Proc. Natl. Acad. Sci. USA
Vol. 79, pp. 6196–6200, October 1982
Biochemistry

Relationship between the total size of exons and introns in protein-coding genes of higher eukaryotes

(gene structure/gene size/molecular evolution)

HIROTO NAORA AND NICHOLAS J. DEACON

Molecular Biology Unit, Research School of Biological Sciences, The Australian National University, Canberra, A.C.T. 2601, Australia

Communicated by D. G. Catcheside, July 19, 1982

ABSTRACT **We have attempted to ascertain the correlation between the genetic information content in the exons and the surrounding intron sequences with regard to their spatial arrangement within a gene. A comparison is made of the sizes, taken from recent publications, of exons and introns of ≈80 different protein-coding chromosomal genes, mostly from higher eukaryotes. The exons of these genes do not show very marked variation in size and can be classified into three major discrete and two minor additional size groups, whereas individual introns vary considerably in size within and between genes. Notwithstanding, the overall length of all introns present within a given gene is a function of the total size, mostly corresponding to the total genetic information content, of the exons. Three cases that violate this exon-size dependency of introns are genes coding for (*i*) histone H1, feather keratin, and interferons, (*ii*) tubulin and actin, and (*iii*) silk fibroin. The exons of these genes are larger than 0.7 kilobase pair in total size and the genes show a strong sequence homogeneity among the repetitious family members or internal repeats of coding sequences within the gene. We propose that conservation of sequences, which is required by the family members, internal repeats, or the entire gene, would actually motivate the removal of introns.**

A unique feature of eukaryote genes is the mosaic arrangement of exons and introns, with some exceptions, from which a primary transcript is formed. The intron-derived sequences must then be removed to form mature mRNA (1). In general, a protein molecule is composed of a set of functional "units," which are encoded by one or more exons separated by introns and which could serve for rapid protein evolution (2). Recent examinations of protein-coding sequences suggest that these sequences in exons were evolved as multiple tandem repeats of the primordial and subsidiary building blocks, followed by various alterations during a long evolutionary period (3, 4). The accumulation in exons of multiple repeats of building block sequences caused an increase of genetic information content within the gene and thus must have affected the gene's organization, including the storage system. Therefore, it seems possible that a complex correlation exists between the increased genetic information content in the exons and the surrounding intron sequences with regard to their spatial arrangement. During the past few years, gene structure has been enthusiastically examined (1) but a systematic search has not yet been carried out on the general pattern of the possible exon–intron correlation. Knowledge of such a correlation should contribute towards a better understanding of the evolution of genes.

Using the structural data on various genes, we have compared the sizes of exons and introns within an individual gene. In general, our results show that genes greater than ≈0.55 kilobase pair (kbp) possess introns. Total intron size is a function of total exon size, showing an initial rapid increase in intron/exon ratio from 0:1 to 6:1 in the total exon size range 0.55 to 0.8 kbp. The genes that violate this correlation are those that show a strong sequence homogeneity among family members of repetitious genes or among internal repeats of the coding sequences within the gene—e.g., silk fibroin gene.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U. S. C. §1734 solely to indicate this fact.

RESULTS

The sizes of exons and introns of ≈80 different protein-coding (but not tRNA- and rRNA-coding) genes—mostly from higher eukaryotes—are taken from recent publications. In some cases, yeast mitochondrial genes are included but these are discussed separately from the chromosomal genes. We define the size of a gene to be that of an assumed transcription unit, starting from the capping or homologous 5′-terminal site and ending at the sequence corresponding to the poly(A) addition or homologous 3′-terminal site (5). The sizes of exons and introns referred to here were determined primarily from the nucleotide sequence analysis data by using Chambon's consensus sequence (1) or by an R-looping or heteroduplex analysis, or both.

Sizes of Individual Exons. The genes cited here vary considerably in size from 0.4 to 38 kbp. However, the exons of these genes do not show very marked variation in size and can be classified into three major discrete and two minor additional size groups, as shown in Fig. 1A. Mean lengths (±SEM) are 52 ± 2, 140 ± 2, 223 ± 3, and 299 ± 5 base pairs (bp) and larger than 340 bp. Curiously, the first three values roughly correspond to the sizes of DNA associated with the nucleosome-linker region, core particle, and entire nucleosome, respectively (15), but the significance of their correlation remains to be elucidated. In this paper, we shall refer to the size classes as "50," "140," "200," "300," and ">300" bp. The histograms of the first three size classes exhibit somewhat sharp peaks.

The 140-bp sequences are the most abundant class of exons present in the protein-coding genes examined so far with the exception of globin genes which possess more 200-bp exons. However, 200-bp exons (exon II) of α- and β-globin genes have been considered to consist of two fused 140-bp class exons (6, 16).

