J Mol Evol (1988) 27:45–55

Journal of
Molecular Evolution

Structure of Vertebrate Genes: A Statistical Analysis Implicating Selection

M.W. Smith

Department of Biology, Johns Hopkins University, Charles and 34th Streets, Baltimore, Maryland 21218, USA

Summary. This paper conducts a statistical analysis of the size distributions of exons and six other gene parts [the transcription unit, introns, intervening DNA (sum of introns), mRNA (sum of exons), and leader and trailer regions of mRNA] as well as the number of exons, the percentage of introns, the placement of introns within the gene, and the potential for frameshifts from coding exon shifts. The first seven variables measured in base pairs fit log-normal distributions. Significant correlations between the sizes of intervening DNA and mRNA, the sizes of leader and trailer regions, and the sizes of introns and flanking exons exist. Introns occur at nonrandom frequencies within the codon frame, in untranslated regions, and relative to the frameshift potential from exon movement or duplication. These nonrandom patterns in gene structure demonstrate that models of gene evolution must incorporate selective processes.

Key words: Exons — Introns — Molecular evolution — Statistical analysis — Vertebrates — Gene structure — Selection

Introduction

The modular structure of eucaryotic genes has stimulated discussion about its significance and evolution (Blake 1978, 1983, 1985; Gilbert 1978, 1985; Cavalier-Smith 1985; Doolittle 1985; Rogers 1985; Sudhof et al. 1985; Cech 1986). These modules include not only segments of coding regions (exons) separated by intervening sequences (introns) but also leader and trailer regions bearing or near controlling elements (e.g., promoters and enhancers).

One way of gaining further insight into the forces that have molded the number and sizes of these modules or gene parts is to conduct statistical studies of eucaryotic gene structure. Naora and Deacon (1982b) described the frequency distributions of the sizes of exons and introns and speculated that classes of exon size exist. They also illustrated a positive relationship between the sizes of intervening DNA and mRNA. Their study suffers from the lack of a rigorous and comprehensive statistical analysis, a heterogeneous set of genes from different phyla, and a biased sample overly influenced by highly related genes. In a more rigorous study, Blake (1983) demonstrated a correlation between polypeptide size and number of exons.

Blake's (1985) review noted the need for additional statistical characterization of the sizes of gene parts. The present study of intron and exon organization and structure was undertaken using sequenced vertebrate genes. The design of data set restrictions included only independent genes with introns. The features examined were the sizes in base pairs of the transcription unit, intervening DNA (sum of introns), mRNA (sum of exons), mRNA leader untranslated region, mRNA trailer untranslated region, exons, and introns. Additional variables were number of exons, percent introns (intervening DNA/transcription unit), and intron locations within the transcription unit. The analysis identified nonrandom patterns that implicate selective constraints.

Offprint requests to: M.W. Smith

Table 1. Locations of intron interruptions in the genes analyzed

Gene	Species[a]	No. of exons	Location of introns[b]	Reference
α_1-acid glycoprotein	Rn	6	02110	Reinke and Feigelson 1985
Actin β	Gg	6	90010	Kost et al. 1983
Adenosine deaminase	Hs	12	02221000201	Valerio et al. 1985
Albumin	Rn	15	12020202020 209	Sargent et al. 1981
Aldolase B	Gg	9	91010010	Burgess and Penhoet 1985
Antifreeze	Pa	2	2	Davies et al. 1984
Apolipoprotein-C3	Hs	4	912	Protter et al. 1984
Calmodulin	Gg	8	9011120	Simmen et al. 1985
Casein	Rn	9	90000009	Jones et al. 1985
Cholecystokinin	Rn	3	91	Deschenes et al. 1985
Chymotrypsin-B	Rn	7	102010	Bell et al. 1984
Crystallin-δ	Gg	17	90000222112 00002	Ohno et al. 1985
Corticolipotropin-β	Mm	3	90	Notake et al. 1983
Cytochrome P450-C	Rn	7	901122	Sogawa et al. 1984
Elastase-1	Rn	8	1022100	Swift et al. 1984
Enkephalin-A	Rn	3	90	Rosen et al. 1984
Factor VIII	Hs	26	21111112010 1120011110 0100	Gitschier et al. 1984
Globin-1β1	Xl	3	20	Meyerhof et al. 1984
Glyceraldehyde-3-phosphate dehydrogenase	Gg	12	92000202002	Stone et al. 1985
Gastrin	Hs	3	91	Ito et al. 1984
Growth hormone somatomammotropin	Hs	5	1000	Selby et al. 1984
Growth hormone releasing factor	Hs	5	9222	Mayo et al. 1985
HLA-DR	Hs	5	1119	Das et al. 1983
Haptoglobin-1	Hs	5	2111	Bensi et al. 1985
Hypoxanthine phosphoribosyltransferase	Mm	9	02000210	Melton et al. 1984
Insulin	Cp	3	91	Chan et al. 1984
Int-1	Mm	4	210	Ooyen and Nusse 1984
Interferon-γ	Hs	4	000	Gray and Goeddel 1982
Interleukin-3	Mm	5	0000	Miyatake et al. 1985
Intermediate filament—keratin	Hs	8	0200021	Marchuk et al. 1984
Lactin	Rn	5	1000	Cooke and Baxter 1982
Luteinizing hormone	Rn	3	00	Jameson et al. 1984
Lysozyme	Gg	4	112	Jung et al. 1980
MHC antigen-DCβ	Hs	5	1111	Larhammar et al. 1983
Metallothionein-2	Mm	3	11	Searle et al. 1984
Myosin light chain-2	Rn	7	001120	Nudel et al. 1984
Natriodilatin	Hs	3	00	Nemer et al. 1984
Natriuretic factor	Hs	3	00	Seidman et al. 1984
Nerve growth factor α	Mm	5	1210	Evans and Richards 1985
Oxytocin	Bt	3	01	Ruppert et al. 1984
Ovalbumin-Y	Gg	8	9000100	Heilig et al. 1982
Ovomucoid	Gg	8	1020202	Stein et al. 1980
Pituitary glycoprotein hormone α	Hs	4	910	Fiddes and Goodman 1981
Pepsinogen	Hs	9	20102200	Sogawa et al. 1983
Parathyroid hormone	Rn	3	92	Heinrich et al. 1984
Renin-1	Hs	10	201022200	Miyazaki et al. 1984
Rhodopsin	Hs	5	1200	Nathans and Hogness 1984
Ribosome protein L30	Mm	5	9021	Weidemann and Perry 1984
Ribosome protein L32	Mm	4	902	Dudov and Perry 1984
Seminal vesicle secretion IV	Rn	3	19	Harris et al. 1983
Somatostatin	Hs	2	0	Shen and Rutter 1984
Superoxide dismutase	Hs	5	0120	Levanon et al. 1985
Tubulin-bα1	Hs	4	010	Hall and Cowan 1985
Thy-1	Mm	3	11	Chang et al. 1985
Thymidine kinase	Gg	7	022000	Kwoh and Engler 1984
Triose phosphate isomerase	Hs	7	120101	Brown et al. 1985
Trypsin	Rn	5	1210	Craik et al. 1984
Vasopressin	Bt	3	01	Ruppert et al. 1984
Whey acidic protein	Rn	4	111	Campbell and Rosen 1984

