

## POINT OF VIEW

### A SYSTEMIC CONCEPT OF THE GENE\*

Maria Inês de Moura Campos Pardini and Romeu Cardoso Guimarães

#### ABSTRACT

The univocal correspondence between one gene and one polypeptide has been challenged by many examples of ambiguities. A rapidly expanding list of one-to-many or many-to-one correspondences includes: genomic rearrangements, alternative processing of transcripts, overlapping translation frames, RNA editing, alternative translation modes, and polyprotein cleavage.

The genomic message requires interpretation through decoding by a sophisticated information retrieval system which should also carry some kind of information. The full meaning of the whole cell, as a unit, is emphasized.

The gene is a combination of (one or more) nucleic acid (DNA or RNA) sequences, *defined by the system* (the whole cell, interacting with the environment, or the environment alone, in subcellular or pre-cellular systems), that gives origin to a product (RNA or polypeptide).

#### INTRODUCTION

The gene is the term for the units of inheritance, responsible for the mendelian characters. It was introduced by Johannsen (1909, in Mayr, 1982) but since then it has been defined in many ways (Kitcher, 1982; Falk, 1986).

---

\* This article is derived from the Master's Dissertation of the first author, presented to the Instituto de Biociências de Botucatu in 1989.

After the identification of its location in the chromosomes and in the DNA molecules of cells, fine genetic analysis introduced the term cistron (Benzer, 1955). Bacterial molecular genetics further clarified the definition, with the demonstration of the colinearity of genetic maps and protein sequences (Yanofsky, 1967). A simple situation was configured by the notion of correspondence between one DNA sequence and one polypeptide.

Advances in the direct study of nucleic acids, especially in eukaryotes, produced an increasingly long list of exceptions to the univocal correspondence rule. It is obvious now that these "exceptions" are so many that the rule itself needs revision. In this paper, we will review some of the most striking challenges to the prevailing concept and delineate an updated concept. It is expected that the new approach to the definition of the gene will be of both didactic and practical value.

## AMBIGUOUS DNA SEQUENCES AND GENES

Examples are abundant of situations where more than one DNA sequence codes for one RNA or polypeptide and where one DNA sequence codes for more than one RNA or polypeptide, or combinations of these. In all cases, the ambiguity has been related to the expression of function derived from the genomic information. In the situations where one DNA sequence produces multiple functions, it becomes clear that the biological information is very densely packed in some segments of the inherited genomes.

### 1. Genomic rearrangements

Antibodies produced by B lymphocytes and T-cell receptors are coded for by DNA sequences put together, during ontogeny, from segments which were far apart in germ cells. The joining mechanism, mediated by the *D* and *J* segments, seems to be the same for all kinds of *V* and *C* regions. Therefore, a *V* region may become part of different immunoglobulins or T-cell receptors, diverse as to the *C* regions, and *vice-versa* (Leder *et al.*, 1974).

Some unicellular organisms change the type of protein produced through translocations of DNA segments. Well known cases are the switches in African trypanosome variant surface antigens (De Lange *et al.*, 1983), and in yeast mating types (Strathern and Herskowitz, 1979; Nasmyth *et al.*, 1981).

### 2. Alternative processing of transcripts

Diverse kinds of mRNA can be generated through differential processing, usually alternative splicing routes, of transcripts from one same DNA segment. There are cases

where the products are partial homologues, of the same protein type, as the membrane-bound *versus* secreted immunoglobulins (Early *et al.*, 1980) or the different tissue-specific isoproteins or isoenzymes (Young *et al.*, 1981). This process occurs in different protein types. Instances are known of choices associated to both the transcription initiation (Benyajati *et al.*, 1983) and termination (Maki *et al.*, 1981) steps. More drastic departures from homology are the cases where segments from transcripts become exons in one processing pathway and introns in another (Crabtree and Kant, 1982; Nawa *et al.*, 1984).

### 3. Overlapping translation frames

These cases were the earliest discovered "exceptions" to the univocal correspondence rule, detected in viruses (Zigg, 1980) and bacteria (Normark *et al.*, 1983). One DNA sequence may code for up to three entirely non-homologous products when its transcription or translation slips from one nucleotide to another, so that the translation acquires a changed triplet reading frame (Shaw *et al.*, 1978).