Short exons (50 bp) that code for a signal peptide are generally present close to the 5′-terminal region, but in one case, in the internal region (17). The short exons also appear repeatedly in the middle portion of some genes where they code for the internally repeated peptide unit (10, 11, 18, 19).

The 300-bp size class, which can be seen at the shoulder of the 200-bp size class (Fig. 1A), and the >300-bp size class, including the large fibroin exon (20), represent a minority of the exons. Some of the large exons—e.g., the last exon of the ovalbumin gene (21)—include a long 3′-untranslated region or are

Abbreviations: bp, base pair(s); kbp, kilobase pair(s).

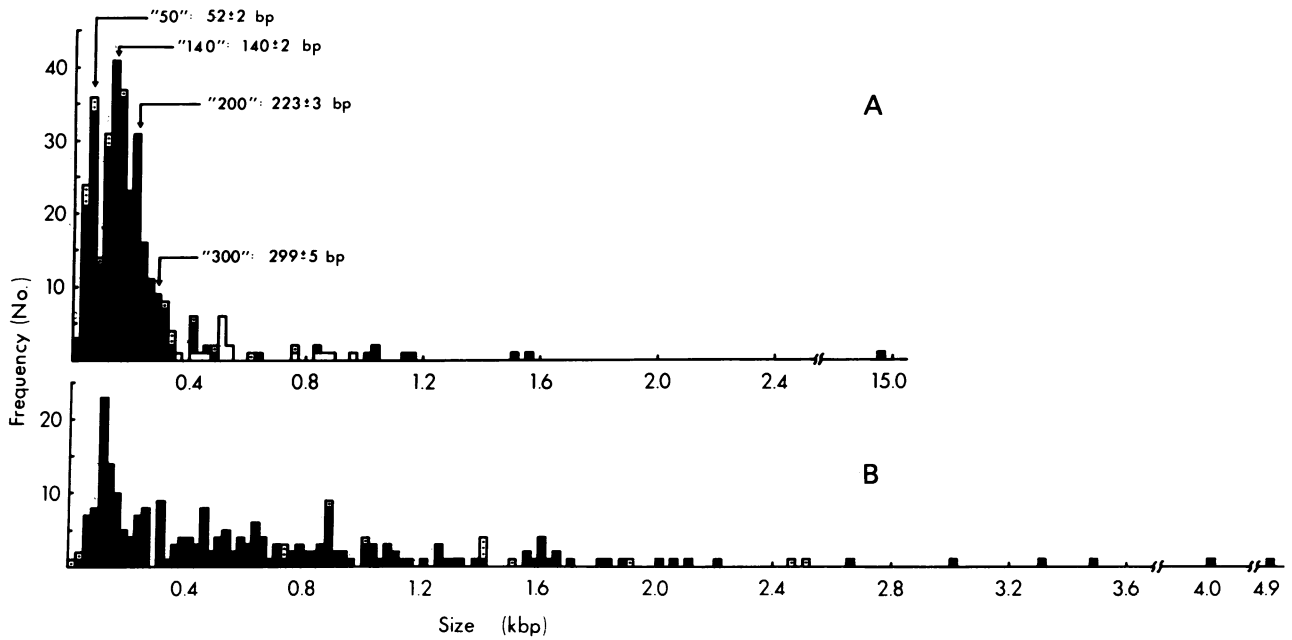

FIG. 1. Size distribution of individual exons (*A*) and introns (*B*) of protein-coding genes. The sizes of exons and introns are taken from the protein-coding genes referred to in Fig. 2 and also from genes for soy bean leghemoglobin (6), French bean phaseolin (7), and human and mouse immunoglobulin heavy and light chains (V, D, J, and C regions) (8–14). Mean lengths (±SEM) of 50-, 140-, 200-, and 300-bp size classes (*A*) were obtained with 62, 125, 68, and 19 exons, respectively. The 300-bp class is only seen as the shoulder of the 200-bp peak and can be seen more clearly on a larger scale. Exons and introns of mitochondrial genes are shown in the dotted areas but are excluded from calculations of the mean length of each class. The open areas represent intron-free genes.

derived from fusion of two or more smaller exons—e.g., exon II of rat preproinsulin I gene (22).

The size distribution of the exons of plant chromosomal genes [coding for leghemoglobin (6) and phaseolin (7)] and some yeast mitochondrial genes [cytochrome *b* and cytochrome oxidase subunit 1 (23–25)] does not differ much from that of chromosomal genes of animal cells, but some yeast mitochondrial genes tend to comprise exons that are slightly larger than 140-bp exons (Fig. 1*A*).

Sizes of Individual Introns. By contrast, it was found that individual introns of various types of genes vary considerably in size from 0.04 to 4.9 kbp, although short introns ranging from 100 to 200 bp in length are slightly more abundant (Fig. 1*B*). Such a variation seems to be unrelated to the type of genes. It seems evident that a wide range of variation is a general characteristic of individual introns present within eukaryote genes.