Table 1. Continued

Gene	Species[a]	Location of introns[b]	Reference
Partially defined genes			
c-*abl*	Mm	10	Wang et al. 1984
Adenine phosphoribosyltransferase	Mm	2101	Dush et al. 1985
Complement factor B	Hs	11110121010	Campbell and Porter 1983
Collagen	Gg	011000	Yamada et al. 1983
Cytochrome *c*	Rn	01	Scarpulla 1984
Dihydrofolate reductase	Mm	21202	Crouse et al. 1982
Fibronectin	Rn	210	Tamkun et al. 1984
Fibrinogen	Hs	211	Fornace et al. 1984
Glucagon	Hs	222	Bell et al. 1983
Insulin-like growth factor II	Hs	20	Dull et al. 1984
Kininogen	Hs	90010010000	Kitamura et al. 1985
Microglobin β_2	Mm	11	Parnes and Seidman 1982
c-*myc*	Mm	91	Stanton et al. 1984
Myosin heavy chain-α	Rn	001	Mahdavi et al. 1984
Myosin light chains 1 and 3	Gg	011119	Nabeshima et al. 1984
Nicotinic acetycholine receptor δ	Gg	10022112010	Nef et al. 1984
Plasminogen activator	Hs	011121212110	Ny et al. 1984
Platelet-derived growth factor	Hs	101	Chiu et al. 1984
Phosphoglycerate kinase	Hs	2220220011	Michelson et al. 1985
Protein C	Hs	1011101	Foster et al. 1985
Prothrombin	Hs	11122	Degen et al. 1983
Pyruvate kinase	Gg	9100120021	Lonberg and Gilbert 1985
c-*ras*-1 (Harvey)	Hs	020	Sekiya et al. 1984
Relaxin	Hs	1	Hudson et al. 1983
c-*src*	Gg	9122111111	Takeya and Hanafusa 1983
Tachykinin	Bt	901111	Nawa et al. 1984
Vasoactive intestinal peptide	Hs	222	Bodner et al. 1985

Note: The first data set of fully defined genes was used in analyses of continuous variables. The second set of partially defined genes was used, in combination with the first set, for intron location analysis. Genes are listed with the species, number of exons (no. of exons; first set only), location of introns in a 5′ to 3′ orientation, and a reference. Locations of breaks by introns between codons were scored as 0 for none, between nucleotides one and two as 1, between nucleotides two and three as 2, and within the untranslated region as 9

[a] Species (abbreviations) are *Bos taurus* (Bt), *Cavia porcellus* (Cp), *Gallus gallus* (Gg), *Homo sapiens* (Hs), *Mus musculus* (Mm), *Pseudopleuronectes americanus* (Pa), *Rattus norvegicus* (Rn), and *Xenopus laevis* (Xl)

[b] Number of introns interrupting (location): 171 (0), 148 (1), and 95 (2) were significantly different from a random 33% ($\chi^2 = 22.0$; $p < 0.005$, 2 *df*). Frequencies of interruptions at both coding exon ends (5′ end, 3′ end of coding exon): 59 (0,0), 38 (0,1), 32 (0,2), 42 (1,0), 53 (1,1), 23 (1,2), 36 (2,0), 23 (2,1), and 22 (2,2) were not significantly different from random trinomial expecteds ($\chi^2 = 9.4$; $p > 0.05$; 2 *df*). The nine categories were reduced to three based on the remainder, when exon size is divided by three. These three categories were no (none), a plus one (+1), or a minus one (−1) remainder. These size classes of coding exons occurred at the following frequencies (type): 134 (none), 97 (+1), and 97 (−1), which were significantly different from the random expecteds ($\chi^2 = 4.8$; $p < 0.005$; 1 *df*)

Materials and Methods

Genes Analyzed and Definitions. Vertebrate protein-coding and intron-containing genes with polyadenylated messages were collected in a comprehensive, yet not exhaustive, search of the literature to July 1985. Rearranged (e.g., immunoglobulins), non-polyadenylated (e.g., histones), and alternatively spliced (e.g., troponin T) genes were not included. Criteria for inclusion in the analysis were sequenced exons, identified cap site, polyadenylation site, and intron sizes. An independent sample was systematically selected as the most recently described gene from several of a gene family or from the same gene chararacterized in several species ($n = 59$; Table 1). Size in base pairs (bp) of the mRNA leader untranslated region, trailer untranslated region, each exon, each intron, the transcription unit (cap to polyadenylation sites), mRNA (sum of exons), intervening DNA (sum of introns), percent introns (intervening DNA/transcription unit × 100), and the number of exons were determined for each of these genes. Intervening DNA and mRNA sizes add to the transcription unit size.

An additional 27 vertebrate genes, which were partially sequenced and/or differentially spliced, formed a second combined data set ($n = 76$). These additional genes were not related to genes in the first data set and had a maximal number of contiguous sequenced exons. The genes were scored for the locations of introns within the mRNA (Table 1). The scoring was breaks between codons = 0, between nucleotides one and two = 1, between nucleotides two and three = 2, and in an untranslated region = 9. Exons at the transcription unit ends were usually associated with one splice junction, middle exons with two, and introns with one.

