA) in protein synthesis was suggested by T Caspersson in 1941 and by J Brachet in 1942, the concept of transcription or production of messenger RNA as the primary step in the function of genes was firmly established in late 50's and early 60's. Development of cellular autoradiography using special chromosomes like polytene chromosomes of fruitflies and midges (W Beermann and his group) and lampbrush



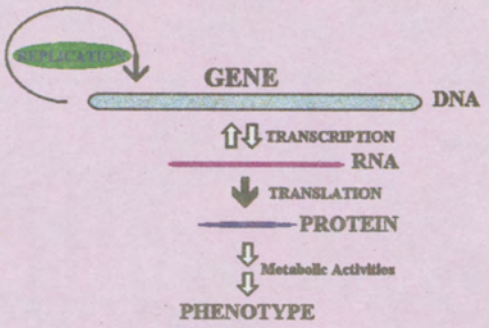
OPERON MODEL

In bacteria, genes involved in a given anabolic or catabolic pathway are organized into units called OPERONS to facilitate their coordinated expression. In the case of INDUCIBLE OPERONS, the regulator gene produces an active REPRESSOR which specifically binds to a DNA sequence named the OPERATOR of the given OPERON and thereby prevents RNA polymerase from transcribing the STRUCTURAL GENES (STR. GENES 1-3 in this example) constituting that Operon. In the presence of the INDUCER, the Repressor binds with higher affinity to the Inducer and due to allosteric changes, adopts a different conformation which prevents its binding to the Operator site. As a consequence the RNA polymerase now can bind and transcribe the Structural Genes in the Operon into a long message corresponding to the three genes (therefore called a POLY-CISTRONIC mRNA) which on translation produces the corresponding three POLYPEPTIDES.

chromosomes of salamanders (H G Callan and others) definitively established that RNA is synthesized at sites of active genes (puffs and loops). For a geneticist, the gene was still a conceptual entity, but to cytologists and cell and molecular biologists, a gene was a distinct physical entity. Early in the 60's, F Jacob and J Monod proposed the famous *operon* concept and also hypothesized the presence of short-lived messenger RNA (mRNA) as the intermediate between gene (DNA) and protein (see *Box 1*). These developments helped establish the *Central dogma* of molecular biology (see *Box 2*) and also in understanding

For a geneticist, the gene was still a conceptual entity, but to cytologists and cell and molecular biologists, gene was a distinct physical entity.

BOX 2



CENTRAL DOGMA OF MOLECULAR BIOLOGY

The genetic information is stored in the base sequence of the DNA molecule which can either generate more copies of itself through REPLICATION (for passing on exact replica of the genetic information to daughter cells or gametes) or be copied by the process of TRANSCRIPTION into RNA which acts as intermediary in the flow of genetic information from the nucleotide language of nucleic acids to the amino acid language of proteins through TRANSLATION. Replication is catalyzed by *DNA-dependent DNA polymerase* while transcription is through *DNA-dependent RNA polymerase.* The information in RNA(messenger) is used to guide the assembly of amino acids of a polypeptide during translation in a specific sequence determined by the GENETIC CODE. The amino acid sequence of a protein (polypeptide) determines its functions. The proteins control all metabolic activities and, therefore, the phenotype. The discovery of *RNA-dependent DNA polymerase* showed that the genetic information may also travel from RNA to DNA (REVERSE TRANSCRIPTION) as in the case of oncogenic Retroviruses.

the physical and chemical basis of the nature of genetic information, its transmission, expression, mutation and recombination. The most important yet very simple concept in the central dogma was that all information transfer in a cell was template dependent: *replication* provides for identical copies of the genetic information to be passed on to daughter cells; *transcription* copies the information from DNA into RNA (as part of the gene activity) and finally, using the genetic code, the nucleic acid language in RNA is *translated* into the amino acid language of proteins. The simplest modern view of a gene was thus emerging as the segment of DNA that transcribes an RNA which in turn determines the order of amino acids in the polypeptide chain which by its activity regulates a specific cell function and thus, ultimately the phenotype. In terms of

The most important yet very simple concept in the Central dogma was that all information transfer in a cell was template dependent.

mutation, it was obvious that any base in the *functional* part of the gene may change independent of the neighbouring bases or in concert with them and result in an altered (mutated) function. Likewise, recombination between homologous chromosomes could occur, at least theoretically, between any two neighbouring bases. However, hopes of finally providing a single universal definition of gene did not materialize since the types of genes and the anatomy that has been discovered during the past three decades are too varied to fit into a typical gene. This can perhaps be a little confusing for a student of genetics but is very exciting for those who wish to learn more about genes.