An interesting observation is that the size of the intron present between two exons is entirely unrelated to the sizes of these two exons (data not shown). This is in sharp contrast to the positive correlation between the intergenic distance between two genes—i.e., two independent transcription units—and the total size of these genes (5). This indicates that the introns within the gene are not really equivalent, in terms of spatial arrangement, to the intergenic spacer sequences present between two clustered functional genes.

Relationship Between Total Sizes of Exons and Introns of the Gene. The relationship between the total sizes of exons and introns of a given gene is shown in Fig. 2. For example, in the rabbit β1 globin gene, introns 1 + 2 = 0.699 kbp is plotted together with the corresponding point for the total size of the exons of the gene—i.e., exons I + II + III = 0.589 kbp. Although there is a certain amount of scattering of the points, Fig. 2 clearly indicates, with some exceptions (see below), that the *total* size of introns present within the gene increases as the *total* size of the exons becomes larger—hereafter called exon-size dependency. The correlation is not strictly linear. The information available on the mechanism that underlies this nonlinear correlation is currently limited and awaits further exploitation. Our observation suggests that, although individual introns show a great variation in size within and between genes, the overall length of all introns present within a given gene is dependent upon the total size, corresponding mostly to the total genetic information content of the exons. An alternative, though not very likely, possibility that cannot be completely ruled out is that the total size of introns conversely determines the size of exons.

It is noted that most, if not all, of the small genes (less than ≈0.55 kbp in total length) do not possess any introns. The intron-free small genes are reiterated to some extent—e.g., histone H4. In general, genes larger than ≈0.55 kbp in total length possess introns. The total length of introns in these genes increases rapidly so that the total length ratios of introns/exons increase from 0:1 to 6:1 in the total exon range 0.55 to 0.8 kbp, and there is a scattering of the points corresponding to these sizes of exons. Such variation suggests that other unknown factors (or factor), in addition to exon size, may be involved in the fine adjustment of intron size or the gene in this size range would simply accept a variable, but somewhat restricted, intron size (or both).

It should be mentioned here that most, if not all, of the medium and large genes that clearly show an exon-size dependency are those for which the family members or internal repeats, if present, retain only a *weak* sequence homogeneity among them (see below).

The mitochondrial genes of *Saccharomyces cerevisiae*—i.e., cytochrome *b* (3.3 to ≈7 kbp in length) and cytochrome oxidase subunit 1 (10 kbp in length) genes—possess several introns, the sizes of which are as would be expected from the exon sizes in a manner similar to chromosomal genes (Fig. 2). However, cytochrome *b* genes from other species of yeast or mammals do not contain any introns. Similarly, cytochrome oxidase subunit 1 genes from mammals have no introns (25, 68) (see *Discussion*).