Table 2. Average sizes and distributions of the gene parts

Variable	Sample size	Median mean	95% Q 95% CI	5% Q 5% CI	Skewness kurtosis	D p
Exons	356	133	567	42	0.435	0.066
		138	630	30	1.605	<0.01
Introns	297	600	4849	89	0.339	0.050
		603	7513	48	−0.006	0.071
mRNA leader	59	62	184	22	−0.469	0.077
		62	231	17	1.426	>0.15
mRNA trailer	59	196	1070	50	0.299	0.070
		205	1153	37	−0.479	>0.15
Transcription unit	59	4226	31,572	981	1.103	0.079
		4448	27,836	711	3.500	>0.15
mRNA	59	869	2102	453	0.954	0.104
		974	2954	321	2.584	0.113
Intervening DNA	59	3159	30,076	470	0.734	0.072
		3174	27,899	361	2.052	>0.15
Percent introns	59	76	95	43	−0.360	0.071
		75	97	40	−0.428	>0.15

Determined from the data for each variable: (nonparametric) median, 95% quantile (95% Q), 5% quantile (5% Q), and sample size; (parametric) mean, 95% and 5% confidence intervals (95% CI and 5% CI) transformed back to the linear scale; and skewness, kurtosis, the Kolmogorov–Smirnov D-statistic, and probability (p) of a worse fit to the normal distribution. A $\log_{10}$ transformation was most normal for all variables except the arc sine transformation of percent introns

Statistical Analysis. The Statistical Analysis System (SAS Institute Inc. 1982) on the Johns Hopkins University IBM4341 computer was used extensively. Null hypotheses were formulated and then tests were constructed by consulting Sokal and Rohlf (1981). Positively skewed size variable (Y) distributions were transformed by $1/Y$, $\sqrt{Y}$, $1/\sqrt{Y}$, and $\log_{10}(Y)$. Percent introns was angular transformed with arc sine Y. A null hypothesis of normality was evaluated with the Kolmogorov-Smirnov D-statistic, and by comparing confidence intervals and means to quantiles and medians. Subsequent analyses used transformed data.

Analyses of variance (ANOVAs) were performed using the General Linear Models procedure on SAS. Exon and intron sizes were tested relative to their position at the end, beginning, or middle of each gene. A second ANOVA design examined the relationships of intervening DNA, mRNA, transcription unit, mRNA leader, and mRNA trailer sizes to the number of exons.

Regressions tested the continuous variables: transcription unit, intervening DNA, mRNA, mRNA leader, mRNA trailer sizes, and the number of exons. When pairings were between mathematically dependent variables (e.g., transcription unit and intervening DNA sizes), no regression was calculated. Relationships between intron and flanking exon sizes were also analyzed. Data falling more than three standard errors from the calculated line in initial regressions were dropped as deviants by more than 99%.

The locations of intron interruptions were tested for a random frequency of occurrence by χ^2 analyses. Single interruptions were evaluated at the three possible locations within the codon frame. In a second analysis, only exons with codons at both ends were tested. Intron interruptions at the 5′ and 3′ coding exon ends occurred at three 5′ locations times three 3′ locations in nine possible combinations. A χ^2 analysis compared the observed frequencies in the nine categories to expecteds calculated from a trinomial square distribution of single codon interruption frequencies. These nine coding exon categories were grouped by the remainder when exon size was divided by three and compared to trinomial squared expecteds. Observed frequencies of intron interruptions in the mRNA leader or mRNA trailer were tested against expecteds calculated from the average sizes of these regions.

Results

Average Sizes of Gene Components

Generalities of gene structure and organization were summarized from data in the literature (Tables 1 and 2). Introns (603 bp) were on average about four times the size of exons (138 bp). The mRNA untranslated region following the termination codon (mRNA trailer = 205 bp) was about three times the size of the region preceding the start codon (mRNA leader = 62 bp). The mRNA (974 bp) plus intervening DNA (3174 bp) sizes sum to within 300 bp of the transcription unit size (4448 bp). Three-fourths of the average primary gene transcript consisted of introns. The mean sizes were based on the best fitting parametric (distribution assuming) statistics. Parametric means and confidence intervals generally approximated the equivalent nonparametric (distribution independent) medians and quantiles (Table 2).

Frequency Distributions

The sizes of the seven gene parts were most normal when transformed (Table 2, Fig. 1). Variables measured in base pairs (exons, introns, intervening DNA, mRNA, mRNA leader, mRNA trailer, and transcription unit) best fit a lognormal distribution. The sizes of these six variables each ranged over several orders of magnitude. The seventh gene measurement, percent introns, was most normal with an arc sine transformation. Distributions from all vari-

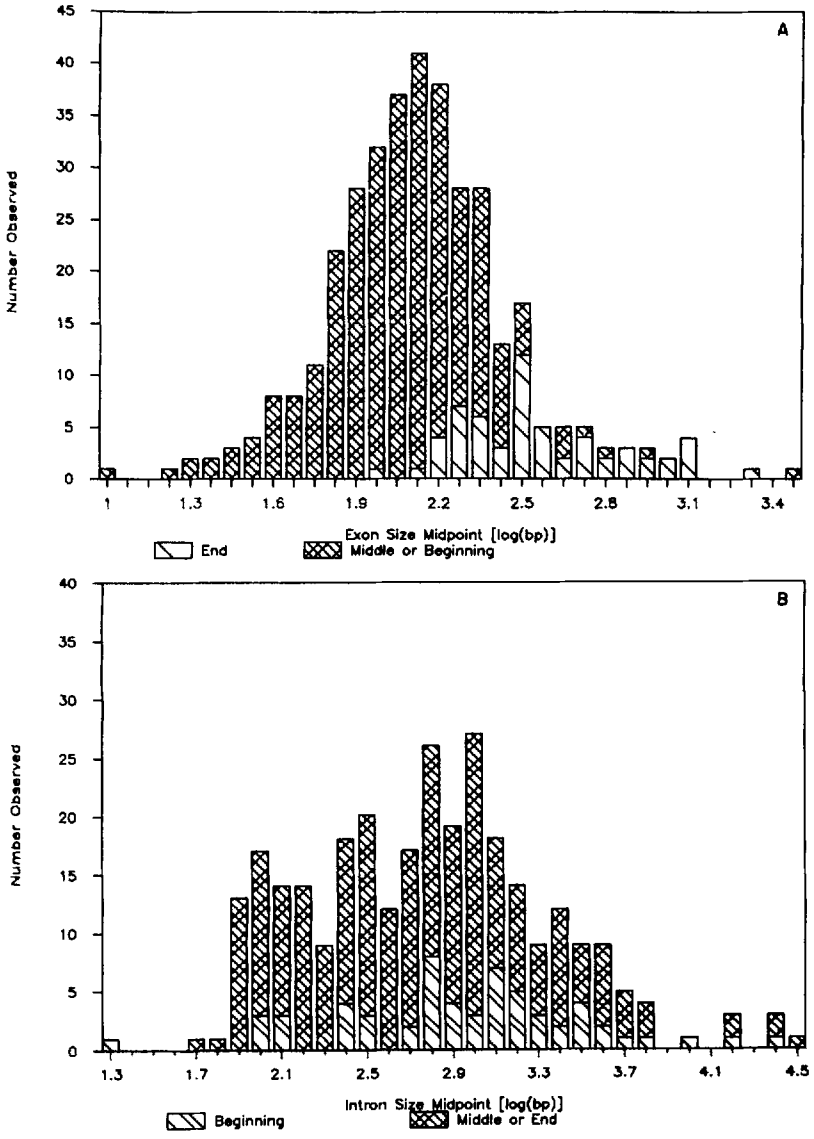

Fig. 1. Frequency distributions of **A** $\log_{10}$ base pairs exon (n = 356) and **B** intron (n = 297) sizes

