PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Introductory Remarks to the Fourth Session: Gene Organization and Evolution

W. F. Bodmer

Phil. Trans. R. Soc. Lond. B 1981 **293**, 173-176
doi: 10.1098/rstb.1981.0070

Phil. Trans. R. Soc. Lond. B **293**, 173–176 (1981)
Printed in Great Britain

Introductory remarks to the fourth session: gene organization and evolution

By W. F. Bodmer, F.R.S.
Imperial Cancer Research Fund, Lincoln's Inn Fields,
London WC2A 3PX, U.K.

The essence of evolution is change in the linear sequence of nucleotides of the DNA, giving rise to altered or new genes and their corresponding products. It is these changes that determine the complex three-dimensional structure of proteins and their conglomerates. Thus, to seek an understanding of protein evolution we must first seek to understand evolution at the DNA level and the constraints that this may place on the evolution of protein structure. In this paper I shall review briefly some of the implications of recent advances in our knowledge of gene structure at the DNA level for the understanding of gene organization and its evolution (see also Bodmer 1981).

Genes in pieces

The simple notion of a 1:1 relation between DNA sequence and amino acid sequence was shattered by the discovery a few years ago that genes occur in pieces that may be put together and shuffled in various ways (see Crick (1979) for a review). The nucleotide sequences that code for the amino acid sequences of protein products are interrupted by intervening non-coding sequences in most higher organisms. Many different coding regions, corresponding perhaps to different domains of a protein, may contribute to making a given gene product. As Blake (1978) first put it succinctly: genes in pieces imply proteins in pieces. Moreover, a given coding region may contribute to more than one gene product, as is well documented for the immunoglobulins (Early *et al.* 1980). The generally assumed mechanism for producing the correct messenger from which the protein sequence is read involves first, the production of a primary transcript from the whole DNA sequence including intervening regions, and then RNA processing by splicing out the transcripted sequences corresponding to the non-coding regions. However, the possibility that in some cases transcripts may be made directly from non-contiguous DNA sequences, for example by looping out the part of the DNA strand to be omitted, is not ruled out.

It is now reasonable to propose that the minimal functional unit is the structural DNA sequence that, uninterrupted by any intervening sequences, codes for a protein domain. Domains defined in this way may well be different and much smaller than those usually considered in protein structure analysis. As Crick (1979) emphasized, shuffling around such structures is a convenient way of combining properties of parts of various different proteins into a new protein. Obvious examples would be the signal peptides found at the end of newly synthesized integral membrane and secreted proteins, the active sites of certain classes of enzymes, particular functional regions of a protein such as that embracing the haem group in haemoglobin and regions with special structural features such as the collagen-like sequence found in complement component C1q (see Porter & Reid 1979). If such structural domains are inserted into intervening sequences, it is likely that the exact position of their insertion is not critical. This lack of a requirement for precision of such transposition events may be a great advantage of intervening sequences, allowing more flexible combinations of various different structural domains to be formed.

Prokaryotes so far appear to have no intervening sequences and as a result have a much more precisely controlled genetic organization. Through this, however, they probably forego the flexibility of evolutionary change that must have been critical for the evolution of the higher eukaryotes. A major change from prokaryotes to eukaryotes is the organization of their genetic material into a nucleus surrounded by a nuclear membrane, and it was perhaps this development that allowed flexibility in the control of gene expression through complex processing of messenger RNA. This, in turn, may well have led to the more flexible organization of gene structure, involving intervening sequences. The probable evolutionary price to pay for this flexibility has been a substantial overall increase in the DNA content per genome.

Gene clusters and their evolution

The significance of gene duplication for the evolution of new genetic functions has been discussed since Bridges's original description of the phenomenon (Bridges 1919). The simple theory is that once a gene has been duplicated, copies can diverge from their original and so acquire new functions without jeopardizing those fulfilled by the original gene. Gene clusters in this sense are quite distinct from the classical bacterial operon. The basic and most intensively studied molecular model for a mammalian gene cluster is the haemoglobin β cluster (see, for example, Weatherall *et al.* 1979). Gene clustering is now turning out to be a widespread and almost universal phenomenon in higher eukaryotes. Thus, the overall organization of the eukaryotic genome must be viewed in terms of gene clusters of greater or lesser complexity. It is the gene clusters, rather than individual genes, that must define the basic genetic functions.