### 4. RNA editing

A transcript may be extensively changed in its primary structure by the action of products of other genes which insert new nucleotides or modify the preexisting ones (Simpson and Shaw, 1989). The polypeptide produced from such mRNA differs from any that could be derived from the genomic sequences, and a cDNA from such edited mRNA may not find homologies in the genome. This surprising mechanism was, again, discovered in trypanosomatid protozoans but is not restricted to them (Powell *et al.*, 1987; Chen *et al.*, 1987; Thomas *et al.*, 1988).

### 5. Alternative translation modes

Other regulatory mechanisms may occur at the step of choosing the initiator or terminator codons in one mRNA (Herman, 1989). If the alternative initiations are in the same frame, partially homologous products are formed. If the reading frames are different, non-homologous polypeptides are produced. This mechanism may be a second way of decoding the "overlapping genes", described above, which need not be defined only by the choice of transcription initiation. Slipping over an internal terminator codon may lengthen the size of a polypeptide or correct for nonsense mutation (Atkins *et al.*, 1972, 1979).

Earlier proposals that gene identity could be univocal and unambiguous at the nucleic acid level (Guimarães, 1986), required that the fully processed and mature RNA would encompass only one function. The processes described in this paragraph, and below, invalidate this argument.

### 6. Polyprotein cleavage

While in bacterial operons one single transcription initiation produces one polycistronic mRNA and its sequential translation produces the different polypeptides (Jacob and Monod, 1961), cases of tandemly joined polypeptides are known in eukaryotes, where one translation product is split into its multiple components through proteolytic cleavage.

The first of these examples is proopiomelanocortin; various tissue-specific products are encoded in one single exon (Douglass *et al.*, 1984).

Beyond these immediate post-translational events we leave the realm of classic Molecular Biology and enter that of General Biochemistry. Post-translational modifications of proteins are multitude in the complex network of metabolic interactions that will produce the mendelian characters. The ambiguities in the one gene - one phenotype rule, at this "metabolic pathways" level, have already been analysed by Hull (1974).

## THE SYSTEMIC CONCEPT

With the already very long list of evidences that the correspondence between DNA sequences and products is equivocal and ambiguous, a revision of the concept of the gene is required. The relationship between encoded information and the product of its decoding is complex, varying with the spatial and temporal conditions of occurrence.

The concept of the gene should clearly incorporate these dynamics. It is not adequate to envisage the genetic messages as entities true to themselves, just waiting for activation signals to be expressed.

The genome contains information, as its main depositary, but each function or expression has to be interpreted from it.

In present day living beings (cells), the whole system works in the retrieval of the pieces of information necessary to build a product, at each moment and site. The stored information, in genomes, requires some kind of sophisticated reading to become meaningful. An analogy might be revealing, despite some exaggeration, between the genome and the disordered "pile of texts" library model. If only the texts were adequately labelled, an efficient information retrieval system would be able to produce ordered outputs.

It is the integrated working cell, interacting with the environmental signals, that has to be rescued to its full meaning as the information retrieval system. Cells, not just genes, are transmitted through generations. The complexity of living systems might, at present, not be accessible to full description, but the perplexity put forth by the acceptance of its uniqueness should not lead to a disregard of its contribution to understanding genetics.

Cellular systems have evolved to a point where most of the process of decoding one particular DNA sequence involves other genes, at the many "regulatory" steps. Nevertheless, some degree of information resides in other components of the system, including even the external environment. In subcellular and further back to pre-cellular conditions, the role of the environment has a greater contribution. Environmental "information" would also be more important in ontogenetic processes, as opposed to its lower role in sexual reproduction through generations, where the role of genomic information is overwhelming.

A general concept would say:

- the gene is a combination of (one or more) nucleic acid (DNA or RNA) sequences, defined by the system (the cell, interacting with the environment, or the environment alone, in cases of subcellular or pre-cellular systems), that corresponds to a product (RNA or polypeptide).