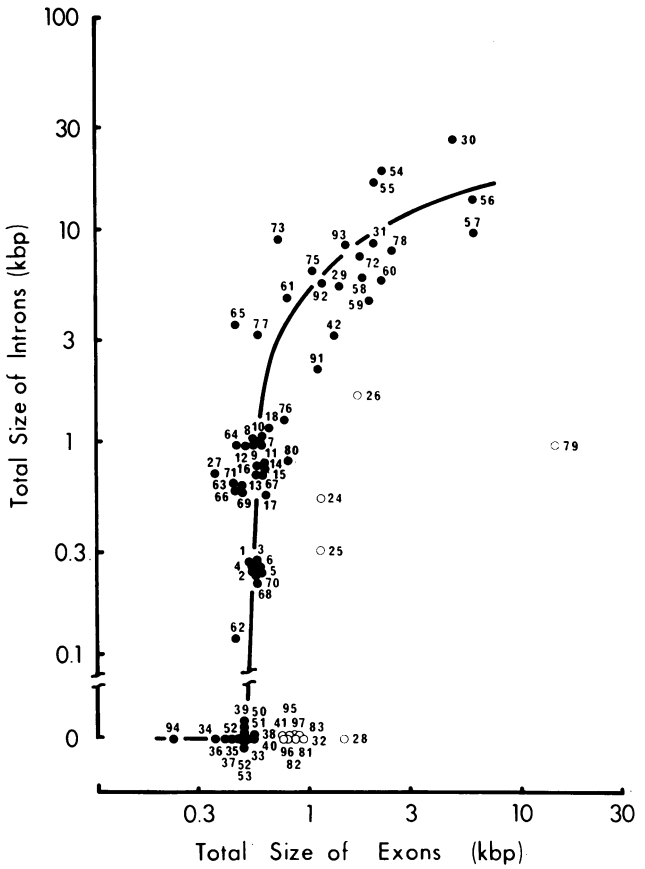

FIG. 2. Relationship between total sizes of exons and introns of protein-coding genes. The total sizes of exons of the chosen genes are plotted together with the corresponding points of the total sizes of introns. Both axes are expressed in a logarithmic scale. Data are derived from the publications cited in parentheses. Closed (●) and open (○) circles represent the genes that show and violate an exon-size dependency, respectively. The points are numbered and correspond to the following genes (an * indicates approximate size): 1, globin, human, $\psi\alpha 1$ (26); 2, $\alpha 2$ (26); 3, $\alpha 1$ (27); 4, globin, mouse, $\alpha 1$ (28); 5, globin, chicken, α^{D*} (29–31); 6, α^{A*} (29–31); 7, globin, human, ε (32); 8, $^{G}\gamma$ (33); 9, $^{A}\gamma$ (33); 10, δ (34); 11, β (35); 12, globin, rabbit, $\psi\beta 1$ (36); 13, $\beta 1$ (37); 14, globin, mouse, β^{maj} (38); 15, β^{min} (38); 16, globin, chicken, β^* (39, 40); 17, globin, *Xenopus laevis*, α (41); 18, β (41); 24, actin, sea urchin (42); 25, yeast (43); 26, *Drosophila melanogaster* (44); 27, metallothionein-I (45); 28, β-tubulin, human, small gene (the introns assumed to be absent) (46); 29, β-tubulin, human, largest gene (46); 30, pro-α2-collagen, chicken (18, 19); 31, δ-crystallin, chicken (47); 32, histone, *D. melanogaster*, H1 (48); 33, H3 (48); 34, H4 (48); 35, H2A (48); 36, H2B (48); 37, histone, newt, H4 (49); 38, H2A (49); 39, H2B (49); 40, H3 (49); 41, H1 (49); 42, transplantation antigen*, mouse (50); 50, cuticle, *D. melanogaster*, I (51); 51, II (51); 52, III (51); 53, IV (51); 54, α-fetoprotein, mouse (52); 55, albumin, mouse (52); 56, vitellogenin, *X. laevis*, A1 (53); 57, A2 (53); 58, ovalbumin, chicken (21); 59, Y, chicken (21); 60, X, chicken (21); 61, ovomucoid, chicken (54); 62, preproinsulin, rat, I (22); 63, II (22); 64, human (55); 65, chicken (56); 66, chorion, *Antheraea polyphemus*, 18 a and b (57); 67, 401a and b (57); 68, 10a (57); 69, 292a (57); 70, 10b (57); 71, 292b (57); 72, amylase*, mouse (58); 73, prolactin, rat (59); 75, corticotropin/β-lipotropin, bovine (60); 76, growth hormone, rat (61); 77, lysozyme, chicken (62); 78, conalbumin, chicken (63); 79, fibroin, *Bombyx mori* (20); 80, growth hormone, human (64); 81, feather keratin, chicken (65); 82, interferon, human, IFN-β (66); 83, IFN-α (67); 91 and 92, mitochondrial cytochrome *b*, *S. cerevisiae*, strain D273-10B (23) and strain L14-4A (23), respectively; 93, mitochondrial cytochrome oxidase, subunit 1 (24); 94 and 95, mitochondrial ATPase, subunit 9 (25) and subunit 6 (25), respectively; 96 and 97, mitochondrial cytochrome oxidase, subunit 2 (25) and subunit 3 (25), respectively.

Exceptional Cases. There are three exceptional cases that violate the above exon-size dependency. It is of particular interest that in all of these cases examined so far the total size of exons is larger than 0.7 kbp and the genes show a *strong* sequence homogeneity among the repetitious family members or internal repeats of coding sequences within the gene, suggesting that these genes or individual family members have strongly conserved their exon sequences.

Case 1: Large genes containing no introns. These include genes for histone H1 of *Drosophila* (48) and newt (49), chicken feather keratin (65), and human interferons (IFN-α and IFN-β) (66, 67) and are characterized by the *total absence* of introns, despite the large size of the exons.

All of these genes, with the exception of the interferon IFN-β gene, are reiterated to some extent and exhibit a strong homogeneity in the coding sequences among their family members. This was clearly demonstrated in histone genes (48). The interferon (IFN-α) gene family also shows extensive sequence similarities in the coding regions—i.e., up to 94% (69).

If we accept Gilbert's proposal (70) that introns facilitate the evolution of new proteins by shuffling the exons coding for functional protein units, it would be extremely difficult for multigene family members to maintain strong sequence homogeneity if they contain introns, because of unnecessary exon shuffling promoted by the presence of introns. Therefore, it seems possible that, because the presence of introns would become harmful to a multigene family, a selection pressure has forced these genes to eliminate their introns. This may suggest that these genes possessed some introns at an early stage but lost them during evolution. In fact, some possible spliced sites were found to be in the conserved regions of human interferon (IFN-α) genes (details to be published elsewhere).

