Nucleic Acids Research, 1999, Vol. 27, No. 15 **3219–3228**

Intron–exon structures of eukaryotic model organisms

Michael Deutsch and Manyuan Long*

Department of Ecology and Evolution, The University of Chicago, 1101 East 57th Street, Chicago, IL 60637, USA

Received February 22, 1999; Revised and Accepted June 8, 1999

ABSTRACT

To investigate the distribution of intron–exon structures of eukaryotic genes, we have constructed a general exon database comprising all available intron-containing genes and exon databases from 10 eukaryotic model organisms: *Homo sapiens*, *Mus musculus*, *Gallus gallus*, *Rattus norvegicus*, *Arabidopsis thaliana*, *Zea mays*, *Schizosaccharomyces pombe*, *Aspergillus*, *Caenorhabditis elegans* and *Drosophila*. We purged redundant genes to avoid the possible bias brought about by redundancy in the databases. After discarding those questionable introns that do not contain correct splice sites, the final database contained 17 102 introns, 21 019 exons and 2903 independent or quasi-independent genes. On average, a eukaryotic gene contains 3.7 introns per kb protein coding region. The exon distribution peaks around 30–40 residues and most introns are 40–125 nt long. The variable intron–exon structures of the 10 model organisms reveal two interesting statistical phenomena, which cast light on some previous speculations. (i) Genome size seems to be correlated with total intron length per gene. For example, invertebrate introns are smaller than those of human genes, while yeast introns are shorter than invertebrate introns. However, this correlation is weak, suggesting that other factors besides genome size may also affect intron size. (ii) Introns smaller than 50 nt are significantly less frequent than longer introns, possibly resulting from a minimum intron size requirement for intron splicing.

INTRODUCTION

In order to understand the structure and evolution of genes and genomes in this era of genomics, it is important to know the general statistical characteristics of the intron–exon structures of eukaryotic genes. On the one hand, designing a research project involving genomic structures requires an understanding of general characteristics of genes and genomes. On the other hand, floods of information from exponentially growing databases of DNA and protein sequences often overwhelm researchers who study individual genes or gene families, rendering it difficult to place a newly sequenced gene or a newly determined gene family in a general picture of eukaryotic genes and genomes. When one determines the intron–exon structure of a newly characterized gene, one wonders if it is a normal structure or if it represents an entirely novel structure. Finally, developing sensitive bioinformatics tools to find genes and open reading frames in eukaryotic genome sequences, an important task in genomics studies, also depends on a complete statistical description of intron–exon structures. An updated statistical description of intron–exon structures has been lacking and is imperative for the theoretical study of the origin and evolution of genes and genomes.

It has been a decade since the first compilation of intron–exon structures in eukaryotic genes was published (1). A number of authors published analyses of some characteristics of nuclear introns in a few particular organisms in the late 1980s and early 1990s (2–5). However, the databases have evolved in both size and content in recent years. The first change is the astronomical growth of the sequence databases as a consequence of sequencing the entire genomes of many organisms. The second feature is that more and more redundant genes have entered the databases. For example, the genome of *Saccharomyces cerevisiae* contains 35% proteins from the same gene families (6). How to efficiently define and exclude such genomic redundancy, which may bring bias to the analysis of intron–exon structure, has become a technical challenge. This investigation has considered these new factors in an attempt to portray the general features of gene structures in various model organisms.

We analyzed the statistical distribution of spliceosomal introns and exons of nuclear genes in various model organisms using a DNA sequence database released recently, GenBank 106. These observations, based on a large number of genes (we only chose those model organisms that have many genes sequenced), may be viewed as a general description of gene structures in those organisms. We found from these statistics that, not surprisingly, species have evolved considerably different intron–exon structures. Remarkably, we observed that such changes are correlated with the evolution of genomes and are constrained by functional properties of intron splicing processes. Such correlations bear some implications for some significant issues in gene evolution.

MATERIALS AND METHODS

GenBank sequence database

GenBank release 106 contains sequences of 1.5×10^9 nt in 2.2×10^6 entries. We downloaded all flat files that contain eukaryotic genes, including gbmam.seq, gbinv.seq, bvrt.seq, gbpln.seq, gbrod.seq, gbpri1.seq and gbpri2.seq, to our alpha

*To whom correspondence should be addressed. Tel: +1 773 702 0557; Fax: +1 773 702 9740; Email: mlong@midway.uchicago.edu

All species	Exons				Introns			
	Avg ±	SD	(Min ,	Max)	Avg ±	SD	(Min ,	Max)
Number per gene	4.1 ±	3.5	(2 ,	34)	3.1 ±	3.5	(1 ,	33)
Number per kb cds					3.7 ±	4.2	(1 ,	83)
Size	72.8 ±	116.9	(1 ,	4992)	1113.5 ±	3583.3	(13 ,	65130)
Total size per gene	338.7 ±	317.8	(4 ,	3655)	1466.9 ±	7178.3		
Total size per kb cds					2242.3 ±	12035.9		

