Front. Biol. China (2006) 4: 362–366
DOI 10.1007/s11515-006-0047-2

RESEARCH ARTICLE

Statistical characteristics of eukaryotic intron database

HE Miao[1], LI Jidong[1], ZHANG Shanghong (✉)[2]
1. School of Life Sciences, Sun Yat-Sen University, Guangzhou 510275, China
2. The Key Laboratory of Gene Engineering of Ministry of Education, and Biotechnology Research Center, Sun Yat-Sen University, Guangzhou 510275, China

Abstract A database called eukaryotic intron database (EID) was developed based on the data from GenBank. Studies on the statistical characteristics of EID show that there were 103, 848 genes, 478,484 introns, and 582,332 exons, with an average of 4.61 introns and 5.61 exons per gene. Introns of 40–120 nt in length were abundant in the database. Results of the statistical analysis on the data from nine model species showed that in eukaryotes, higher species do not necessarily have more introns or exons in a gene than lower species. Furthermore, characteristics of EID, such as intron phase, distribution of different splice sites, and the relationship between genome size and intron proportion or intron density, have been studied.

Keywords eukaryote, intron, database, statistical characteristics

1 Introduction

Introns are an important kind of noncoding sequences. They are widely distributed in eukaryotic genes, usually representing a large part of the length, even up to 90% of a gene (Doolittle, 1987). In addition, introns exist in organelle genomes as well as archaeal and bacterial genomes (Zhang, 1998). The origin, evolution, and function of introns have been a major theoretical problem in the study on genes and genomes. There are many debates on this issue.

Currently, one opinion considers that introns would be byproducts of the evolution of gene and genome. This view regards their origin as "intron-late," which means introns would be derived from existing genomic sequences after the appearance of functional genes or eukaryotes (Doolittle, 1987). Meanwhile, the other opinion considers that introns would play an important role in the origin and evolution of primitive genes and genomes, and that they would also have important functions in the regulation of modern genes and genomes. Therefore, this view regards their origin as "intron-early," meaning introns would exist in primitive genomes (Doolittle, 1987; Zhang, 1998).

As for the functions of introns, some examples indicate that they may play various roles: acting as a promoter (Tee et al., 1995), an enhancer (Lou et al., 1996), or a *trans*-acting factor (Mattick, 1994), involving in RNA editing (Herbert, 1996), etc. In addition, many small nucleolar RNAs (snoRNAs) are encoded by introns (Maxwell and Fournier, 1995). These findings would certainly shed light on further understanding of the origin and functions of introns.

Bioinformatics makes it possible for us to gather information about introns systemically, and to analyze it statistically. This practice is very useful for further study on introns (Sakharkar et al., 2000). In this paper we analyzed the statistical characteristics of introns in the eukaryotic intron database (EID) we developed previously (He et al., 2004), including the numbers, length distribution, density, phases, and splicing sites of introns. The analysis would provide some useful information for the study on the origin and evolution of introns.

2 Development of the eukaryotic intron database and analysis methods

2.1 Data source

We downloaded all EID data and annotations of sequences from GenBank (Release 125). The sequence data came from many species including invertebrates, primates, rodents, other mammals, other vertebrates, and plants.

Translated from *Acta Scientiarum Naturalium Universitatis Sunyatseni*, 2005, 44(6): 79–82 [译自: 中山大学学报, 2005, 44(6): 79–82]

E-mail: lsszsh@zsu.edu.cn

2.2 Structure of the eukaryotic intron database

EID is composed of one main database and several subdatabases. The main database includes three parts: dEID, pEID, and hEID, representing the three text-file databases, DNA-EID, protein-EID, and header-EID, respectively. All of them were in FASTA format. The subdatabases included nuclear-mRNA intron database, organelle intron database, and those for introns in primates, rodents, other mammals, other vertebrates, invertebrates, or plants.

2.3 Statistical methods

Programs for retrieving the text files according to the common file format in all databases and for statistical analysis were written in PERL. Statistical characteristics, such as gene numbers, intron numbers, exon numbers, and other intron parameters, were obtained from the analysis. We employed the GD: GRAPH module of PERL to visualize the statistical results.

3 Statistical results from the eukaryotic intron database

3.1 Basic statistical results

There are 103,848 genes in the database. These genes were composed of a total number of 478,484 introns and 582,332 exons. On average, there were 4.61 introns and 5.61 exons per gene. The numbers of introns in different genes from different species or from the same species might vary greatly. For example, the gene encoding type VII collagen in humans had 117 introns.