The chorion genes in silkmoth are arranged on chromosomes in a manner similar to that of the histone genes, but they contain one intron of a reasonable size (0.2–0.7 kbp) (Fig. 2). However, the family members show some sequence divergence (57). Our interpretation is that because some diversification among family members was acceptable or even beneficial, the chorion genes might retain one intron.

Case 2: Tubulin and actin genes. Both tubulin and actin—large "housekeeping" proteins—are encoded by multigene families. Human β-tubulin genes show an extensive size variation, yet, except for some family members, tubulins are well conserved (46). The largest gene (6.8 kbp) possesses four exons interrupted by three introns and the total size of these introns does correspond to the value expected from its exon size (Fig. 2). We propose that the human β-tubulin genes are in the process of intron elimination and therefore the various sizes represent the family members at different stages of this processing. It seems likely, therefore, that the 6.8-kbp gene has been slightly processed or has not yet been processed, whereas the ones close to the size of mRNA have become the intron-free genes. Some members are smaller than mRNA and probably have become nonfunctional genes (46).

A similar example may be seen in actin genes that have different numbers of introns (42–44, 71). The total sizes of these introns are shorter than those that would be expected from the sizes of total exons (Fig. 2). The protein-coding regions of the genes are conserved within an organism, although some members are more closely related than others (71). It should be noted that the introns are present in various locations in different members or in different organisms, or both (71, 72). Recently, Schuler and Keller proposed that the primordial actin gene had at least six exons and five introns, but the present-day actin genes have lost many introns during evolution (72). This proposal is in accordance with our interpretation that the members of the actin gene family are in the process of, or have completed, intron elimination, in a different manner from member to mem-

ber or organism to organism (or both). Furthermore, it also seems possible that the necessity to maintain less stringent sequence homogeneity among the family members might allow the actin genes to retain one or two short introns.

Case 3: Fibroin gene. The gene for fibroin is fundamentally an extensively reiterated array of 18-bp "crystalline" repeats, which in turn comprise larger repeats (20, 73). At least 10 larger crystalline repeats are arranged alternatively with "amorphous" coding domains. It should be noted that a strong sequence homogeneity of all the repeats within the gene and the alternating arrangement of crystalline and amorphous coding sequences are vital for the proper functioning of the fibroin protein (73). The fibroin gene (16 kbp) contains only one short (0.97 kbp) intron at a region distal from the repetitious sequence gene core (20). Here again, if there were many introns or long introns, or both, in the midst of a highly repetitive sequence gene core, the presence of introns would be harmful to sequence homogeneity of the repeats and the special sequence arrangement.

Like the fibroin gene, the pro-α2-collagen gene consists of multiple internal repeats of short nucleotide sequences (18, 19). However, the sequence homogeneity of repeats is far less stringent than that of the fibroin gene (18, 19). This implies that diversification would be acceptable, even beneficial, because of the less stringent functional constraints on collagen proteins. It is probably for this reason that the pro-α2-collagen gene still possesses full-sized introns.

DISCUSSION

It is controversial at present whether introns were present together with exons in the ancestral eukaryotic genes or whether they were inserted into preformed intron-free genes during evolution. Furthermore, we do not know yet whether the intron sequences were merely the by-products of the mechanism that produced repeats of the coding-sequence building blocks (3) and fused repeats. The findings of this paper are compatible with either hypothesis. More information is required before this can be discussed further.

The following conclusion may be drawn from our observation. If a chromosomal protein-coding gene is larger than $\approx$0.55 kbp, it is, with some exceptions, composed of more than one short ($\approx$50 to 200 bp; mostly $\approx$140 bp) exon interrupted by introns, of which the total size is a function of the total genetic information content of the exons. Our observation that there is an abundance of exons $\approx$140 bp in length is unlikely to be the result of an artifact caused by an uneven collection of gene types, as the proteins encoded by the genes referred to here vary considerably in size, molecular structure, cellular distribution, and biological function.