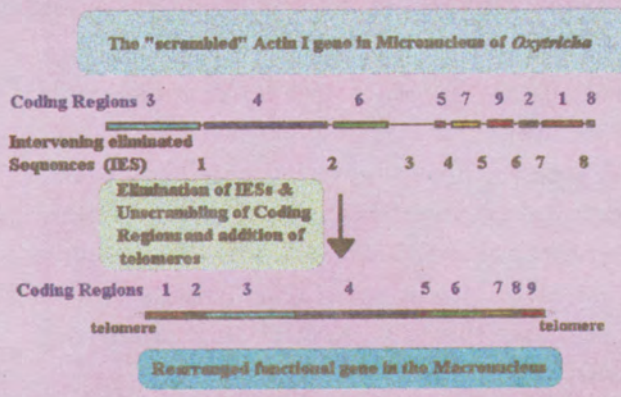
Genes Exist in Many Designs

The *operon* model itself suggested different classes of genes like the *regulatory* and *structural* genes. In addition, control elements like the operator and promoter were also shown to exist. Each of these could mutate independently and yet result in a similar end-result at the phenotypic level. These concepts led to defining the *gene* as a unit of transcription which would include not only the protein-coding region of DNA but also the flanking DNA sequences which control the activity of the transcribed region. The cracking of the genetic code and the concept of collinearity of the sequence of bases in a gene and of the amino acids in the polypeptide were exciting discoveries of the 60s and appeared, in those days, to have finally unraveled the mystery of the relation between gene and phenotype. However, as more precise methods of study became available and more genes were examined in detail, the apparently universal definition of a typical eukaryotic transcription unit or gene once again became inadequate due to the ever-increasing numbers of unconventional genes being discovered.

Gene as a unit of transcription includes not only the protein-coding region of DNA but also the control sequences flanking it on either sides.

The protein-coding *structural genes* are now known to exist in a wide variety of designs. The discovery of *introns* (regions of a gene that are transcribed but are removed or spliced out during post-transcriptional processing of the precursor of mRNA and

Box 3



Assembly of functional *Actin I* gene in the macronucleus of *Oxytricha*, a ciliate, by "unscrambling" of coding regions and elimination of the intervening non-coding regions. Ciliates have two kinds of nuclei, the micronucleus and the macronucleus, the former is transcriptionally inactive and is responsible only for reproduction while the latter is transcriptionally active "vegetative" nucleus that carries out all the "somatic" functions. The macronucleus carries "gene"-sized DNA molecules that are derived from the genomic DNA of the micronucleus by a variety of DNA processing events. The coding regions for the actin protein are scrambled in *Actin I* gene in micronucleus of *Oxytricha* with the different coding segments being demarcated by short "Intervening Eliminated Sequences" (IESs). This gene is non-functional till "rearranged" and "unscrambled" during macronucleus development. As shown above, this involves selective removal of the IESs, rejoining of the different coding regions in correct order and addition of telomeric sequences at both the ends. Following such rearrangements, all the gene-sized fragments in macronucleus are amplified.

are thus not translated) appeared to demolish the concept of collinearity of gene and protein. Besides the existence of protein-coding genes that carry introns (referred to as *split genes*) and those that do not, a variety of other types have since been discovered. These include *repeated genes* (present in multiple copies in the haploid genome either at a single location, as in the case of genes that code for histone proteins or at dispersed sites, as in the case of certain repetitive DNA sequences that are transcribed), *assembled genes* (DNA segments that are initially inherited as non-contiguous separate entities but are assembled into contiguous sequences sometime during development by splicing of different pieces and/or through recombination at DNA level, see *Box 3*), *overlapping genes* (a contiguous segment of DNA produces different transcripts corresponding to different regions either on the same strand or from opposing strands; for