ables provided estimates of means, confidence intervals, skewness, and kurtosis. Large sample sizes of exons and introns ($n \approx 300$) allowed detailed examination of their distributions.

The lognormal distributions of exon and intron sizes were not the same shape (Fig. 1). Exon sizes had a more peaked distribution than introns (Table 2; kurtosis = 1.6 and 0.0, respectively). The calculated intron size confidence intervals poorly matched the observed quantiles. Intron sizes have a sharp lower cutoff at approximately 75 bp, except for three smaller calmodulin introns, and more than one mode may be present. Both qualitative and quantitative differences exist in the shapes of exon and intron distributions.

Relationships Among the Sizes of Gene Parts

Exon and intron sizes were dependent on position within the gene (Table 3, Fig. 1). For example, the first intron was almost twice the size of a middle or end intron (least squares adjusted means = 911 bp, 573 bp, and 539 bp, respectively). In a stronger relationship, the last exon was about three times the size of a middle or first exon (356 bp, 127 bp, and 105 bp, respectively). Exon and, to a lesser extent, intron positions were important features for describing gene structure.

Number of exons was also related to gene structure (Tables 3 and 4). Transcription unit, intervening DNA, and mRNA sizes were each related to the number of exons in a gene. The model sums of squares were consistently larger when ANOVA analyses were compared to similar regressions.

Interrelationships were found between sizes of the measured gene elements (Fig. 2, Table 4). Messenger RNA untranslated region sizes were correlated to each other and to transcription unit size. Intervening DNA and mRNA sizes were interrelated. Intron size and either following or preceding

Table 3. Relationships between the sizes of gene parts and categories: number of exons, or position at beginning, end, or middle of gene

Category and size	Sums of squares		Sample size	*F*-value	*p*
	Model	Total			
Number of exons					
Intervening DNA	3.840	8.432	53	5.38	0.0002
mRNA	0.779	2.269	53	3.36	0.0057
mRNA leader region	0.176	4.548	53	0.26	0.9666
mRNA trailer region	0.935	7.385	53	0.93	0.4918
Transcription unit	2.714	5.473	53	6.32	0.0001
Position in gene					
Intron	2.293	92.456	297	3.74	0.0249
Exon	12.774	40.312	356	81.88	0.0001

Analysis of variance sums of squares (model and corrected total), sample size, *F*-value, and a significance level (*p*) are presented. Continuous size variable values were transformed to a $\log_{10}$ scale. The number of exons data set was genes with less than 10 exons

Table 4. Regression analysis on gene parts

Independent and dependent variables	Intercept SE	Slope SE	Sample size	*F*-value	*p*
Intron					
Following exon	1.935	0.083	296	5.54	0.0193
	0.100	0.035			
Preceding exon	1.817	0.085	296	8.66	0.0035
	0.082	0.029			
mRNA[a]					
Intervening DNA	0.802	0.898	58	13.92	0.0004
	0.717	0.241			
Transcription unit	0.803	0.947	58	27.89	0.0001
	0.535	0.179			
mRNA leader region[b]					
mRNA trailer region	1.199	0.264	58	9.56	0.0031
	0.200	0.085			
Number of exons					
Intervening DNA	2.864	0.113	53	24.72	0.0001
	0.122	0.023			
mRNA	2.665	0.058	53	23.58	0.0001
	0.064	0.012			
mRNA leader region	1.765	0.005	53	0.06	0.8149
	0.109	0.020			
mRNA trailer region	2.013	0.057	53	5.46	0.0234
	0.132	0.025			
Transcription unit	3.085	0.099	53	32.76	0.0001
	0.093	0.017			
Transcription unit[a]					
mRNA leader	0.906	0.244	58	5.37	0.0242
	0.382	0.105			
mRNA trailer	0.967	0.367	58	7.84	0.0070
	0.477	0.131			

Slope, intercept, standard errors, sample size, *F*-value, and significance level (*p*) calculated from each set of dependent and independent variables. All variables, except number of exons, were transformed to $\log_{10}$ scale. The data for number of exons is the same set used in Table 3

[a] Factor VIII gene was outside of data range and dropped

[b] Luteinizing hormone was dropped as a greater than 3 SE deviate in the initial regression