The distribution pattern of the lengths (sizes) of introns and exons in the database are shown in Fig. 1. Exons with lengths of 90–120 nucleotides had the highest frequencies. This result was consistent with that obtained by Dorit et al (1990) with an earlier version of GenBank. Introns with lengths of 40–120 nucleotides had the highest frequencies. The longest intron resided in the zinc-finger protein gene (AF178030) of human chromosome 8 with a length of 152 kb. The shortest intron was only 8 nucleotides in length. It was found in a ciliate. Compared with the distribution of the lengths of exons, the distribution of the lengths of introns showed two peaks: 50 nucleotides and 90 nucleotides, respectively.

In addition, statistical results of the subdatabases for each of the nine model species (*Homo sapiens, Rattus norvegicus, Mus musculus, Gallus gallus, Drosophila melanogaster, Caenorhabditis elegans, Arabidopsis thaliana, Zea mays, and Schizosaccharomyces pombe*) are shown in Table 1.

In eukaryotes, higher species did not necessarily have more introns or exons in a gene than lower species. The average numbers of introns or exons per gene did not vary considerably with species, except that *S. pombe* had smaller values. The average numbers of introns or exons per gene were the largest in *H. sapiens*: 5.77 and 6.77, respectively.

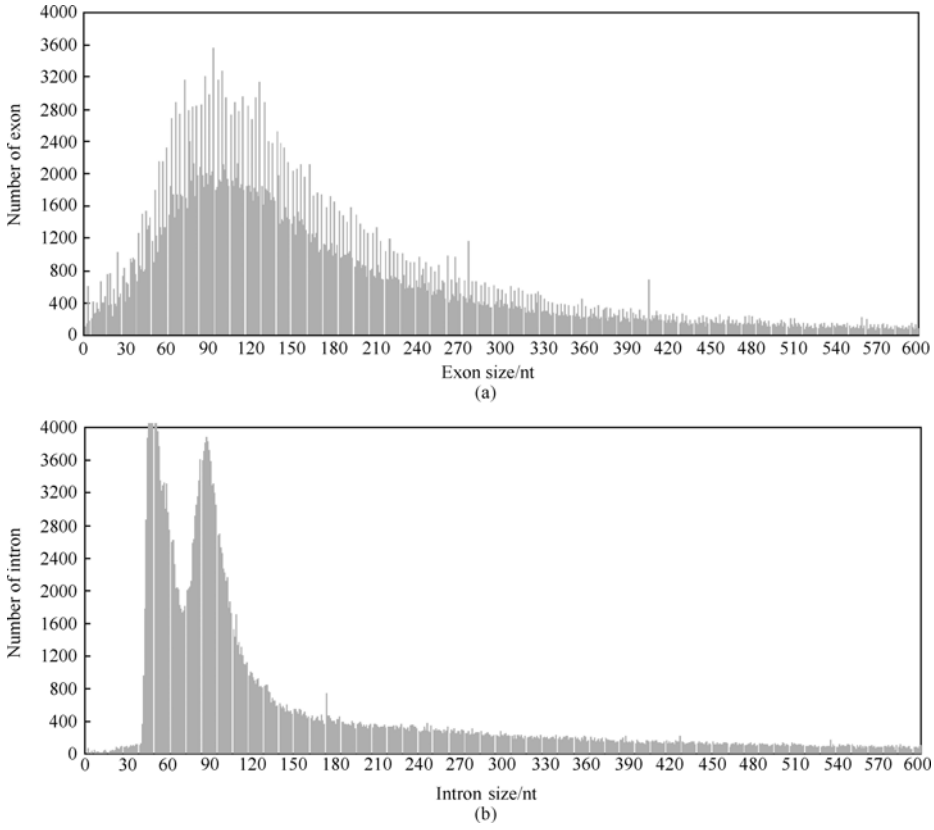

Fig. 1 Distribution of exon size (a) and intron size (b) in eukaryotic intron database (EID)

Table 1 Statistical results of nine model species in eukaryotic intron database (EID)

Database or subdatabase	Number of genes	Number of introns	Number of exons	Introns per gene	Exons per gene
Main database	103 848	478 484	582 332	4.61	5.61
Homo sapiens	16 524	95 379	111 903	5.77	6.77
Rattus norvegicus	881	4 021	4 902	4.56	5.56
Mus musculus	3 296	17 310	20 606	5.25	6.25
Gallus gallus	442	2 133	2 575	4.83	5.83
Drosophila melanogaster	10 961	36 730	48 054	3.35	4.38
Caenorhabditis elegans	21 064	116 548	139 216	5.53	6.61
Arabidopsis thaliana	24 596	129 650	154 428	5.27	6.18
Zea mays	163	741	910	4.55	5.58
Schizosaccharomyces pombe	156	366	524	2.35	3.36