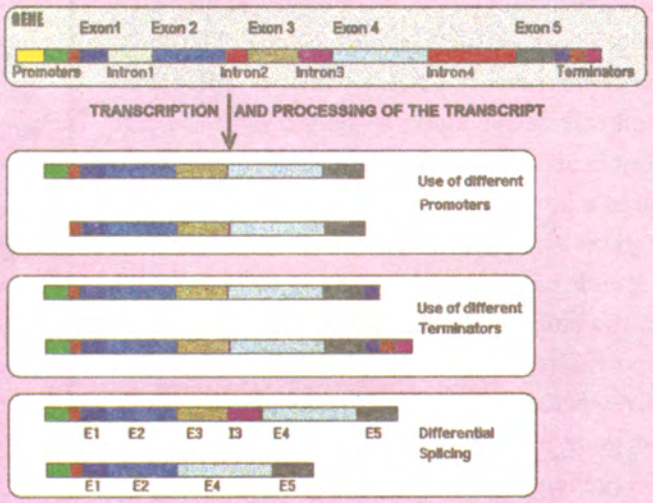
example both DNA strands of the *c-myc* gene of mouse are transcribed in overlapping fashion),*transposable or jumping genes* (DNA sequences that have the propensity to be *mobile* or to *jump* from one location to another like the transposable elements discovered by Barbara McClintock, see *Resonance*, October, 1996), *polyprotein genes* (same RNA produces different polypeptides through translation along its length as in the case of the gene for many neuroendocrine peptides that result from proteolytic cleavage of a large precursor protein or of genes for some ribosomal proteins that also carry poly-ubiquitin coding sequences on their 3' end),*nested genes* (an entirely different gene is embedded within the boundaries of another gene either in its intron or on the opposite strand; for example the large intron of the *dunce* gene of *Drosophila* harbours at least two other genes including the *Sgs-4* gene coding for salivary gland secretion protein), *pseudogenes* (genes that share identical coding region with a functional gene but themselves are not capable of transcription due to absence of the regulatory elements and/or introns) and so on.

In terms of processing of the primary transcripts, the same gene may produce different products in the same or different cells due to alternative splicing of the intron containing transcripts or due to initiation and/or termination of transcription at variable sites (see *Box 4*). A remarkable example of variable rearrangement of the genetic sequence to generate new genes is the case of immunoglobulin genes that produce the endless variety of antibodies in our body: in these cases the precursor gene is modular with several components, each present in multiple but variable copies. The functional immunoglobulin gene is produced by a series of splicing events at the DNA level such that one member of each module is randomly joined with others so that each of the mature lymphocytes in our blood-stream produces a unique antibody. As a result, the DNA base sequence of the immunoglobulin gene is different not only between cells that produce antibodies and those that do not but also amongst the different antibody producing lymphocytes themselves.

The DNA base sequence of the immunoglobulin genes is different not only between cells that produce antibodies and those that do not but also amongst the different antibody producing lymphocytes themselves.

Box 4

Same gene may produce different transcripts through the use of different promoters or different terminators or through differential splicing



Production of different transcripts, and therefore, of different polypeptide products, by alternative splicing and/or the use of alternative PROMOTERS and TERMINATORS. A gene may have several different promoters in the upstream region and several terminators (motifs that instruct the RNA polymerase to terminate transcription at defined sites). The transcription unit iself may include a number of EXONS (E1 to E5 in the above example) and introns (I1 to I4 in the above example). In many cases, the process of SPLICING (removal of introns to generate the mature mRNA) is regulated so that the same primary transcript may give rise to different species of mRNAs by regulating the removal or non-removal of some introns and/or exons. Alternatively the use of alternative promoters and/or terminators produces transcripts that differ in their functional properties. These have very significant functional consequences. For example, a key decision by a *Drosophila* egg to develop as male or as female depends upon the pattern of splicing of the transcripts of a gene called *transformer*.

The central theme of template dependence in transcription is violated by the discovery of a process called editing of RNA.

The central theme of template dependence in transcription is violated by the increasingly known instances of a process called *Editing of RNA*. This process results in the mature mRNA having a different base sequence from what was initially transcribed by the DNA template. Consequently, editing of RNA generates new information that was not present in the gene at the DNA level. The upstream (5' to the transcription unit) or downstream (3' to the transcription unit) regulatory elements, although generally not considered as part of the gene because they are not transcribed, are also very important in our current

BOX 5

LACK OF CORRELATION BETWEEN GENE SIZE (in kilobases) AND PROTEIN SIZE (no. of aminoacids) - SOME HUMAN GENE EXAMPLES

GENE NAME	SIZE (kb)	mRNA SIZE (kb)	No. of INTRONS	PROTEIN SIZE
Histone H4	0.6	0.3	0	104aa
b-globin	1.5	0.6	2	147aa
Insulin	1.7	0.4	2	52aa
Protein Kinase C	11	1.4	7	671aa
Albumin	25	2.1	14	585aa
Catalase	34	1.6	12	527aa
Factor VIII	186	9	25	625aa
Thyroglobulin	300	8.7	36	2767aa
Dystrophin	>2000	17	>50	3685aa