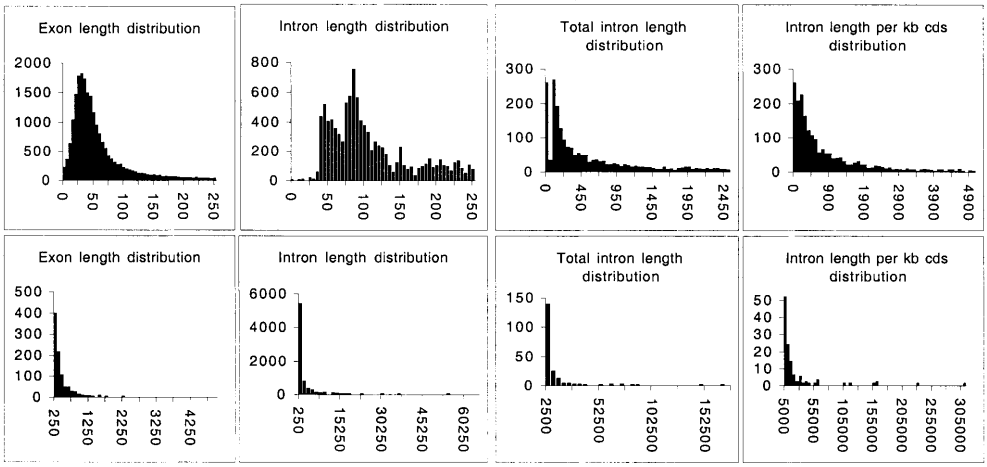

Figure 1. Intron and exon length distributions in the overall database. The distributions of individual intron (intron length distribution) and exon lengths, of the total intron content per gene (total intron length distribution) and of the total intron content per kb of coding sequence are summarized and graphed. Each graph consists of two parts, with smaller lengths above longer ones. Exon lengths are given in amino acids, intron lengths in nucleotides. The horizontal and vertical axes represent lengths and frequencies, respectively.

WDPS 500au workstation (Digital). All further analyses are based on the information stored in these files.

Exon databases

Using the method developed by Long *et al.* (7), we withdrew all entries in the GenBank files that contain intron–exon structures to form a raw intron–exon database. This raw database includes information on locus names, definition of intron–exon structures, species name, protein sequences and DNA sequences. Following the method of Long *et al.* (7), we then calculated all essential parameters, such as the sizes of introns and exons in the regions of the coding sequence (CDS), 3'-UTR and 5'-UTR. In order to avoid errors brought about by erroneous intron submissions, we also collected the dinucleotides around 5' and 3' splice sites as the feature table defined. In the analysis we only used those sequences that had correct GT..AG signals around splice sites within introns. This also deleted a minor class of introns that contain different splice site sequences. The deletion, however, did not change the statistical results significantly, because it only represents a small fraction (<1%) of the total introns (8).

In addition to the overall intron–exon database created from all available sequences, we also created intron–exon databases for the 10 model organisms that have many genes sequenced. These organisms are: *Homo sapiens*, *Mus musculus*, *Gallus gallus*, *Rattus norvegicus*, *Arabidopsis thaliana* (cress), *Zea mays* (corn), *Schizosaccharomyces pombe*, *Aspergillus*, *Caenorhabditis elegans* and *Drosophila*.

Removing redundancy

Many genes now have more than one copy in the database, either orthologous genes from different species or paralogous genes from a gene family in the same species. In some cases, the same genes have been sequenced and reported twice by different laboratories. An extreme case is that there are thousands of immunoglobulin sequences in the database. The uneven distribution of these redundant sequences in the databases will introduce bias into an analysis of intron–exon structures, e.g. the intron number. To avoid this potential bias, we purged the intron–exon databases using the method of Long *et al.* (7). The purging is based on pairwise comparison of protein sequences. When two protein sequences have a similarity greater than or equal to 20%, calculated by fasta3 (9), we keep one sequence and drop the other one in two ways. (i) If we are interested in the number of exons and introns per gene, we compare all the genes in the same gene families as defined by the 20% similarity criterion (all sequences that in comparison have similarity >20% are grouped together and taken as a family). We take a gene with the most common intron and exon numbers as representative of the family. Only the families that contain more than two sequences were considered for the purpose of comparison. (ii) In order to describe intron and exon lengths, we kept the genes that contain the highest intron and exon numbers as representatives of each family. This procedure created the largest unbiased sample of introns and exons from independent or quasi-independent genes.