In cases of unambiguous coding at the nucleic acid level, the defining role of the system is diminished, but the general concept proposed is not overcome (Figure 1).

Correspondence	Inherited NA segments	Gene products	Definition of the gene	Regulation (activation/repression)	Strategies: development of
UNI-VOCAL	I	1			Speed (bacteria)
MULTI-VOCAL	I. > 1 Genomic rearrangements	1			
	II. 1 Overlapping genes A. homologous products a. utilization of introns and exons b. initiation/termination of transcription/translation B. non-homologous products a. change in translation frames	> 1			Complexity (eukaryotes)

Figure 1 - The type of correspondence, univocal or multivocal, between the inherited nucleic acid (NA) segments and their products, distinguishes two types of genes. In the univocal type, predominant in bacteria, the non-genetic components of the system (ST) are only regulatory and the gene is defined in the NA structure. In the multivocal type, more common in eukaryotes, the ST also participates in the definition of the gene.

Man-made nucleic acid sequences would have to be submitted to a functional test, *in vitro* or *in vivo*, to be able to be called genes. Their status, as genes of one or another type, would be relative to the systems where they succeeded in working (Figure 2).

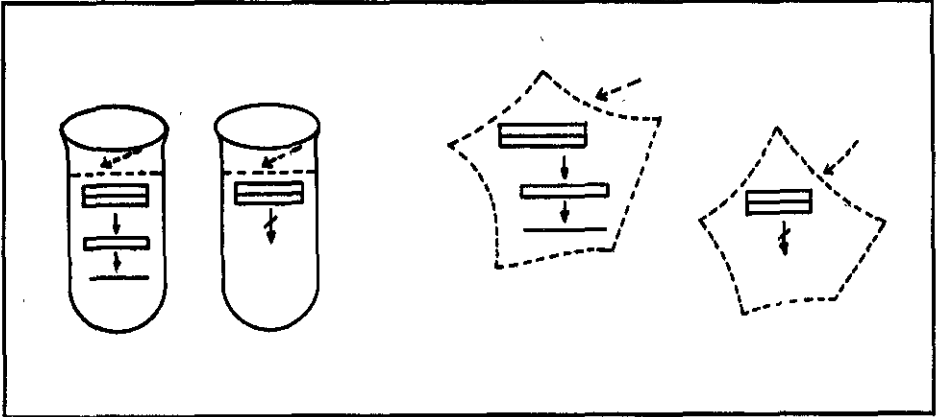





Figure 2 - The systemic concept of the gene. The question is asked: is this nucleic acid sequence a gene? The answer is obtained from tests made with decoding systems. The sequence (  ) will be a gene for the system(s) where function is produced (  , RNA;  , protein). To the left, *in vitro* (test tubes or sub-, a- or pre-cellular ponds) systems; to the right, cellular systems.

When a nucleic acid sequence is examined, it is not possible to say that it is a gene that actually produces an RNA in a cell, even if we can be sure that it does so *in vitro*. If it shows homology to a known gene, and is present in cellular or viral genomes, it could be called a gene, but if no proof is obtained for a functional role, it might end up classified as a pseudogene (Li, 1983). The same rationale applies to open reading frames, which could code for polypeptides, but remain in this limbo status until their translation is documented.

The solution to the ambiguity problem, proposed by Lewin (1990) "When the sequences representing proteins overlap or have alternative forms of expression, we may reverse the usual description of the gene. Instead of saying "one gene-one polypeptide" we may describe the relationship as "one polypeptide-one gene". Thus we may regard the sequence actually responsible for the production of the polypeptide (including introns as well as exons) as the gene, while recognizing that from the perspective of another protein, part of this same sequence may also belong to its gene" - ought to be regarded only as a technical solution, adequate to the observer's needs. It still follows the classical mendelian approach, of trying to uncover the inheritance mechanisms, starting from the phenotype.

This solves the problem of reaching the actual configuration of the message that was translated into the known product. In some instances only, the procedure will also identify the pieces of DNA that participated in building that message. Some cases of RNA editing will escape this attempt.