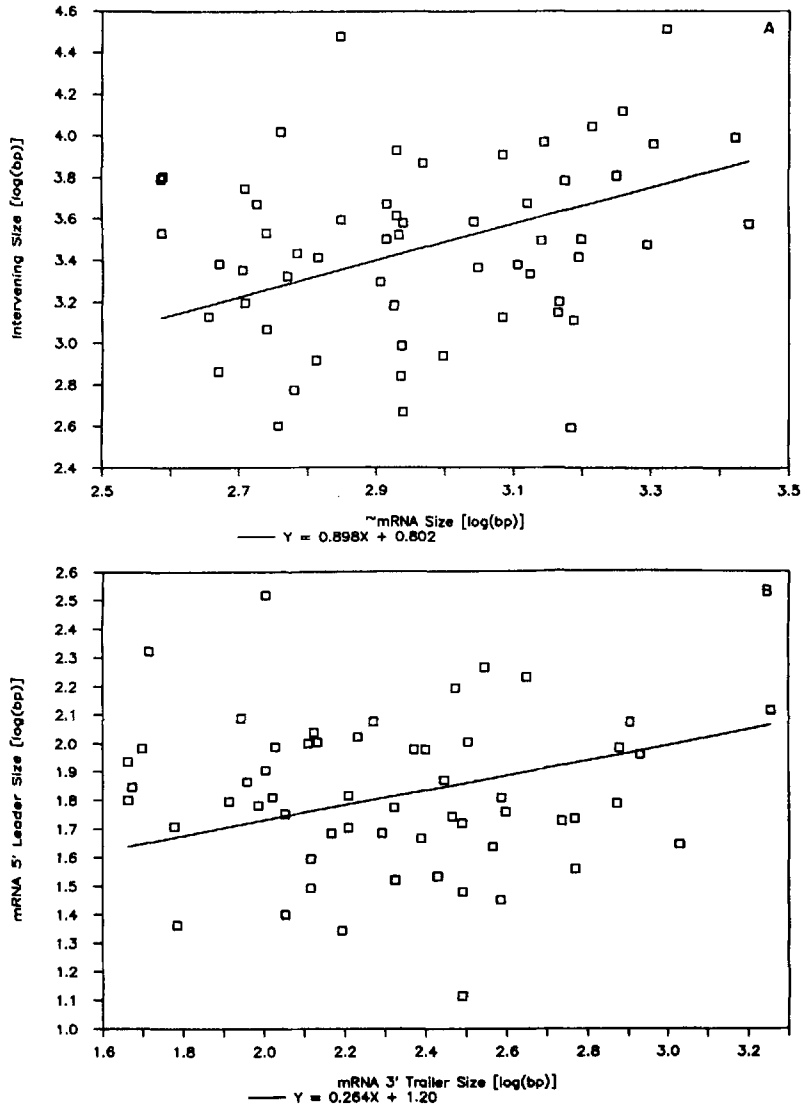

Fig. 2. A Relationship between intervening DNA (sum of introns) and ~mRNA (sum of exons) sizes; **B** relationship between mRNA 5′ leader and 3′ trailer untranslated region sizes. Variables were measured in $\log_{10}$ base pairs.

exon sizes were weakly related. The previously untested correlation matrix was indicative of size interrelationships in gene structure.

Intron Locations in mRNA Transcripts

Location of introns within the transcript was another important factor in describing gene structure (Table 1, footnote b). Introns were found between codons at significantly high frequencies (41%) and infrequently between codon nucleotides two and three (23%). The codons at middle coding exon ends were interrupted by introns in nine possible ways (see Materials and Methods), which is not significantly different from the random trinomial square expecteds. The same middle coding exons were assessed for frameshift generating potential upon insertion, deletion, or duplication within a gene. Middle coding exon sizes with remainders of zero were found in significant excess (134) of the 115 predicted. The small mRNA leader untranslated region was interrupted by introns at a frequency (83%) much higher than would be predicted by the mRNA leader and trailer target sizes ($\chi^2 = 58.3$; $p < 0.005$; 1 *df*).

Discussion

In this study gene structure and organization have been analyzed with a perspective distinct from the nucleotide sequence. The vertebrate genes available in the literature are primarily from man and rodents, and are those that researchers have known about and cloned. Even though these biases exist, they are not sufficient to explain the nonrandom patterns found. The generalities of gene structure have been presented as averages derived from a very diverse data set. Selective constraints on vertebrate gene evolution are postulated to account for the patterns observed.

Distributions of the Sizes of Gene Parts

The frequency distributions of variables measured in base pairs suggest interacting selective constraints. The distributions of the exon, intron, transcription unit, intervening DNA, mRNA, mRNA leader, and mRNA trailer sizes were each best described by a lognormal distribution. A simple normal distribution would be predicted from a series of additive factors. Lognormal distributions of the sizes of gene parts strongly implicate interacting selective constraints.

Some previous speculations on exon sizes are inconsistent with the data obtained. No evidence for several exon size classes was found (Naora and Deacon 1982b). The study by Naora and Deacon (1982b) contained many related genes and exons. In the present study, the more extensive analysis of exon sizes was not consistent with the uniform distribution reported for the chicken pyruvate kinase gene (Lonberg and Gilbert 1985). Using Lonberg and Gilbert's own criteria on this larger data set, untransformed exon sizes had a standard deviation (261 bp) greater than the mean (197 bp), which further rejects the hypothesis of a uniform exon size distribution. The log transformation best approximates a normal exon size distribution.

Selective forces were evident by examination and comparison of exon and intron size distributions (Fig. 1, Table 2). Since exons and introns are found side by side in the gene, similar distribution shapes are predicted from identical size constraints. Exons had a more peaked distribution than introns, which is consistent with stabilizing selection against extremes in exon size. Exons with sizes falling within the middle of the size distribution may be more easily processed by the mRNA splicing pathways. The intron size frequency distribution has more than one mode and a sharp lower cutoff at 75 bp. The loss of splicing activity in artificial β-globin constructs with less than 80 bp of intron verifies a minimum size requirement near the lower cutoff (Wierenga et al. 1984). The differences in peakedness, number of modes, and the lower intron size cutoff demonstrate the importance of differential selection on exon and intron sizes.

Gene Size Relationships

Exon and intron sizes were uniquely related to their positions within the gene (Table 3, Fig. 1). The last exon, containing the large mRNA trailer untranslated region, was about three times the size of the other exons within a gene. In a weaker relationship, the first intron was almost twice the size of other introns in a gene. The position effects on exon and intron sizes at opposite ends of the gene implicate differential selection.

The number of exons partially described gene structure (Tables 3 and 4). Relationships between the number of exons and the sizes of the transcription unit, mRNA, and intervening DNA were examined. In each case, ANOVAs yield larger sums of squares than identical regressions. ANOVA is not sensitive to nonlinear relationships. The more robust ANOVAs and the previous finding of lognormal distributions (Table 2) strongly suggest multiple interacting selective forces.

There were additional interrelationships between the sizes of the gene parts (Table 4). Intervening DNA and mRNA sizes were positively correlated. Untranslated region sizes were correlated to one another and to the transcription unit size. The relationship between untranslated region sizes suggests that these regions play a similar role in a cellular function (e.g., determining message stability). Significant relationships, while weak, were found between intron size and following or preceding exon size. The territorial gene model, where larger genes are flanked by correspondingly large amounts of intergenic DNA (Naora and Deacon 1982a), marginally applies to individual exons and introns. The interrelationships of the sizes of vertebrate gene parts demonstrate overall structure and organization.

Intron Locations in mRNA Transcript

Selection is most parsimonious with the nonrandom locations of introns (Table 1). Introns were found most frequently between codons (41%) and at a low frequency between codon nucleotides two and three (23%). Introns must be spliced out of the mRNA leader untranslated regions at a frequency (84%) two and a half times larger than the expected proportion based on their average target sizes. The nonrandom frequencies of intron locations are inconsistent with a simple intron insertion model lacking selection.