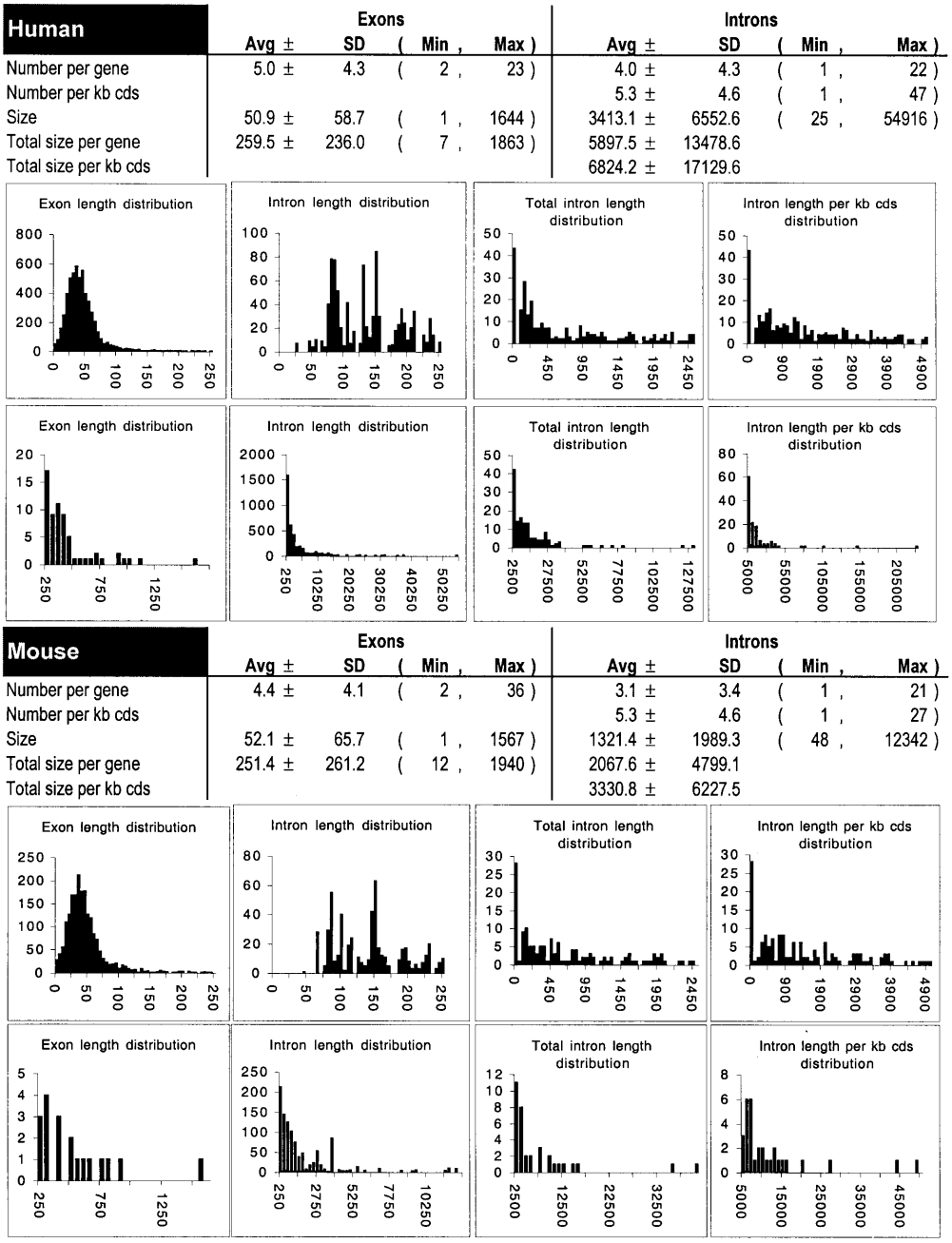

Figure 2. (Above and following pages) Intron and exon length distributions in model organisms and 5'-/3'-UTRs. The distributions of individual intron (intron length distribution) and exon lengths, of the total intron content per gene (total intron length distribution) and of the total intron content per kb of coding sequence are summarized and graphed. Each graph consists of two parts, with smaller lengths above longer ones, except in certain cases where a small sample size made this unnecessary (intron lengths for *Aspergillus*, total intron length and intron length per kb coding sequence for *S.pombe* and total intron length for corn). Exon lengths are given in amino acids, intron lengths in nucleotides. The horizontal and vertical axes represent lengths and frequencies, respectively.

Searching for homologous genes

In order to analyze the distribution of introns in homologous genes across the model organisms, we generated homologous gene families using GBPURGE (7) at a criterion of 30% similarity. In pairwise comparisons, if two sequences had a similarity of 30% or higher, we grouped them into a single family. From each gene family containing sequences of at least five model species, we kept one sequence from each organism, choosing a sequence with the most common number of introns in that species. We then generated databases of homologous genes for each species. We used this data set to analyze the relationship between introns and genome size. We also used the general

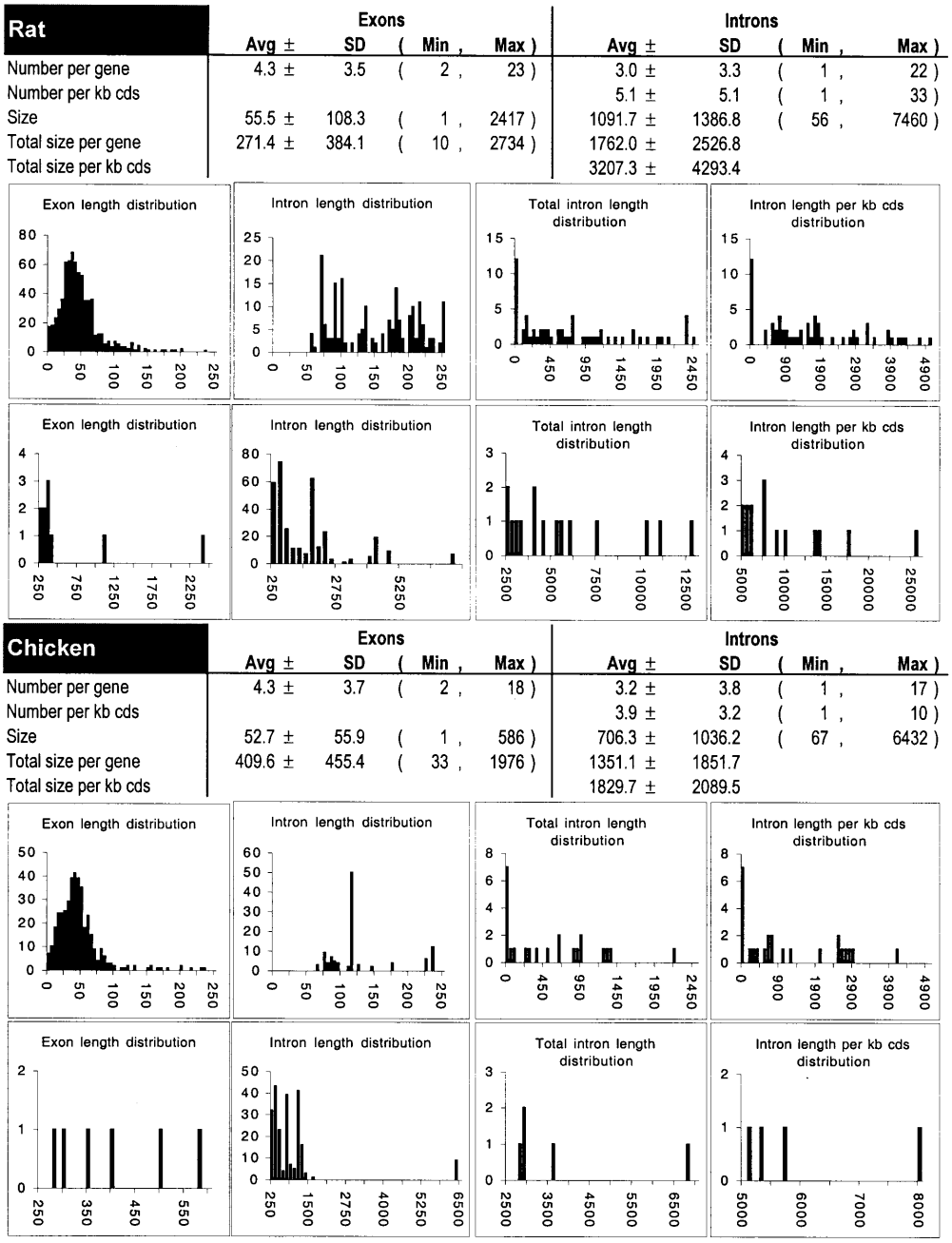

Figure 2. *Continued*.

databases of 10 model species for similar analysis for the purpose of comparison.