An interesting observation is that there is a clear correlation between conservation of the stringent sequence homogeneity among the family members or internal repeats of the genes and violation of the exon-size dependency. Tandemly repetitious intron-free gene families are subject to unequal crossover (74) and hence promote the rate of sequence correction between family members. Indeed, some of the "exceptional" genes are tandemly reiterated and are free of introns—e.g., histone and feather keratin. As mentioned earlier, most of the genes that exhibit an exon-size dependency do not show such a stringent sequence homogeneity among the family members or internal repeats. Introns were rapidly diversified in terms of size and nucleotide sequences (1). The presence of such rapidly diversifying introns might permit rapid diversification of genes or family members. When stringent functional constraints are required for the products of family members or for the entire length of the product, the introns must be eliminated to avoid diversification. Therefore, we propose that conservation of sequences, which is required by the family members, repeats, or the entire gene, would actually motivate the removal of introns. The removal of intron sequences from intron-containing genes has been observed to cause severe destabilization or total loss of the messenger activity of transcripts (75, 76) or a remarkable degree of size variation (polymorphism) of the gene (77). The organism should develop a mechanism to retain the stability of the messenger activity or minimize size variations of the gene when eliminating introns. Therefore, elimination of introns, though maintaining necessary functions, is probably a costly event.

Our proposal may partly account for the total absence of introns in prokaryote genes. Prokaryotes, perhaps for the sake of streamlining their energetic burden, lack the "luxury" of large genomes with room for evolutionary experiments with alterations to gene sequences. For reasons such as these, involving the haploid nature and "singleness" of prokaryote genes, sequence conservation in all prokaryote genes might be as vital as that in the family members of the exceptional (e.g., histone H1) genes, which have eliminated all introns. This may also be true for mitochondrial genes—i.e., if they are interpreted in a similar way to that for the presence of various forms of β-tubulin and actin genes (see above), it is possible that most mitochondrial genes have lost introns to conserve their sequences, but some might merely have accepted various sizes of introns or are in the process of intron elimination. However, mitochondrial genes are different from chromosomal genes in several aspects (68)—thus, other unknown factors (or factor) also may be involved. This is of course a preliminary interpretation which requires further examination.

Our results make it possible to estimate the length of extra DNA required for the expression of sequences that code for any given protein from the known size of the polypeptide chain. This can be done by summing the size of total exon DNA, extra sequences (e.g., untranslated regions), total intron DNA (from Fig. 2), and territorial DNA sequences (5). For example, a chicken ovomucoid molecule (M_r 28,000, secretory protein) consists of three main domains, each of which is composed of two smaller subdomains, one approximately 17–23 and the other approximately 37–46 amino acid residues in length (54). In other words, the coding sequence for the ovomucoid polypeptide is composed of three sets of 50-bp and 140-bp exons, tandemly reiterated (total length of coding sequence, 0.56 kbp). With the addition of sequences [0.263 kbp = 0.069 + 0.006 + 0.053 + 0.135 kbp in the ovomucoid gene (78)], coding for a signal peptide, the initiation and termination signals, and the 5′- and 3′-untranslated regions of mRNA, the total length of exons becomes 0.82 kbp.

Fig. 2 gives the value of $\approx$4 kbp (actually 4.7 kbp for the introns of the ovomucoid gene) for the size of the introns that should associate with 0.82 kbp of exons, if the gene is not exceptional with regard to its spatial requirement for introns. Thus, the complete gene (an independent transcription unit) is equal to 5.5 kbp in length. As reported previously (5), the 5.5-kbp chromosomal gene of higher eukaryotes requires another 4 + 4 = 8-kbp (approximate) territorial DNA sequence. Thus, 5.5 + 8 = 13.5-kbp DNA sequences (i.e., 13.5/0.56 = 24-fold larger DNA sequences) are finally needed for the protein molecule encoded by only a 0.56-kbp DNA sequence. Nevertheless, a eukaryote cell still possesses a further excess of "nonfunctional" DNA sequences on chromosomes.

We thank Dr. D. Clark-Walker, Department of Genetics, The Australian National University, and Prof. V. Holoubek, University of Texas, for valuable discussions and comments and Mrs. G. Hines for her assistance in preparing this manuscript.