A testable prediction from the primordial, exon shuffling, model is an excess of exons that can easily move or duplicate without generating a downstream frameshift. For example, internally duplicating an exon of multiple three size would not cause downstream frameshifts. A significant excess (134 observed vs 115 expected) of protein-coding exons with sizes that were multiples of three was observed. The nonrandom locations of introns within the codon frame and the untranslated regions and the slight excess of coding exons less likely to cause frameshifts are most consistent with selection during evolution.

Acknowledgments. Thanks go to many colleagues at Johns Hopkins University for their interest, criticism, and helpful dis-

cussions. In particular I thank D. Powers, my fellow lab members, and my father, M.H. Smith, for advice and encouragement during the preparation of this manuscript. This work was supported by Maryland Sea Grant R/F-44 and NSF grant DEB 82-07006 to D.A. Powers. The author is a recipient of a Maryland Sea Grant predoctoral traineeship. This paper is contribution #1384 from the Department of Biology.

References

Bell GI, Sanchez-Pescador R, Laybourn PJ, Najarian RC (1983) Exon duplication and divergence in the human preproglucagon gene. Nature 304:368–371

Bell GI, Quinto C, Quiroga M, Valenzuela P, Craik CS, Rutter WJ (1984) Isolation and sequence of a rat chymotrypsin B gene. J Biol Chem 259:14265–14570

Bensi G, Raugei G, Klefenz H, Cortese R (1985) Structure and expression of the human haptoglobin locus. EMBO J 4:119–126

Blake CCF (1978) Do genes-in-pieces imply proteins-in-pieces? Nature 273:267–268

Blake CCF (1983) Exons—present from the beginning? Nature 306:535–537

Blake CCF (1985) Exons and the evolution of proteins. Int Rev Cytol 93:149–185

Bodner M, Fridkin M, Gozes I (1985) Coding sequences for vasoactive intestinal peptide and PHM-27 peptide are located on two adjacent exons in the human genome. Proc Natl Acad Sci USA 82:3548–3551

Brown JR, Daar IO, Krug JR, Maquat LE (1985) Characterization of the functional gene and several processed pseudogenes in the human triosephosphate isomerase gene family. Mol Cell Biol 5:1694–1706

Burgess DG, Penhoet EE (1985) Characterization of the chicken aldolase B gene. J Biol Chem 260:4604–4614

Campbell RD, Porter RP (1983) Molecular cloning and characterization of the gene coding for human complement protein factor B. Proc Natl Acad Sci USA 80:4464–4468

Campbell RS, Rosen JM (1984) Comparison of the whey acidic protein genes of the rat and mouse. Nucleic Acids Res 12: 8685–8697

Cavalier-Smith T (1985) Selfish DNA and the origin of introns. Nature 315:283–284

Cech TR (1986) The generality of self-splicing RNA: relationship to nuclear mRNA splicing. Cell 44:207–210

Chan SJ, Episkopou V, Zeitlin S, Karathanasis SK, MacKrell A, Steiner DF, Efstratiadis A (1984) Guinea pig preproinsulin gene: an evolutionary compromise? Proc Natl Acad Sci USA 81:5046–5050

Chang HC, Seki T, Moriuchi T, Silver J (1985) Isolation and characterization of mouse *Thy*-1 genomic clones. Proc Natl Acad Sci USA 82:3819–3823

Chiu I-M, Reddy EP, Givol D, Robbins KC, Tronick SR, Aaronson SA (1984) Nucleotide sequence analysis identifies the human c-*sis* proto-oncogene as a structural gene for platelet-derived growth factor. Cell 37:123–129

Cooke NE, Baxter JD (1982) Structural analysis of the prolactin gene suggests a separate origin for its 5′ end. Nature 297:603–606

Craik CL, Choo Q-L, Swift GH, Quinto C, MacDonald RJ, Rutter WJ (1984) Structure of two related rat pancreatic trypsin genes. J Biol Chem 259:14255–14264

Crouse GF, Simonsen CC, McEwan RN, Schimke RT (1982) Structure of amplified normal and variant dihydrofolate reductase genes in mouse sarcoma S180 cells. J Biol Chem 257: 7887–7897

Das HK, Lawrence SK, Weissmann SM (1983) Structure and nucleotide sequence of the heavy chain gene of HLA-DR. Proc Natl Acad Sci USA 80:3543–3547

Davies PL, Hough C, Scott GK, Ng N, White BN, Hew CL (1984) Antifreeze protein genes of the winter flounder. J Biol Chem 259:9241–9247

Degen SJF, MacGillivray TTA, Davie EW (1983) Characterization of the complementary deoxyribonucleic acid gene coding for human prothrombin. Biochemistry 22:2087–2097

Deschenes RJ, Haun RS, Funckes CL, Dixon JE (1985) A gene encoding rat cholecystokinin. J Biol Chem 260:1280–1286

Doolittle RF (1985) The genealogy of some recently evolved vertebrate proteins. Trends Biochem Sci 10:233–237

Dudov KP, Perry RP (1984) The gene family encoding the mouse ribosomal protein L32 contains a uniquely expressed intron-containing gene and an unmutated processed gene. Cell 37:457–468

Dull TJ, Gray A, Hayflick JS, Ullrich A (1984) Insulin-like growth factor II precursor gene organization in relation to insulin gene family. Nature 310:777–781

Dush MK, Sikela JM, Khan SA, Tischfield JA, Stanbrook PJ (1985) Nucleotide sequence and organization of the mouse adenine phosphoribosyltransferase gene: presence of a coding region common to animal and bacterial phosphoribosyltransferases that has a variable intron/exon arrangement. Proc Natl Acad Sci USA 82:2731–2735

Evans BA, Richards RI (1985) Genes for the α and γ subunits of nerve growth factor are contiguous. EMBO J 4:133–138

Fiddes JC, Goodman HM (1981) The gene encoding the common alpha subunit of the four human glycoprotein hormones. J Mol Appl Genet 1:3–18

Fornace AJ Jr, Cummings DE, Comeau CM, Kant JA, Crabtree GR (1984) Structure of the human γ-fibrinogen gene. J Biol Chem 259:12826–12830

Foster DC, Yoshitake S, Davie EW (1985) The nucleotide sequence of the gene for human protein C. Proc Natl Acad Sci USA 82:4673–4677