RESULTS

Database construction

The original and purged databases are summarized in Table 1. These results show that in the current databases most of the sequences (>70%) are redundant; either paralogous genes from the same gene family (superfamily) or orthologous genes from different species. Purging of these redundant sequences efficiently avoided the bias brought about by redundant sequences.

Intron–exon structures in the overall database

Protein coding region. Figure 1 summarizes the distribution of intron–exon structures in protein coding regions for the overall database. This distribution gives a clear picture of the eukaryotic genes. An average gene contains 3.7 introns in 1 kb of protein coding region, but with considerable variation: a gigantic gene, human collagen type VII (13), contains 117 introns; the Fugu

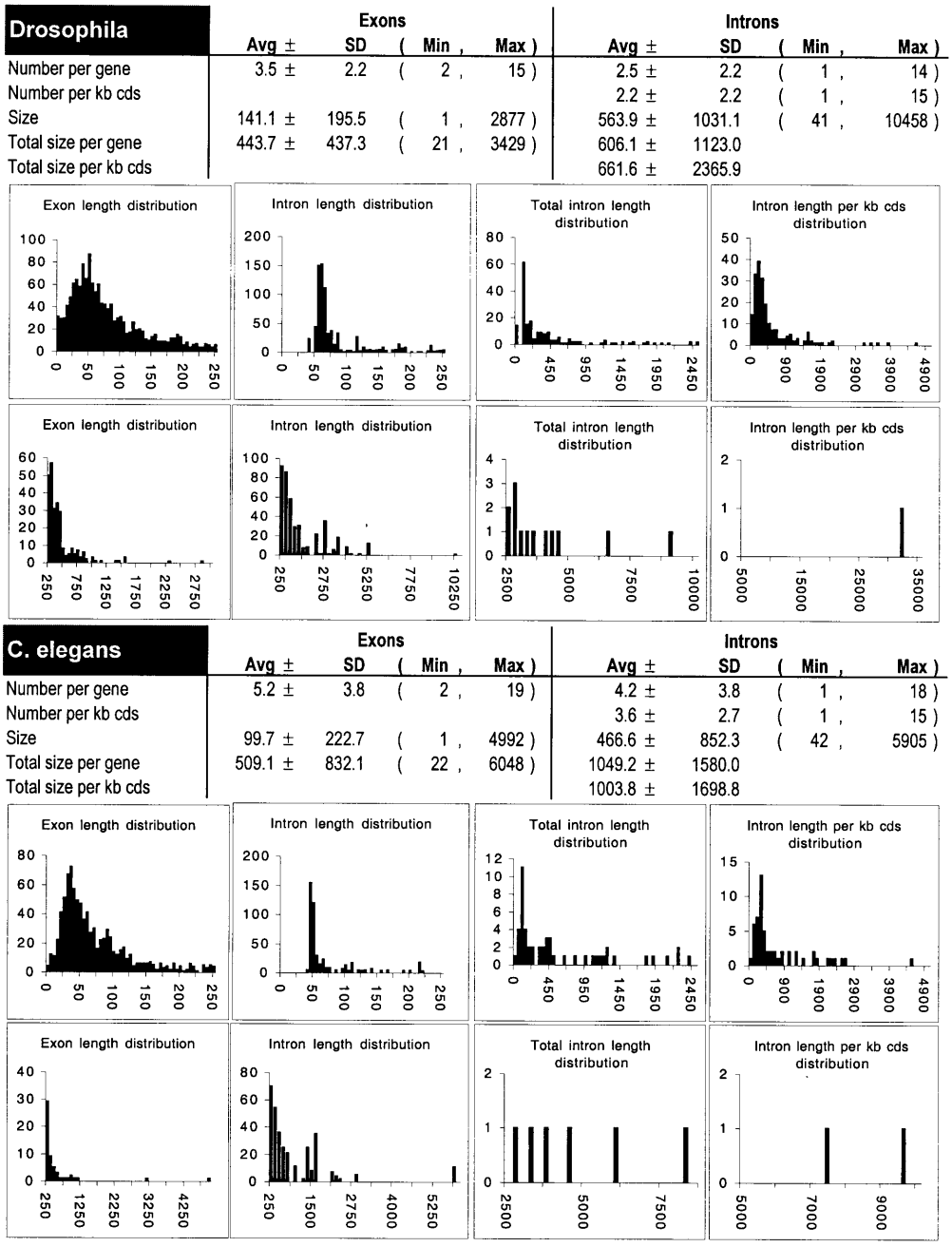

Figure 2. *Continued*.

fish gene homologous to the Huntington's disease gene contains 66 introns (22).

Figure 1 shows that exon lengths are distributed much more tightly than intron lengths. Most exons are 30–40 residues long, which is consistent with previous observations on smaller samples (2,7). A common intron is 40–125 nt long, however, this statistic shows huge variation (in the database the largest recorded is 108 kb; human gene GenBank accession no. AC003992). The longest introns, although not in the database, are >300 kb in the dystrophin gene (10), which contains >700 kb of intron sequences (5). Human gene CIT987-SKA-34504 (M. D. Adams *et al*., GenBank accession no. AC002302) contains introns of 151 kb. The smallest introns were 18–20 nt long in the nucleomorph, a eukaryotic endosymbiont (11), and 21 nt in *Paramecium tertaurelia*, a ciliated protozoan (12).