1. Breathnach, R. & Chambon, P. (1981) *Annu. Rev. Biochem.* **50**, 349–383.
2. Artymiuk, P. J., Blake, C. C. F. & Sippel, A. E. (1981) *Nature (London)* **290**, 287–288.
3. Ohno, S. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 7657–7661.
4. Ohno, S., Kato, K., Hozumi, T. & Matsunaga, T. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 132–136.
5. Naora, H. & Deacon, N. J. (1982) *Differentiation* **21**, 1–6.
6. Jensen, E. Ø., Paludan, K., Hyldig-Nielsen, J. J., Jørgensen, P. & Marcker, K. A. (1981) *Nature (London)* **291**, 677–679.
7. Sun, S. M., Slightom, J. L. & Hall, T. C. (1981) *Nature (London)* **289**, 37–41.
8. Bernard, O. & Gough, N. M. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 3630–3634.
9. Blomberg, B., Traunecker, A., Eisen, H. & Tonegawa, S. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 3765–3769.
10. Sakano, H., Hüppi, K., Heinrich, G. & Tonegawa, S. (1979) *Nature (London)* **280**, 288–294.
11. Early, P., Huang, H., Davies, M., Calame, K. & Hood, L. (1980) *Cell* **19**, 981–992.
12. Hieter, P. A., Max, E. E., Seidman, J. G., Maizel, J. V., Jr., & Leder, P. (1980) *Cell* **22**, 197–207.
13. Seidman, J. G., Max, E. E. & Leder, P. (1979) *Nature (London)* **280**, 370–375.
14. Sakano, H., Rogers, J. H., Hüppi, K., Brack, C., Traunecker, A., Maki, R., Wall, R. & Tonegawa, S. (1979) *Nature (London)* **277**, 627–633.
15. McGhee, J. D. & Felsenfeld, G. (1980) *Annu. Rev. Biochem.* **49**, 1115–1156.
16. Gō, M. (1981) *Nature (London)* **291**, 90–92.
17. Lingappa, V. R., Lingappa, J. R. & Blobel, G. (1979) *Nature (London)* **281**, 117–121.
18. Dickson, L. A., Ninomiya, Y., Bernard, M. P., Pesciotta, D. M., Parsons, J., Green, G., Eikenberry, E. F., de Crombrugghe, B., Vogeli, G., Pastan, I., Fietzek, P. P. & Olsen, B. R. (1981) *J. Biol. Chem.* **256**, 8407–8415.
19. Wozney, J., Hanahan, D., Tate, V., Boedtker, H. & Doty, P. (1981) *Nature (London)* **294**, 129–135.
20. Tsujimoto, Y. & Suzuki, Y. (1979) *Cell* **18**, 591–600.
21. Heilig, R., Perrin, F., Gannon, F., Mandel, J. L. & Chambon, P. (1980) *Cell* **20**, 625–637.
22. Lomedico, P., Rosenthal, N., Efstratiadis, A., Gilbert, W., Kolodner, R. & Tizard, R. (1979) *Cell* **18**, 545–558.
23. Nobrega, F. & Tzagaloff, A. (1980) *J. Biol. Chem.* **255**, 9828–9837.
24. Bonitz, S. G., Coruzzi, G., Thalenfeld, B. E. & Tzagaloff, A. (1980) *J. Biol. Chem.* **255**, 11927–11941.
25. Clark-Walker, G. D. & Sriprakash, K. S. (1981) *J. Mol. Biol.* **151**, 367–387.
26. Proudfoot, N. J. & Maniatis, T. (1980) *Cell* **21**, 537–544.
27. Michelson, A. M. & Orkin, S. H. (1980) *Cell* **22**, 371–377.
28. Nishioka, Y. & Leder, P. (1979) *Cell* **18**, 875–882.
29. Deacon, N. J., Shine, J. & Naora, H. (1980) *Nucleic Acids Res.* **8**, 1187–1199.
30. Engel, J. D. & Dodgson, J. B. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 2596–2600.
31. Dodgson, J. B., McCune, K. C., Rusling, D. J., Krust, A. & Engel, J. D. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 5998–6002.
32. Baralle, F. E., Shoulders, C. C. & Proudfoot, N. J. (1980) *Cell* **21**, 621–626.
33. Slightom, J. L., Blechl, A. E. & Smithies, O. (1980) *Cell* **21**, 627–638.
34. Spritz, R. A., DeRiel, J. K., Forget, B. G. & Weissmann, S. M. (1980) *Cell* **21**, 639–646.
35. Lawn, R. M., Efstratiadis, A., O'Connell, C. & Maniatis, T. (1980) *Cell* **21**, 647–651.
36. Lacy, E. & Maniatis, T. (1980) *Cell* **21**, 545–553.
37. Hardison, R. C., Butler, E. T., III, Lacy, E., Maniatis, T., Rosenthal, N. & Efstratiadis, A. (1979) *Cell* **18**, 1285–1297.
38. Konkel, D. A., Maizel, J. V., Jr., & Leder, P. (1979) *Cell* **18**, 865–873.
39. Naora, H. (1981) *Shigei Med. J.* **2**, 53–67.
40. Villeponteau, B. & Martinson, H. (1981) *Nucleic Acids Res.* **9**, 3731–3746.
41. Patient, R. K., Elkington, J. A., Kay, R. M. & Williams, J. G. (1980) *Cell* **21**, 565–573.