As shown in Fig. 1b, introns with a length of less than 20 nt were rare. Small introns were usually 20–30 nt in length. There was no such constraint on exons. The smallest exons were only several nucleotides in length. The reason for this would be that a sequence of 20–30 nt was required for intron recognition and for providing the splicing signals.

3.2 Relationship between genome sizes and introns

The relationship between the genome sizes of the nine model species studied and their intron proportions (total intron length in kb per kb CDS) or intron densities (total intron number per kb CDS) is shown in Table 2. These results were obtained based on genome size data and our corresponding subdatabases.

As shown in Table 2, the intron proportion in *H. sapiens* was as large as 8.43, followed by those in *M. musculus* and *R. norvegicus*. The intron proportion in *G. gallus* was smaller than those of the above three mammals. The intron proportions in the two invertebrates were similar, and both of them are small. The intron proportions in the two plant species are also small. *S. pombe* has the smallest intron proportion (only 0.14). As for intron densities, vertebrates usually had high densities, while invertebrates had relatively low densities. Although the intron proportions in plants were small, their intron densities were not low. Similar to the intron proportion, the intron density in *S. pombe* was the lowest of the nine model species.

The total nucleotides in the databases were highly correlated with the total intron lengths because the sequences in the databases were all gene sequences. Therefore, the more the sequences, the longer the total intron length. The correlation between genome sizes and the corresponding intron densities or intron proportions was not so high, but there was still a tendency that intron densities and intron proportions increased with genome sizes. On the other hand, it seemed that intron densities varied more in eukaryotes with relatively small genomes, and that intron proportions varied more in eukaryotes with relatively large genomes.

3.3 Statistical analysis on intron phases

Intron phases referred to the positions of an intron relative to the three nucleotides of a codon in a gene. It could be seen from Table 3 that the distribution of intron phases in each model species was not random. Introns with phase 0 accounted for about half of all introns, significantly more than introns with phase 1 or phase 2. This result was consistent with that obtained by Fedorov et al (2002), with an earlier version of GenBank.

Table 2 Relationship between genome sizes and introns

Species (Subdatabase)	Total nucleotide in database /kb	Total intron length /kb	Total exon length /kb	Genome size /Mb	Intron proportion	Intron density
Homo sapiens	166 413	148 769	17 653	3 165	8.43	5.40
Rattus norvegicus	2 405	1 658	747	2 700	2.22	5.38
Mus musculus	13 879	10 550	3 329	3 000	3.17	5.20
Gallus gallus	1 137	715	422	1 200	1.69	5.05
Drosophila melanogaster	37 405	19 489	17 916	180	1.09	2.05
Caenorhabditis elegans	59 247	29 951	29 296	97	1.02	3.98
Arabidopsis thaliana	56 130	21 975	34 155	120	0.64	3.80
Zea mays	384	203	181	3 300	1.12	4.09
Schizosaccharomyces pombe	251	31	220	12.6	0.14	1.66

According to the "intron-late" hypothesis, introns were inserted into genes during genome evolution. Therefore, the distribution of the three intron phases should be random, i.e., the proportions of each phase should be approximately equal. According to the "intron-early" hypothesis, the most primitive introns and exons were derived from early microgenes (Gilbert, 1987). Consequently, the phase of all primitive introns should be 0. Introns with other phases came up later in the course of genome evolution. The observed results were what the "intron-early" scenario expected.

Table 3 Statistical results of intron phases

Database or subdatabase	Phase 0 /%	Phase 1 /%	Phase 2 /%
Main database	49.1	24.3	26.6
Homo sapiens	52.2	25.1	22.7
Rattus norvegicus	54.1	20.4	25.5
Mus musculus	49.4	25.3	25.3
Gallus gallus	51.2	25.1	23.7
Drosophila melanogaster	46.2	25.7	27.9
Caenorhabditis elegans	51.4	23.5	25.1
Arabidopsis thaliana	53.3	24.2	22.5
Zea mays	50.8	23.4	25.8
Schizosaccharomyces pombe	45.4	26.4	28.2

3.4 Statistical analysis on intron splice sites

The most frequent splice site of spliceosomal introns was GT…AG (canonical splice site). Another kind of splice site of spliceosomal introns was AT...AC (Hall and Padgett, 1996). There was a distribution pattern for the splice sites GT…AG, GC…AG (probably derived from GT…AG), and AT...AC. The proportions of every kind of splice sites in the main database are shown in Table 4.