UTR regions. In the database, 2% of genes contain descriptions of introns and exons in the 5'- and 3'-UTRs of the RNA (Fig. 2). Seventy-four genes have 5'-UTR sequences, with seven genes having one 5'-UTR intron. Sixty-nine genes have

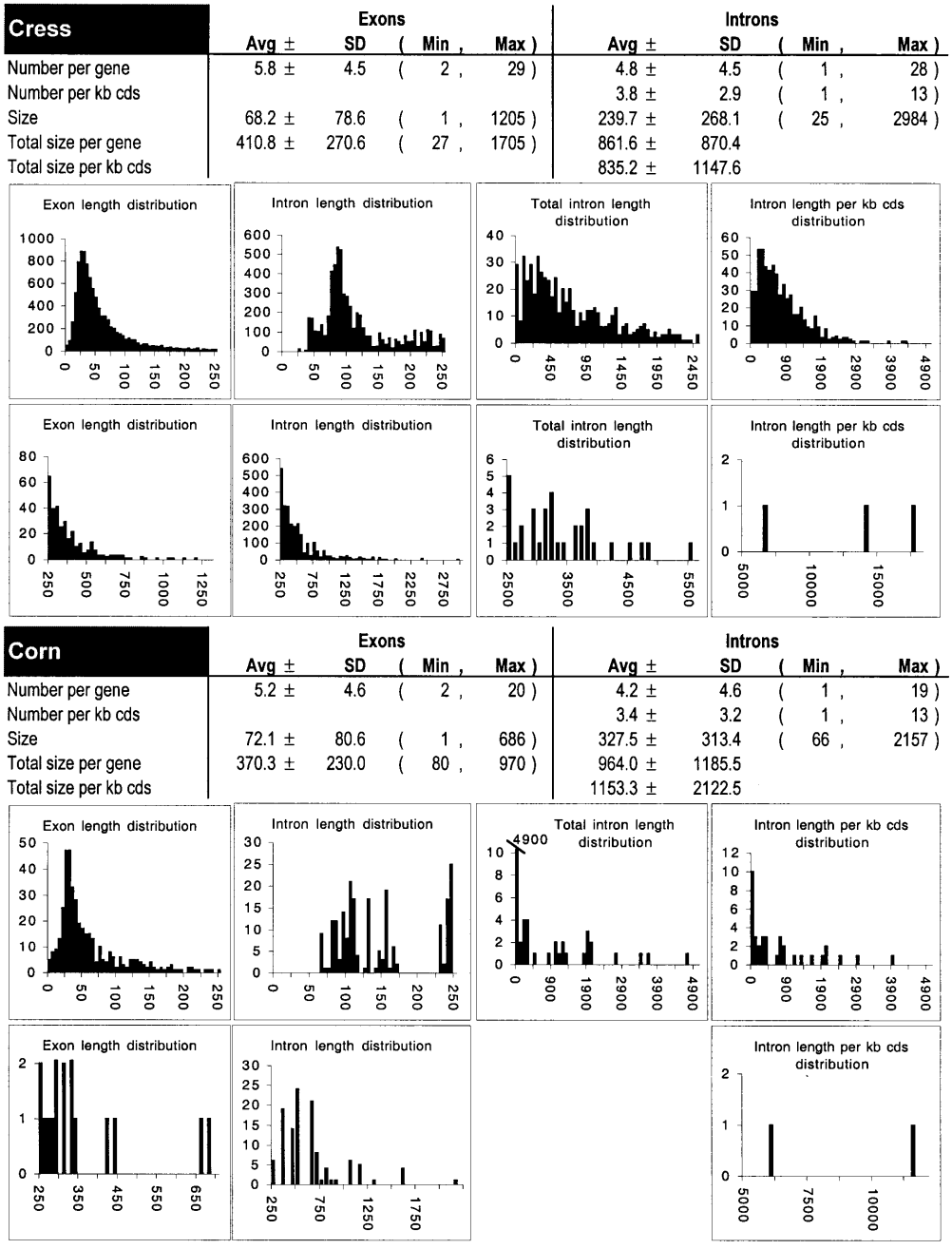

Figure 2. *Continued.*

a single 3'-UTR exon. The lengths of these exons and introns show a variable distribution. For instance, the average 3'-UTR sequence is 340 nt long, with a minimum of 17 nt and maximum of 1376 nt; the lengths of the seven 5'-UTR introns range from 96 to 8214 nt.

Intron–exon structures in the 10 model organisms

Figure 2 shows the intron–exon structures of the 10 model organisms. Like the overall database, these organisms show a tighter distribution of exon lengths than of intron lengths, as well as minimum intron lengths.

Human genes have the most introns and the largest introns of the 10 species. An average human gene has four introns; the highest recorded number is 116 introns, in the collagen type VII gene (13). The mean intron length is 3413 bases while the most common length is 75–150 nt. The longest introns are recorded in the BSC gene (its first intron is >71 kb) and the dystrophin gene (several introns >100 kb). The intron length in