(The above gene sizes include the coding regions and some flanking regulatory sequences)

Relation between the size of gene and its protein product is variable: While there is no intron in the *Histone H4* gene, the gene for *Dystrophin* (a mutation in this gene is responsible for Muscular Dystrophy) is one of the largest known with a very large number of introns.

concepts of genes due to their vital role and their complex organization. Thus in terms of mutational characterization of the gene, a mutation with a profound effect on the phenotype may be in the upstream or downstream regulatory elements, or in the coding region or in the non-coding 5' or 3' untranslated regions (UTR) or in introns. Likewise, from the point of view of recombination, while base pairs remain the theoretical unit of recombination, the distance between adjacent genes may vary from zero to hundreds or thousands of kilobases. The size of the gene itself may vary from a few hundred base pairs to several hundred kilobase pairs (see *Box 5*)! Likewise, sizes of the upstream or downstream regulatory regions also vary considerably.

All Genes Do Not Code for Proteins

Not all genes code for proteins. The ribosomal and transfer RNAs (rRNA and tRNA) are transcriptional products of

The non-protein coding genes function through their transcripts in an as yet unknown fashion.

corresponding genes which, although essential for translation, are not translated into proteins. Another set of genes produces transcripts (the snRNAs or the small nuclear RNAs) that are also not translated but are essential for the processing (splicing) of the intron containing precursor RNAs. It was also found that eukaryotes have a number of different DNA-dependent RNA polymerases that function to transcribe different classes of genes. Thus while the RNA polymerase I transcribes ribosomal RNA, the RNA polymerase II transcribes the protein-coding genes and the RNA polymerase III transcribes the tRNAs, the snRNAs etc. An additional complication in the concept of gene is the increasing realization that many genes are transcribed like the typical *protein-coding genes* by RNA polymerase II and produce transcripts of varying sizes, which are processed in much the same way as the protein-coding mRNAs, but are ultimately not translated or are untranslatable. These *non-protein coding genes* function through their transcripts in an as yet unknown fashion.

Yet another intriguing feature is the lack of any correlation between the genome size and the biological complexity of different organisms. For example, the salamanders and frogs have much more DNA per haploid genome than humans, although we believe that humans are biologically more complex and thus may require a greater genetic information content. Even in those multicellular organisms that have a very small genome size, only a small fraction (5% to 30% on various estimates) of the genome is known to actually transcribe. These two features are collectively referred to as the *C-value paradox* (C-value being the total amount of DNA in a haploid genome as in gametes). Two questions thus arise: why should related organisms differ in their genomic DNA content and why should the genome carry so much of DNA which apparently is not transcribed? Failure to have satisfactory answers to either of these questions have prompted some scientists to think in terms of *selfish, parasitic* and/or *junk* DNA etc while others believe that we are still far from understanding the varied ways of functioning of DNA. It is possible that, besides the protein-

Salamanders and frogs have much more DNA per haploid genome than humans, although we believe that man is biologically much more complex.

coding language, the DNA may have yet another language, for example in relation to the higher order organization of chromatin in a eukaryotic nucleus. The chromatin organization may define certain *supra-genic* aspects of biological information that we inherit through the generations and has a profound role in gene activity.

Our understanding of the structure and organization of the genetic material has greatly increased in recent years: we have deciphered the total DNA base sequence of genomes of some organisms and we are witnessing a gene revolution due to marvels of genetic engineering. However, in spite of these remarkable achievements, our understanding of *gene* continues to be incomplete. This incomplete understanding of the varied nature of gene continues to challenge and excite biologists. To be able to understand and appreciate the wonderful and most efficient machines that the living organisms are, we will need to comprehensively understand what genes are! This comprehensive understanding may not be very far off in view of the rapid progress in this field.

The chromatin organization may define certain *supra-genic* aspects of biological information that we inherit through the generations and has a profound role in gene activity.

Suggested Reading

- **Peter Portin. The concept of the gene: short history and present status. *Quarterly Review of Biology*, Vol. 68.pp. 173-223, 1993.**
- **Benjamin Lewin. *Genes V.* Oxford Univ. Press. Oxford. New York. Tokyo, 1994.**

Address for Correspondence
S C Lakhotia
Cytogenetics Laboratory
Department of Zoology
Banaras Hindu University
Varanasi 221 005, India

Euler could repeat the Aeneid from the beginning to the end, and he could even tell the first and last lines in every page of the edition which he used. In one of his works there is a learned memoir on a question in mechanics, of which, as he himself informs us, a verse of Aeneid gave him the first idea.

David Brewster