A schematic view of the genetic organization hierarchy is illustrated in figure 1. This shows the gene cluster as the basic functional unit, while the structural DNA domain is the minimal function unit from which genes and their clusters are built. Gene clusters may be very complex, as in the immunoglobulins or the major histocompatibility systems HLA in man and H2 in the mouse, or relatively simple, as in the haemoglobins, or perhaps even just involving a single product, as may be the situation, for example, for some of the glycolytic enzymes. It is known from somatic cell genetic studies in particular that at least in most cases, the glycolytic enzymes are not controlled by closely linked genes (see, for example, Evans *et al.* (eds) 1979).

The evolution of a gene cluster may be quite complex and, apart from the usual types of changes such as base pair substitutions and deletions, may involve one or more of the following phenomena.

1. Domains may be duplicated individually within a cluster, and transposed within the cluster or to other gene clusters.

2. Domains may be deleted or inverted.

3. Unrelated sequences can be inserted into a gene cluster by transposition.

4. There are six ways to read a DNA sequence, three in each direction, and so, including frame shifts and inversions, it may be possible to produce a variety of different amino acid

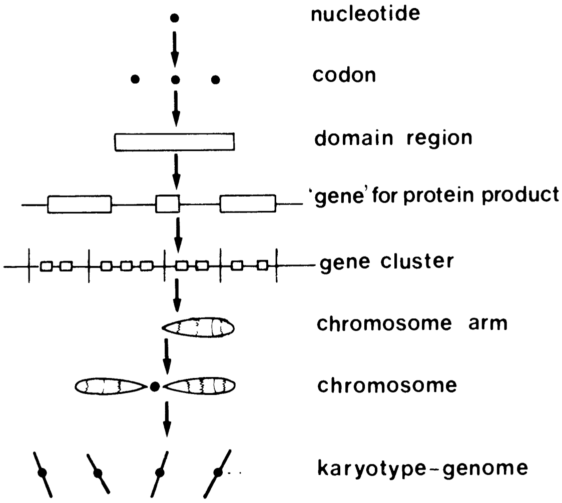

Figure 1. A scheme for the hierarchy of genetic organization.

sequences from the same original nucleotide sequence. In addition, of course, domains after duplication, inversion or transposition will diverge, following the usual processes of base substitution, deletion and addition.

5. Different products may be formed from different combinations or subsets of the same basic set of domains within a gene cluster.

When a new amino acid sequence is formed that has no obvious relation to any previous sequence, such as will happen, for example, in the translation of a frame-shifted stretch of DNA or the translation of a previously untranslated region (Nishioka *et al.* 1980), how is the new amino acid sequence put to use? Can a more or less arbitrary amino acid sequence perform some function, however inefficiently, and then gradually evolve to be more efficient? When such a new sequence evolves, can it function in a way that is related to the function of the products already made by the region? If the concept suggested by Michael Sternberg and colleagues, and others, that there may be a limited number of protein conformations, is correct, then perhaps the notion that a more or less arbitrary amino acid sequence may be able to perform some function is not entirely outrageous. In future it should be possible to test this by using recombinant DNA techniques to produce, for example, substantial frame-shift changes and so make a new protein *in vitro*, or in the bacterium, and test its function. Certainly it is the genetic structure of an organism that moulds the phenotype and both limits and defines the range of evolutionary possibilities.