Besides being limited in fulfilling the conceptual goal, the proposed reversal of the information flow (Lewin, 1990) disobeys physiology. Realistic learning of how cells work would be attained through another reversal, of the results of the technical procedures used, to build the physiological picture. If a concept intends to be used in full for didactic, heuristic and research purposes, it should try to portray or reflect reality faithfully. It seems that the systemic definition satisfies these criteria.

The systemic approach indicates the need for deeper investigation on defining the exact meaning of the component parts of the living systems (Fleischaker, 1990) and on how they interact with each other and with the environment. It is still too early to decide whether all hereditary and ontogenetic information of an organism is provided by its genome, just to be sequentially read or activated, or whether some may exist, for instance, in the form of "positional information" (Basile and Guimarães, 1972) or of self-templating cellular structures whose informational content, to be transmitted, may not depend entirely on the expression of particular nucleic acid sequences (Lewin, 1990).

## ACKNOWLEDGMENTS

Financial support came from CNPq-Brasil. We thank Alfredo Pereira Junior and the Theoretical Biology Group of Botucatu, Crodowaldo Pavan and Pedro Henrique Saldanha for criticisms and discussions.

Publication supported by FAPESP.

## RESUMO

A correspondência unívoca entre um gene e um polipeptídeo foi contestada por muitos exemplos de ambiguidades. Uma lista rapidamente crescente de correspondência um-para-muitos ou muitos-para-um inclui: rearranjos genômicos, processamento alternativo de transcritos, módulos de tradução sobrepostos, editoração de RNA, modos alternativos de tradução e clivagem de poliproteínas.

A mensagem genômica requer interpretação através de descodificação por um sistema sofisticado de recuperação de informações, que também deve conter algum tipo de informação. Enfatiza-se o significado pleno da célula inteira, como uma unidade.

O gene é uma combinação de (uma ou mais) sequências de ácidos nucleicos (DNA ou RNA), *definida pelo sistema* (a célula inteira, em interação com o ambiente, ou somente o ambiente, em sistemas subcelulares ou pré-celulares), que corresponde a um produto (RNA ou polipeptídeo).

## REFERENCES

- Atkins, J.F., Elseviers, D. and Gorini, L. (1972). Low activity of beta-galactosidase in frameshift mutants of *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* 69: 1192-1195.
- Atkins, J.F., Gesteland, R.F., Reid, B.R. and Anderson, C.W. (1979). Normal tRNAs promote ribosomal frameshifting. *Cell* 18: 1119-1131.
- Basile, R. and Guimarães, R.C. (1972). Aspectos biológicos e bioquímicos da diferenciação celular. *Ciênc. Cult.* 24: 105-149.
- Benyajati, C., Spocrel, N., Haymerle, H. and Ashburner, M. (1983). The messenger for alcohol dehydrogenase in *Drosophila melanogaster* differs in its 5' end in different developmental stages. *Cell* 33: 125-133.
- Benzer, S. (1955). Fine structure of a genetic region in bacteriophage. *Proc. Nat. Acad. Sci. USA* 41: 344-354.
- Chen, S.H., Habib, G., Yang, C.Y., Gu, Z.W., Lee, B.R., Weng, S.A., Silberman, S.R., Cal, S.J., Deslypere, L., Rosseneu, M., Gotto, A.M. Jr., Li, W.H. and Chan, L. (1987). Apolipoprotein b-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science* 238: 363-366.
- Crabtree, G.R. and Kant, J.A. (1982). Organization of the rat gamma-fibrinogen gene: alternative mRNA splice patterns produce the gamma A and gamma B (gamma') chains of fibrinogen. *Cell* 31: 159-166.
- De Lange, T., Kooter, F.M., Michels, P.A. and Borst, P. (1983). Telomere conversion in trypanosomes. *Nucleic Acids Res.* 11: 8149-8165.
- Douglass, J., Civelli, O. and Herbert, E. (1984). Polyprotein gene expression: generation of diversity of neuroendocrine peptides. *Annu. Rev. Biochem.* 53: 665-715.
- Early, P., Rogers, J., Davis, M., Calame, K., Bond, M., Wall, R. and Hood, L. (1980). Two mRNAs can be produced from a single immunoglobulin  $\mu$ -gene by alternative RNA processing pathways. *Cell* 20: 313-319.
- Falk, R. (1986). What is a gene? *Stud. Hist. Phil. Sci.* 17: 133-173.
- Fleischaker, G.R. (1990). Origins of life: an operational definition. *Orig. Life Evol. Biosph.* 20: 127-137.
- Guimarães, R.C. (1986). O gene como uma molécula de RNA. *Ciênc. Cult. (Suppl.)* 38: 939-940.
- Herman, R.C. (1989). Alternatives for the initiation of translation. *TTBS* 14: 219-222.
- Hull, D. (1974). *Philosophy of Biological Science*. Prentice-Hall, Englewood Cliffs, 189 p.
- Jacob, F. and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3: 318-356.
- Kitcher, P. (1982). Genes. *Brit. J. Phil. Sci.* 33: 337-359.
- Leder, P., Honjo, T., Packman, S., Shaw, D., Nau, M. and Norman, B. (1974). The organization and diversity of immunoglobulin genes. *Proc. Nat. Acad. Sci. USA* 71: 5109-5115.
- Lewin, B. (1990). *Genes IV*. John Wiley & Sons, New York, p. 109.
- Li, N.H. (1983). Evolution of duplicate genes and pseudogenes. In: Nei, M. and Koehn, R.H. *Evolution of genes and proteins*. Sinauer, Sunderland, 15-33.
- Maki, R., Roeder, W., Traunecker, A., Sidman, C., Wabi, M., Raschke, W. and Tonegawa, S. (1984). The role of DNA rearrangement and alternative RNA processing in the expression of immunoglobulin delta genes. *Cell* 24: 353-365.