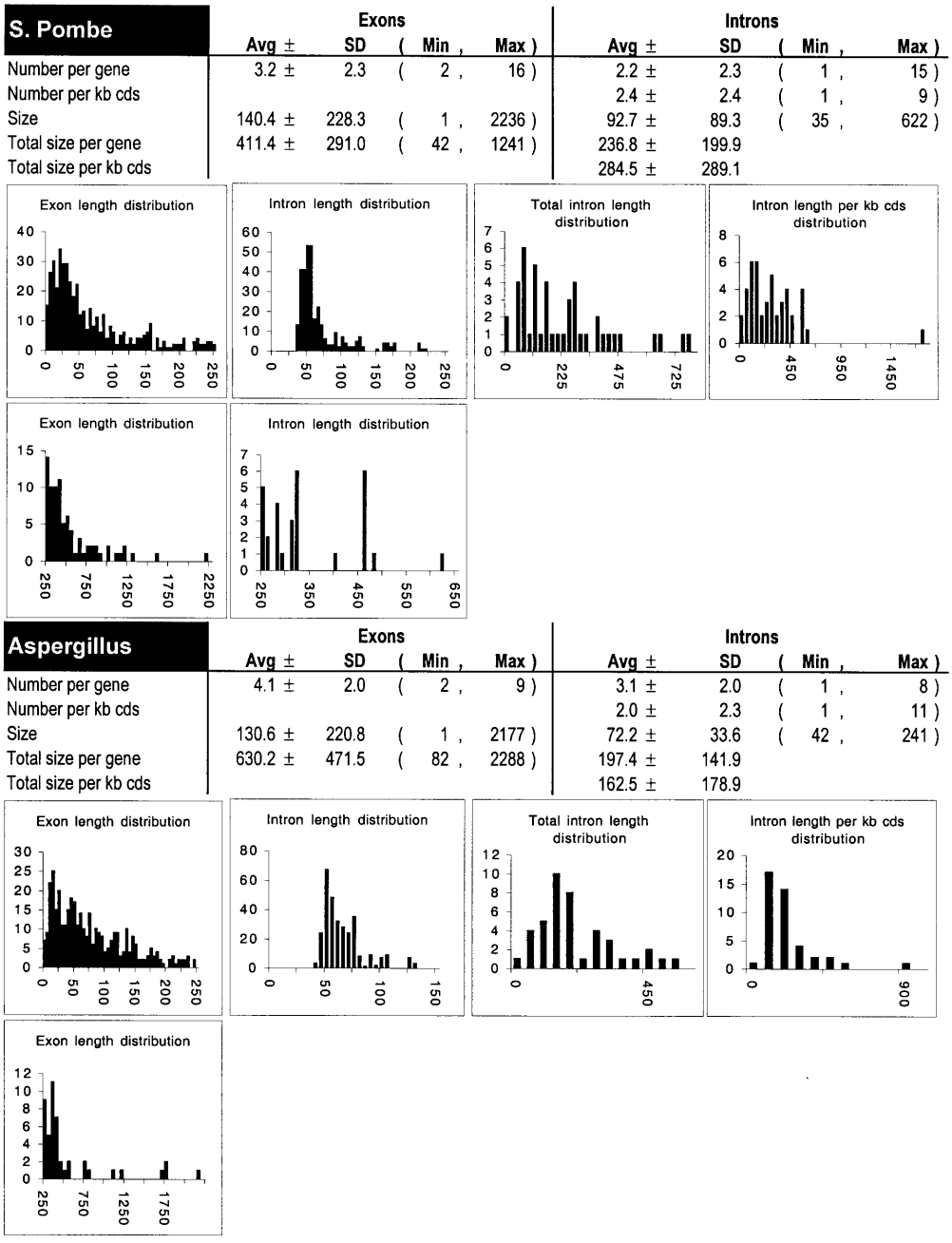

Figure 2. *Continued.*

1 kb of human gene CDS is close to 7 kb (6825 nt) on average. The other two mammalian species (mouse and rat) are similar to humans in the numbers of introns per kb of CDS. Their individual intron lengths and total intron length per kb of CDS or per gene are shorter than human but higher than other non-mammalian species. It appears that these two rodent species have shorter proteins.

The two fungi, *S.pombe* and *Aspergillus*, have the shortest introns. Their average gene contains only two introns per kb of CDS, totalling 160–280 nt. Individual introns on average are only 40–75 nt long. This is very similar to the case of *S.cerevisiae*, the first eukaryotic species whose complete genome was sequenced, which was found to have very few, short introns (14).

Chicken genes contain more and larger introns, next to the mammalian genes. A slightly higher number of introns per gene is compensated for by shorter introns (700 nt on average), giving a total of 1.8 kb of intron sequence per kb of CDS.

Caenorhabditis elegans and *Drosophila*, two invertebrates, do not contain long introns. Their total intron lengths per kb of

CDS average only 1000 and 600 nt, respectively. However, these two species have contrasting intron–exon structures. *Caenorhabditis elegans* genes contain more (4.0 introns per kb of CDS), shorter (467 nt each) introns, while *Drosophila* genes have fewer, longer introns (2.7 introns per kb of CDS, 564 nt each).

Table 1. Number of genes and introns before and after purging

Organism	Unpurged		Purged (more introns)		Purged (most common)	
	Genes	Introns	Genes	Introns	Genes	Introns
Overall database	16,989	58,973	2,903	17,102	2,000	5,604
Human	2,554	11,212	582	4,645	404	1,559
Mouse	1,183	4,151	267	1,563	174	507
Rat	405	1,820	119	519	76	215
Chicken	160	646	60	335	29	84
Drosophila	1,830	4,361	355	1,159	202	475
C. elegans	276	1,356	121	785	57	241
Cress	2,894	14,732	1,253	8,257	589	2,791
Corn	162	618	71	328	38	149
S. pombe	236	561	143	355	45	96
Aspergillus	260	741	95	320	42	128

The genes of the two plant species, corn and cress, contain introns whose lengths per kb of CDS are intermediate, like the two invertebrates. The number of introns is similar in the two plant species (3.9–4.3 per kb of CDS), but the average corn intron (328 nt) is longer than the average cress intron (240 nt).

Correlation between intron size and genome size

In our analysis, we observed a correlation between the size of genomes and the amount of intron sequences in their genes. Table 2 and Figure 3 show correlations between genome size and the number of introns per gene, number of introns per kb of CDS, total intron length per gene and intron length per kb of CDS. The correlation between genome size and total intron length per kb CDS is significant at a marginal level ($P = 0.06$ for $R = 0.6$). This weak correlation may suggest a limited causal relationship between intron content and genome complexity. However, it also indicates that other factors are likely to be involved in the evolution of intron size.

Figure 3 reveals a possible relationship between genome size and intron–exon structure. The calculation is, however, based on the purged databases, in which different species may be represented by different types of genes. For example, the plant databases contain photosynthesis-related genes, while the human database contains immunoglobulin genes. Different types of genes may have different intron–exon structures. A more rigorous comparison requires homologous genes in the different organisms. We obtained sequences from 55 gene families that were represented in at least five of the 10 species. Figure 3 shows the relationship between intron content and genome size in these homologous genes. These results show an elevated correlation between intron content and genome size. The higher correlation values in homologous gene sets indicate a better control of the factors due to non-homologous genes in the analyses of Figure 3.