The number of gene clusters and genome complexity

There has been much argument in the past about the number of functional genes in higher organisms. As already emphasized, genetic complexity in terms of function must now be counted in terms of the numbers of gene clusters rather than the number of individual genes. The new discoveries of molecular genetics provide the basis for at least an approximate estimate of this complexity. To calculate the number of gene clusters we need to know (*a*) the proportion of the total of 3×10^9 base pairs in higher eukaryotes that are functional, (*b*) the number of genes per gene cluster, (*c*) the distribution of sizes of gene products, and (*d*) the distribution of coding ratios, namely the proportion of the DNA sequence within a gene cluster that actually codes for amino acids. Reasonable, though admittedly imprecise, estimates are as follows: (*a*) approximately 50 % of the total DNA in functional clusters, (*b*) an average cluster size of 15 gene products (about the geometric mean between the sizes of the major histocompatibility system and haemoglobin clusters), (*c*) an average gene product size of 300 amino acids or approximately 1000 base pairs, and (*d*) a coding ratio of 1:30 as in the haemoglobins (Jeffreys 1979). From this, one would estimate the mean total number of clusters as $(1.5 \times 10^9)/(30 \times 15 \times 1000)$, or approximately 3300. Multiplying by 15 suggests about 50000 different gene products. These estimates are obviously subject to a considerable margin of error, but it does seem likely that the total number of clusters is in the range of, say, 3000 to a maximum of 15000, while the total number of different protein products is likely to be somewhere between 50000 and 100000. Any given cell type is likely to be synthesizing, at any given time, a small proportion of the products from any given gene cluster. Also, since differentiation is mediated by differential gene expression, a given cell is likely to express products from a relatively small proportion, say 10–20 % at most, of all gene clusters. This certainly means that the number of protein products made by any given cell is unlikely to be more than a few thousand.

A further constraint on overall complexity, as measured in terms of numbers of different genetic functions, is that gene clusters occur in related families. Examples of such families are the haemoglobin α and β clusters and myoglobin, or the immunoglobulin heavy and light chains coupled perhaps with β2 microglobulin, a part of the HLA–ABC product and also of the *Thy1* product. Functional complexity is achieved by using combinations of combinations and fine tuning of variations on major themes, rather than by using a wide variety of very different products. Detailed studies of the structure and function of systems of proteins such as the enzymes of glycolysis, and their organization at the DNA level, are the essential background needed to understand how the genome is organized and how this influences its function and evolution.

References (Bodmer)

Blake, C. C. F. 1978 Do genes-in-pieces imply proteins-in-pieces? *Nature, Lond.* **273**, 267.

Bodmer, W. F. 1981 Gene clusters, genome characterisation and complex phenotypes: When the sequence is known, what will it mean? *Am. J. hum. Genet.* (In the press.)

Bridges, C. B. 1919 Duplication. *Anat. Rec.* **15**, 357.

Crick, F. 1979 Split genes and RNA splicing. *Science, N.Y.* **204**, 264–271.

Early, P., Roger, J., Davis, M., Calame, K., Bond, M., Wall, R. & Hood, L. 1980 Two mRNAs can be produced from a single immuno globulin μ gene by alternative RNA processing pathways (II). *Cell* **20**, 313–319.

Evans, H. J., Hamerton, J. L., Klinger, H. P. & McKusick, V. A. (eds.) 1979 *Human Gene Mapping 5* (*Edinburgh Conference, 5th International Workshop on Human Gene Mapping; Birth defects*) (*Original article series*, vol. 15, no. 11). The National Foundation.

Jeffreys, A. J. 1979 DNA sequence variants in the $^{G}\gamma$-, $^{A}\gamma$-, δ- and β-globin genes of man. *Cell* **18**, 1–10.

Nishioka, Y., Leder, A. & Leder, P. 1980 Unusual α-globin-like gene that has cleanly lost both globin intervening sequences. *Proc. natn. Acad. Sci. U.S.A.* **77**, 2806–2809.

Porter, R. R. & Reid, K. B. M. 1979 Activation of the complement system by antibody–antigen complexes: the classical pathway. *Adv. Protein Chem.* **33**, 1–71.

Weatherall, D. J., Clegg, J. B., Wood, W. G. & Pasvol, G. 1979 Human haemoglobin genetics. In *Human genetics: possibilities and realities* (*Ciba Symp.* no. 66), pp. 147–186.