Gilbert W (1978) Why genes in pieces? Nature 271:501

Gilbert W (1985) Genes-in-pieces revisited. Science 228:823–824

Gitschier J, Wood WI, Goralka TM, Wion KL, Chen EY, Eaton DH, Vehar GA, Capon DJ, Lawn RM (1984) Characterization of the factor VIII gene. Nature 312:326–330

Gray PW, Goeddel DV (1982) Structure of the human immune interferon gene. Nature 298:859–863

Hall JL, Cowan NJ (1985) Structural features and restricted expression of a human α-tubulin gene. Nucleic Acids Res 13: 207–223

Harris SE, Mansson P-E, Tully DR, Burkhart B (1983) Seminal vesicle secretion IV gene: allelic difference due to a series of 20-base-pair direct tandem repeats within an intron. Proc Natl Acad Sci USA 80:6460–6464

Heilig R, Muraskowsky R, Kloepfer C, Mandel JL (1982) The ovalbumin gene family; complete sequence and structure of the Y gene. Nucleic Acids Res 14:4363–4382

Heinrich G, Kronenberg HM, Potts JT Jr, Habener JF (1984) Gene encoding parathyroid hormone. J Biol Chem 259:3320–3329

Hudson P, Haley J, John M, Cronk M, Crawford R, Haralambidis J, Treagear G, Shine J, Niall N (1983) Structure of a genomic clone encoding biologically active human relaxin. Nature 301: 628–631

Ito R, Sato K, Helmer T, Jay G, Agarwal K (1984) Structural analysis of the gene encoding human gastrin: the large intron contains an *Alu* sequence. Proc Natl Acad Sci USA 81:4662–4666

Jameson L, Chin WW, Hollenberg AN, Chang AS, Habener JF (1984) The gene encoding the β-subunit of rat luteinizing hormone. J Biol Chem 259:15474–15480

Jones WK, Yu-Lee L, Clift SM, Brown TL, Rosen JM (1985) The rat casein multigene family. J Biol Chem 260:7042–7050

Jung A, Sippel AE, Grez M, Schutz G (1980) Exons encode functional and structural units of chicken lysozyme. Proc Natl Acad Sci USA 77:5759–5763

Kitamura N, Kitagawa H, Fukushima D, Takagaki Y, Miyata T, Nakanishi S (1985) Structural organization of the human kininogen gene and a model for its evolution. J Biol Chem 260:8610–8617

Kost TA, Theodorakis N, Hughes SH (1983) The nucleotide sequence of the chick cytoplasmic β-actin gene. Nucleic Acids Res 11:8287–8301

Kwoh TJ, Engler JA (1984) The nucleotide sequence of the chicken thymidine kinase gene and the relationship of its predicted polypeptide to that of the vaccinia virus thymidine kinase. Nucleic Acids Res 12:3959–3971

Larhammar D, Hyldig-Nielsen JJ, Serenius B, Andersson G, Rask L, Peterson PA (1983) Exon-intron organization and complete nucleotide sequence of a human major histocompatibility antigen DCβ gene. Proc Natl Acad Sci USA 80:7313–7317

Levanon D, Lieman-Hurwitz J, Dafni N, Wigderson M, Sherman L, Bernstein Y, Laver-Rudich Z, Danciger E, Stein O, Groner Y (1985) Architecture and anatomy of the chromosomal locus in human chromosome 21 encoding the Cu/Zn superoxide dismutase. EMBO J 4:77–84

Lonberg N, Gilbert W (1985) Intron/exon structure of the chicken pyruvate kinase gene. Cell 40:81–90

Mahdavi V, Chambers AP, Nadal-Ginard B (1984) Cardiac α- and β-myosin heavy chain genes are organized in tandem. Proc Natl Acad Sci USA 81:2626–2630

Marchuk D, McCrohon, Fuchs E (1984) Remarkable conservation of structure among intermediate filament genes. Cell 39:491–498

Mayo KE, Cerelli GM, Lebo RV, Bruce BD, Rosenfeld MG, Evans RM (1985) Gene encoding human growth hormone-releasing factor precursor: structure, sequence, and chromosomal assignment. Proc Natl Acad Sci USA 82:63–67

Melton DW, Konecki DS, Brennand J, Caskey CT (1984) Structure, expression, and mutation of the hypoxanthine phosphoribosyltransferase gene. Proc Natl Acad Sci USA 81:2147–2151

Meyerhof W, Klinger-Mitropoulos S, Stadler J, Weber R, Knochel W (1984) The primary structure of the larval β_1-globin gene of *Xenopus laevis* and its flanking region. Nucleic Acids Res 12:7705–7719

Michelson AM, Bruns GAP, Morton CC, Orkin SH (1985) The human phosphoglycerate kinase multigene family. J Biol Chem 260:6982–6992

Miyatake S, Yokota T, Lee F, Arai K-I (1985) Structure of the chromosomal gene for murine interleukin 3. Proc Natl Acad Sci USA 82:316–320

Miyazaki H, Fukamizu A, Hirose S, Hayashi T, Hori H, Ohkubo H, Nadanishi S, Murakami K (1984) Structure of the human renin gene. Proc Natl Acad Sci USA 81:5999–6003

Nabeshima Y, Fujii-Kuriyama Y, Muramatsu M, Ogata K (1984) Alternate transcription and two modes of splicing result in two myosin light chains from one gene. Nature 308:333–338

Naora H, Deacon NJ (1982a) Clustered genes require extragenic territorial DNA sequences. Differentiation 21:1–6

Naora H, Deacon NJ (1982b) Relationship between the total size of exons and introns in protein coding genes of higher eukaryotes. Proc Natl Acad Sci USA 79:6196–6200

Nathans J, Hogness DS (1984) Isolation and nucleotide sequence of the gene encoding human rhodopsin. Proc Natl Acad Sci USA 81:4851–4855

Nawa H, Kotani H, Nakanishi S (1984) Tissue specific generation of two preprotachykinin mRNAs by alternate splicing. Nature 312:729–734

Nef P, Mauron A, Stalder R, Alliod C, Ballivet M (1984) Structure, linkage, and sequence of the two genes encoding δ and γ subunits of the nicotinic acetylcholine receptor. Proc Natl Acad Sci USA 81:7975–7979

Nemer M, Chamberland M, Sirois D, Argentin S, Drouin J, Dixon RAF, Zivin RA, Condra JH (1984) Gene structure of human cardiac hormone precursor, pronatriodilatin. Nature 312:654–656

Notake M, Tobimatsu T, Watanabe Y, Takahashi H, Mishina M, Numa S (1983) Isolation and characterization of the mouse corticotropin-β-lipotropin precursor gene and a related pseudogene. FEBS Lett 156:67–71