DISCUSSION

This analysis provides a general picture of the intron–exon structure of eukaryotic genes. On average, the analyzed genes have 3.7 introns of 40–150 nt each. These statistics are subject to large variation. A number of introns >100 kb or <20 nt exist in the databases and the literature. A human gene contains more than 100 introns, while some genes in the Fugu fish have more than 60 introns. The intron–exon structures of different organisms are variable. These statistical analyses, in general, may provide a fundamental basis for both understanding the structure of a gene that is identified in molecular studies and developing more sensitive tools to find genes or open reading frames in eukaryotic genome sequences. In particular, this investigation also suggests two interesting points, which may increase our understanding of the evolution of intron–exon structures.

The first is that although introns can be very long, the minimum intron size is limited by the length of the splicing signals. Most of the shortest introns observed were 20–30

Table 2. Correlation between intron size (means) and genome size

Species	Genome size (Mbp)	Number of introns per gene		Size of individual introns		Total intron size per kb cds		Sample size	
		All genes	Homologous genes	All genes	Homologous genes	All genes	Homologous genes	# genes	# introns
Human	3400	4.0	5.54	3413.4	1152.4	6824.6	5001.7	50	257
Mouse	3454	3.1	6.68	1321.4	666.1	3331.2	3260.3	32	159
Rat	2900	3.0	3.77	1091.7	566.8	3207.7	2998.8	31	115
Chicken	1200	3.2	3.69	706.3	329.1	1830.1	1920.6	23	79
Drosophila	180	2.5	2.44	563.9	445.3	662.1	779.2	36	73
C. elegans	100	4.2	4.26	466.6	280.7	1004.2	1033.4	50	213
Cress	100	4.8	4.10	239.7	156.9	835.7	733.9	39	143
Corn	5000	4.2	4.09	327.5	270.2	1153.7	872.4	11	35
S. pombe	14	2.2	2.39	92.7	104.0	285.0	256.5	23	54
Aspergillus	13	3.1	5.27	72.2	73.1	162.9	379.2	11	69
R (correlation)		0.22	0.45	0.50	0.57	0.60	0.60		

The correlation coefficients (R) are calculated for the correlation between genome size and the measure of intron size indicated at the top of each column. Statistics from the purged database containing genes with the most common number of introns. Sample sizes given are for homologous genes.

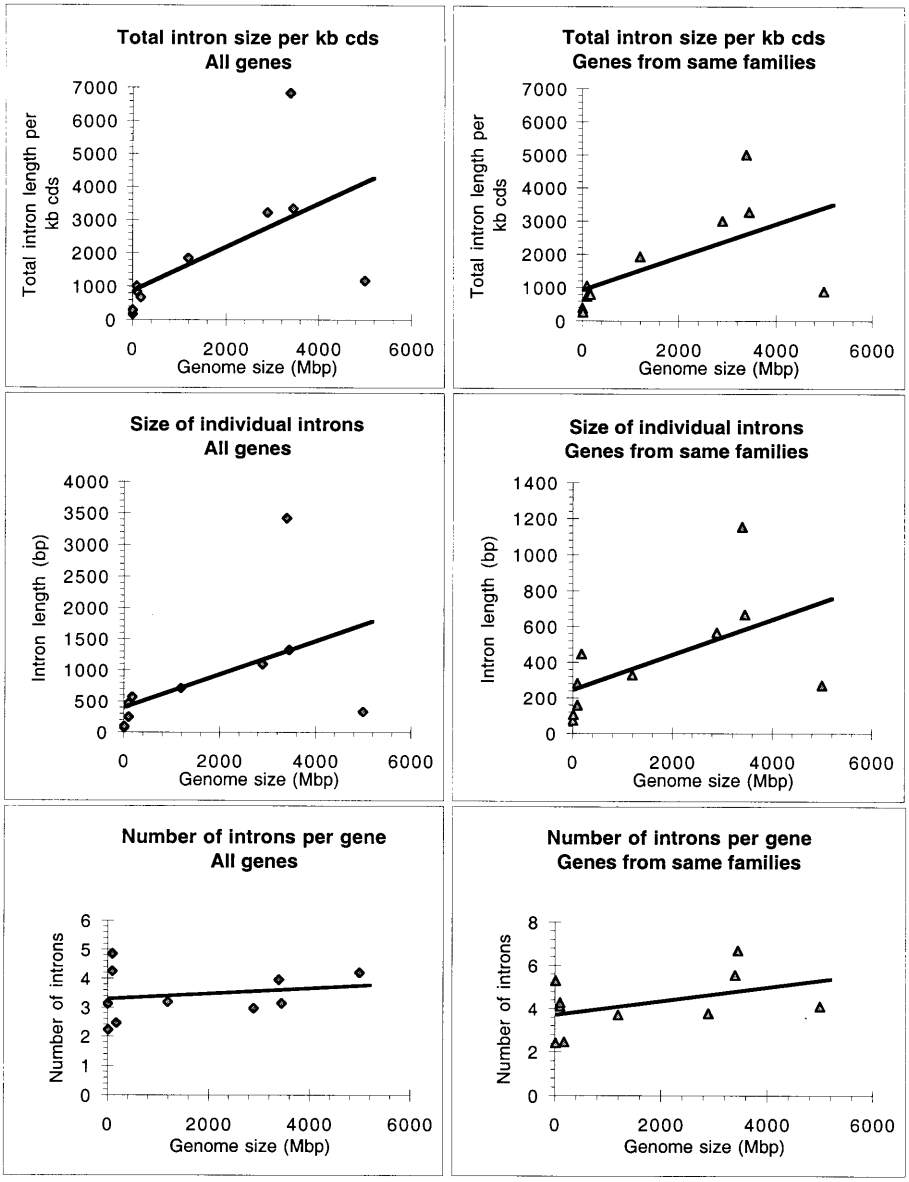

Figure 3. Correlation between genome size and average intron size/number in model organisms. Figures labeled 'All genes' contain data from all the genes for each of the model organisms; figures labeled 'Genes from same families' contain data only from genes present in at least five of the model organisms.

bases long; very few were <20 bases. This indicates that in order to encode adequate splicing signals, introns cannot be too short. In fact, conservation analysis of the splice sites showed that the conserved sequence distribution in introns can be extended over >20 nt (14). The smallest recorded introns, found in protist genes, are 13–20 bases long (12). They also encode the minimum splice sequences (GT..AG dinucleotide sequences are encoded), supporting the general conclusion of this analysis.