42. Overbeek, P. A., Merlino, G. T., Peters, N. K., Cohn, V. H., Moore, G. P. & Kleinsmith, L. J. (1981) *Biochim. Biophys. Acta* **656**, 195–205.
43. Gallwitz, D. & Sures, I. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 2546–2550.
44. Fyrberg, E. A., Kindle, K. L., Davidson, N. & Sodja, A. (1980) *Cell* **19**, 365–378.
45. Durnam, D. M., Perrin, F., Gannon, F. & Palmiter, R. D. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 6511–6515.
46. Cowan, N. J., Wilde, C. D., Chow, L. T. & Wefald, F. C. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 4877–4881.
47. Jones, R. E., Bhat, S. P., Sullivan, M. A. & Piatigorsky, J. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 5879–5883.
48. Kedes, L. H. (1979) *Annu. Rev. Biochem.* **48**, 837–870.
49. Stephenson, E. C., Erba, H. P. & Gall, J. G. (1981) *Nucleic Acids Res.* **10**, 2281–2295.
50. Steinmetz, M., Moore, K. W., Frelinger, J. G., Sher, B. T., Shen, F.-W., Boyse, E. A. & Hood, L. (1981) *Cell* **25**, 683–692.
51. Snyder, M., Hirsh, J. & Davidson, N. (1981) *Cell* **25**, 165–177.
52. Kioussis, D., Eiferman, F., van de Rijin, P., Gorin, M. B., Ingram, R. S. & Tilghman, S. M. (1981) *J. Biol. Chem.* **256**, 1960–1967.
53. Wahli, W., Dawid, I. B., Wyler, T., Weber, R. & Ryffel, G. U. (1980) *Cell* **20**, 107–117.
54. Stein, J. P., Catterall, J. F., Kristo, P., Means, A. R. & O'Malley, B. W. (1980) *Cell* **21**, 681–687.
55. Ullrich, A., Dull, T. J., Gray, A., Brosius, J. & Sures, I. (1980) *Science* **209**, 612–615.
56. Perler, F., Efstratiadis, A., Lomedico, P., Gilbert, W., Kolodner, R. & Dodgson, J. (1980) *Cell* **20**, 555–566.
57. Jones, C. W. & Kafatos, F. C. (1980) *Cell* **22**, 855–867.
58. Young, R. A., Hagenbüchle, O. & Schibler, U. (1981) *Cell* **23**, 451–458.
59. Chien, Y.-H. & Thompson, B. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 4583–4587.
60. Nakanishi, S., Teranishi, Y., Watanabe, Y., Notake, M., Noda, M., Kakidani, H., Jingami, H. & Numa, S. (1981) *Eur. J. Biochem.* **115**, 429–438.
61. Barta, A., Richards, R. I., Baxter, J. D. & Shine, J. (1981) *Proc. Natl. Acad. Sci. USA* **78**, 4867–4871.
62. Jung, A., Sippel, A. E., Grez, M. & Schültz, G. (1980) *Proc. Natl. Acad. Sci. USA* **77**, 5759–5763.
63. Cochet, M., Gannon, F., Hen, R., Maroteaux, L., Perrin, F. & Chambon, P. (1979) *Nature (London)* **282**, 567–574.
64. DeNoto, F. M., Moore, D. D. & Goodman, H. M. (1981) *Nucleic Acids Res.* **9**, 3719–3730.
65. Barone, E. D., Powell, B. & Rogers, G. E. (1981) *Proc. Aust. Biochem. Soc.* **14**, 84 (abstr.).
66. Lawn, R. M., Adelman, J., Franke, A. E., Houck, C. M., Gross, M., Najarian, R. & Goeddel, D. V. (1981) *Nucleic Acids Res.* **9**, 1045–1052.
67. Nagata, S., Mantei, N. & Weissmann, C. (1980) *Nature (London)* **287**, 401–408.
68. Clark-Walker, G. D. (1982) in *DNA and Evolution: Natural Selection and Genome Size*, ed. Cavalier-Smith, T. (Wiley, Chichester, England), in press.
69. Goeddel, D. V., Leung, D. W., Dull, J. T., Gross, M., Lawn, R. M., McCandliss, R., Seeburg, P. H., Ullrich, A., Yelverton, E. & Gray, P. W. (1981) *Nature (London)* **290**, 20–26.
70. Gilbert, W. (1978) *Nature (London)* **271**, 501.
71. Firtel, R. A. (1981) *Cell* **24**, 6–7.
72. Schuler, M. A. & Keller, E. B. (1981) *Nucleic Acids Res.* **9**, 591–604.
73. Gage, L. P. & Manning, R. F. (1980) *J. Biol. Chem.* **255**, 9444–9450.
74. Smith, G. P. (1976) *Science* **191**, 528–535.
75. Wickens, M. P., Woo, S., O'Malley, B. W. & Gurdon, J. B. (1980) *Nature (London)* **285**, 628–634.
76. Gruss, P. & Khoury, G. (1980) *Nature (London)* **286**, 634–637.
77. Manning, R. F. & Gage, L. P. (1980) *J. Biol. Chem.* **255**, 9451–9457.
78. Catterall, J. F., Stein, J. P., Kristo, P., Means, A. R. & O'Malley, B. W. (1980) *J. Cell Biol.* **87**, 480–487.