Nudel U, Calvo JM, Shani M, Levy Z (1984) The nucleotide sequence of a rat myosin light chain 2 gene. Nucleic Acids Res 12:7175–7186

Ny T, Elgh F, Lund B (1984) The structure of the human tissue-type plasminogen activator gene: correlation of intron and exon structures to functional and structural domains. Proc Natl Acad Sci USA 81:5355–5359

Ohno M, Sakamoto H, Yasuda K, Okada TS, Shimura Y (1985) Nucleotide sequence of a chicken δ-crystallin gene. Nucleic Acids Res 13:1593–1606

Ooyen AV, Nusse R (1984) Structure and nucleotide sequence of the putative mammary oncogene *int*-1; proviral insertions leave the protein-encoding domain intact. Cell 39:233–240

Parnes JR, Seidman JG (1982) Structure of wild-type and mutant mouse β_2-microglobin genes. Cell 29:661–669

Protter AA, Levy-Wilson B, Miller J, Bencen G, White T, Seilhamer JJ (1984) Isolation and sequence analysis of the human apolipoprotein CIII gene and the intergenic region between the apo AI and apo CIII genes. DNA 3:449–456

Reinke R, Feigelson P (1985) Rat α_1-acid glycoprotein. J Biol Chem 260:4397–4403

Rogers J (1985) Exon shuffling and intron insertion in serine protease genes. Nature 315:458–459

Rosen H, Douglass J, Herbert E (1984) Isolation and characterization of the rat proenkephalin gene. J Biol Chem 259: 14309–14313

Ruppert S, Scherer G, Schutz G (1984) Recent gene conversion involving bovine vasopressin and oxytocin precursor genes suggested by nucleotide sequence. Nature 308:554–557

Sargent TD, Jagodizinski LL, Yang M, Bonner J (1981) Fine structure and evolution of the rat serum albumin gene. Mol Cell Biol 1:871–883

SAS Institute Inc (1982) SAS user's guide: basics, 1982 ed. SAS Institute, Cary NC

Scarpulla RC (1984) Processed pseudogenes for rat cytochrome *c* are preferentially derived from one of three alternate mRNAs. Mol Cell Biol 4:2279–2288

Searle PF, Davison BL, Stuart GW, Wilke TM, Norstedt G, Palmiter RD (1984) Regulation, linkage, and sequence of mouse metallothionein I and II genes. Mol Cell Biol 4:1221–1230

Seidman CE, Bloch KD, Klein KA, Smith JA, Seidman JG (1984) Nucleotide sequences of the human and mouse atrial natriuretic factor genes. Science 226:1206–1209

Sekiya K, Fushimi M, Hori H, Hirohashi S, Nishimura S, Sugimura T (1984) Molecular cloning and the total nucleotide sequence of the human c-Ha-*ras*-1 gene activated in a melanoma from a Japanese patient. Proc Natl Acad Sci USA 81: 4771–4775

Selby MJ, Barta A, Baxter JD, Bell GI, Eberhardt NL (1984) Analysis of a major human chorionic somatomammotropin gene. J Biol Chem 259:13131–13138

Shen L-P, Rutter WJ (1984) Sequence of the human somatostatin I gene. Science 224:168–171

Simmen RCM, Tanaka T, Ts'ui KF, Putkey JA, Scott MJ (1985) The structural organization of the chicken calmodulin gene. J Biol Chem 260:907–912

Sogawa K, Fujii-Kuriyama Y, Mizukami Y, Ichihara Y, Takahashi K (1983) Primary structure of the human pepsinogen gene. J Biol Chem 258:5306–5311

Sogawa K, Gotoh O, Kawajiri K, Fujii-Kuriyama Y (1984) Distinct organization of methylcholanthrene- and phenobarbital-inducible cytochrome P-450 genes in the rat. Proc Natl Acad Sci USA 81:5066–5070

Sokal RR, Rohlf FJ (1981) Biometry. WH Freeman, New York

Stanton LW, Fahrlander PD, Tesser PM, Marcu KB (1984) Nucleotide sequence comparison of normal and translocated murine c-*myc* genes. Nature 310:423–425

Strein JP, Catterall JF, Kristo P, Means AR, O'Malley BW (1980) Ovomucoid intervening sequences specify functional domains and generate protein polymorphism. Cell 21:681–687

Stone EM, Rothblum KN, Alevy MC, Kuo TM, Schwartz RJ (1985) Complete sequence of the chicken glyceraldehyde-3-phosphate dehydrogenase gene. Proc Natl Acad Sci USA 82: 1628–1632

Sudhof TC, Goldstein JL, Brown MS, Russell DW (1985) The LDL receptor gene: a mosaic of exons shared with different proteins. Science 228:815–822

Swift GH, Craik CS, Stary SJ, Quinto C, Lahaie RG, Rutter WJ, MacDonald RJ (1984) Structure of the two related elastase genes expressed in the rat pancreas. J Biol Chem 259:14271–14278

Takeya T, Hanafusa H (1983) Structure and sequence of the cellular gene homologous to the RSV *src* gene and the mechanism for generating the transforming virus. Cell 32:881–890

Tamkun JW, Schwarzbauer JE, Hynes RO (1984) A single rat fibronectin gene generates three different mRNAs by alternative splicing a complex exon. Proc Natl Acad Sci USA 81: 5140–5144

Valerio D, Duyvesteyn MGC, Dekker BMM, Weeda G, Berkvens TM, van der Voorn L, van Ormondt H, vander Eb AJ (1985) Adenosine deaminase: characterization and expression of a gene with a remarkable promoter. EMBO J 4:437–443

Wang JYJ, Ledley F, Goff S, Lee R, Groner Y, Baltimore D (1984) The mouse c-*abl* locus: molecular cloning and characterization. Cell 36:349–356

Wiedemann LM, Perry RP (1984) Characterization of the expressed gene and several processed pseudogenes for the mouse ribosomal protein L30 gene family. Mol Cell Biol 4:2518–2528

Wieringa B, Hofer E, Weissmann C (1984) A minimum intron length but no specific internal sequence is required for splicing the large rabbit β-globin intron. Cell 37:915–925

Yamada Y, Kuhn K, Crombrugghe BD (1983) A conserved nucleotide sequence, coding for a segment of the C-propeptide, is found at the same location in different collagen genes. Nucleic Acids Res 11:2733–2744

Received October 20, 1986/Revised June 2, 1987