Second, it has been speculated that intron content is correlated with genome size. For example, it was proposed that intronless prokaryotic genes might be a product of selection against introns for more efficent molecular processes of replication, transcription and processing (16). Recently, it was observed that small genome sizes in *Drosophila* were correlated with high deletion rates in the Helena retrosequence and introns (17–19). The correlations we have observed in the genomes of model organisms support, in general, the notion that the existence of non-functional elements should be consistent with the size of the genome. The small introns in bird genes provide another example of this relationship (20). However, the correlations as revealed in this study are only at a marginal level of significance in the total intron size per kb CDS, although the correlations increase in homologous genes. These analyses indicate a possibly true but weak connection between genome size and intron sizes, suggesting that there may be other factors involved as well.

This is the distribution of a large sample, with 2903 independent or quasi-independent genes and 17 102 introns. Cautionary notes, however, should be made. First, these statistics were calculated only from genes that contain introns, so information such as the number of introns per gene is only valid for these intron-containing genes. A complete survey of genes that do not contain introns is so far unavailable, except in the case of yeast (14,21). Second, when introns are very long, many researchers tend not to sequence them, for understandable reasons. As a result, the average intron sizes are underestimated to some extent. However, the smooth and fast decrease in frequency of intron sizes in Figure 1 implies that very large introns may make up a small proportion (<5%) of introns in the genome, although the final statistic awaits the completion of the human genome project. Finally, it is not unreasonable to believe that the mode of intron size distribution is likely a stable measurement of most introns.

ACKNOWLEDGEMENTS

We thank Carl Rosenberg at Falling Rain Genomics Inc. (Lincoln, MA) for developing the GBPURGE package which made automatic purging of gene redundancy possible. We thank Walter Gilbert for discussions. The laboratory of M.L. was supported by the Packard Fellowship in Science and Engineering and the National Science Foundation.

REFERENCES

1. Hawkins,J.D. (1988) *Nucleic Acids Res.*, **16**, 9893–9906.
2. Dorit,R.L., Schoenbach,L. and Gilbert,W. (1990) *Science*, **250**, 1377–1382.
3. Palmer,J.D. and Logsdon,J.M.,Jr (1991) *Curr. Opin. Genet. Dev.*, **1**, 470–477.
4. Mount,S.M., Burks,C., Hertz,G., Stormo,G.D., White,O. and Fields,C. (1992) *Nucleic Acids Res.*, **20**, 4255–4262.
5. Fedorov,A., Suboch,G., Bujakov,M. and Fedorova,L. (1992) *Nucleic Acids Res.*, **20**, 2553–2557.
6. Das,S., Yu,L., Gaitatzes,C., Rogers,R., Freeman,J., Bienkowska,J., Adams,R.M., Smith,T.F. and Lindelien,J. (1997) *Nature*, **385**, 29–30.
7. Long,M., Rosenberg,C. and Gilbert,W. (1995) *Proc. Natl Acad. Sci. USA*, **92**, 12495–12499.
8. Sharp,P.A. (1997) *Cell*, **91**, 875–879.
9. Pearson,W.R. (1994) *Methods Mol. Biol.*, **24**, 307–331.
10. Boyce,F.M., Beggs,A.H., Feener,C. and Kunkel,L.M. (1991) *Proc. Natl Acad. Sci. USA*, **88**, 1276–1280.
11. Gilson,P.R. and McFadden,G.I. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 7737–7742.
12. Russell,C.B, Fraga,D. and Hinrichsen,R.D. (1994) *Nucleic Acids Res.*, **22**, 1221–1225.
13. Christiano,A.M., Hoffman,G.G., Chung-Honet,L.C., Lee,S., Cheng,W., Uitto,J. and Greenspan,D.S. (1994) *Genomics*, **21**, 169–179.
14. Long,M., De Souza,J.S. and Gilbert,W. (1997) *Cell*, **91**, 739–740.
15. Long,M., de Souza,S.J., Rosenberg,C. and Gilbert,W. (1998) *Proc. Natl Acad. Sci. USA*, **95**, 219–223.
16. Doolittle,W.F. (1990) In Stone,E.M. and Schwartz,R.J. (eds), *Intervening Sequences in Evolution and Development*. Oxford University Press, Oxford, UK, pp. 43–62.
17. Petrov,D.A. and Hartl,D.L. (1998) *Mol. Biol. Evol.*, **15**, 293–302.
18. Petrov,D.A., Lozovskaya,E.R. and Hartl,D.L. (1996) *Nature*, **384**, 346–349.
19. Moriyama,E.N., Petrov,D.A. and Hartl,D.L. (1998) *Mol. Biol. Evol.*, **15**, 770–773.
20. Hughes,A.L. and Hughes,M.K. (1995) *Nature*, **377**, 391.
21. Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M., Louis,E.J., Mewes,H.W., Murakami,Y., Philippsen,P., Tettelin,H. and Oliver,S.G. (1996) *Science*, **274**, 546–567.
22. Baxendale,S., Abdulla,S., Elgar,G., Buck,D., Berks,M., Micklem,G., Durbin,R., Bates,G., Brenner,S., Beck,S. and Lehrach,H. (1995) *Nature Genet.*, **10**, 67–76.