- Mayr, E. (1982). *The Growth of Biological Thought. Diversity, Evolution and Inheritance*. Harvard University Press, Cambridge, Mass., p. 736.
- Nawa, H., Kotani, H. and Nakanishi, S. (1984). Tissue-specific generation of two preprotachykinin mRNAs from one gene by alternative RNA splicing. *Nature* 312: 729-734.
- Normark, S., Bergstrom, S., Edlund, T., Grundstrom, T., Jaurin, B., Lindberg, F.P. and Olsson, O. (1983). Overlapping genes. *Annu. Rev. Genet.* 17: 499-525.
- Nasmyth, K.A., Tatchell, K., Hall, B.D., Astell, C. and Smith, M. (1981). A position effect in the control of transcription of yeast mating-type loci. *Nature* 289: 244-250.
- Powell, L.M., Wallis, S.C., Pease, R.J., Edwards, Y.H., Knott, T.J. and Scott, J. (1987). A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell* 50: 831-840.
- Shaw, D.C., Walker, J.E., Northrop, F.D., Barrell, B.G., Godson, G.N. and Fiddes, J.C. (1978). Gene K, a new overlapping gene in bacteriophage G4. *Nature* 272: 510-515.
- Simpson, L. and Shaw, J. (1989). RNA editing and mitochondrial cryptogenes of kinetoplastid protozoa. *Cell* 57: 355-366.
- Strathern, J.N. and Herskowitz, I. (1979). Asymmetry and directionality in production of new types during clonal growth: the switching pattern of homothallic yeast. *Cell* 17: 371-381.
- Thomas, S.M., Lamb, R.A. and Paterson, R.G. (1988). Two mRNAs that differ by two nontemplated nucleotides encode the amino terminal proteins P and V of the Paramyxovirus SV5. *Cell* 54: 891-902.
- Yanofsky, C. (1967). Gene structure and protein structure. *Sci. Am.* 216: 80-94.
- Young, R.A., Hagenbuchle, O. and Schübler, U. (1981). A single mouse alpha-amylase gene specifies two different tissue-specific mRNAs. *Cell* 23: 451-458.
- Ziff, E.B. (1980). Transcription and RNA processing by the DNA tumour viruses. *Nature* 287: 491-499.

(Received July 11, 1991)