Table 4 Distribution of different splice sites in the main database

Splice site	Intron number	Proportion /%
GT…AG	465 086	97.200
NN…NN[a]	10 126	2.116
GC…AG	2 140	0.447
GT…NN	343	0.072
NN…AG	274	0.057
TA…GG	162	0.034
AT…AC	48	0.010
AT…AT	35	0.007

[a]: N represents unknown nucleotide

In addition, statistical analysis on the subdatabases indicated that the proportions of canonical splice site were over 95% in all the nine model species studied. Besides the canonical splice site, the proportions of unknown splice sites (NN…NN, GT…NN, and NN…AG) were not negligible. This might be due to annotation errors related to the splice sites.

4 Discussion

It may be concluded from the present study that with highest frequencies, an exon encodes 30–40 amino acids and that introns with phase 0 are also with the highest frequencies, whether in higher eukaryotes or lower eukaryotes. This result supports in a way the "intron-early" hypothesis. Of course, the conclusion does not preclude the possibility that some introns originated late in the course of genome evolution.

Our results also show that intron content tends to increase with eukaryotic genome sizes. This is what the "large genome" evolutionary scenario anticipates (Zhang, 1998). In such an evolutionary scenario, the amounts of other noncoding sequences, such as repeated sequences, would also increase besides intron content (Zhang, 1998). Related study also indicates that the amounts of extragenic noncoding sequences would increase with genome sizes more rapidly than intron content (Vinogradov, 1999).

It is significant to develop new algorithms and software to analyze the data in the intron databases, to undertake further study on the origin and evolution of introns. On one hand, a study of the statistical characteristics of intron phases and their relation with the conservation of exon-intron junctions is required. This study is useful for the understanding of the correlation both in structure and in function, and the origin, of exon and intron boundary sequences. On the other hand, an understanding of the information about intron-related repeated sequences, genes in introns, and conserved regions in introns is required. This practice will reinforce a profound study on the evolutionary relationship between introns and other genomic sequences, such as the possible evolutionary relationship between introns and repeated sequences (Zhang, 1998).

Acknowledgements This work was supported by the National Natural Science Foundation of China (No. 30270752) and the Guangdong Natural Science Foundation (No. 031616).

References

Doolittle W. F., The origin and function of intervening sequences in DNA: a review. Am. Nat., 1987, 130(6): 915–928

Dorit R. L., Schoenbach L., Gilbert W., How big is the universe of exons? Science, 1990, 250(4986): 1377–1382

Fedorov A., Merican A. F., Gilbert W., Large-scale comparison of intron positions among animal, plant, and fungal genes. Proc. Natl. Acad. Sci. USA, 2002, 99(25): 16128–16133

Gilbert W., The exon theory of genes. Cold Spring Harb. Symp. Quant. Biol., 1987, 52: 901–905

Hall S. L. and Padgett R. A., Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre-mRNA introns. Science, 1996, 271(5256): 1716–1718

He M., Li J. D., Zhang S. H., Development of Eukaryotic Intron Database (EID). Acta Scientiarum Naturalium Universitatis Sunyatseni, 2004, 43(sup.): 50–55［何淼，李继东，张尚宏．真核生物内含子数据库的构建．中山大学学报（自然科学版），2004，43(增刊): 50–55]

Herbert A., RNA editing, introns and evolution. Trends Genet., 1996, 12(1): 6–9

Lou H., Gagel R. F., Berget S. M., An intron enhancer recognized by splicing factors activates polyadenylation. Genes Dev., 1996, 10(2): 208–219

Mattick J. S., Introns: evolution and function. Curr. Opin. Genet. Dev., 1994, 4(6): 823–831

Maxwell E. S. and Fournier M. J., The small nucleolar RNA. Annu. Rev. Biochem., 1995, 35: 897–934

Sakharkar M. K., Kangueane P., Woon T. W., Tan T. W., Kolatkar P. R., Long M., de Souza S.J., IE-Kb: intron exon knowledge base. Bioinformatics, 2000, 16(12): 1151–1152

Tee M. K., Babalola G. O., Aza-Blanc P., Speek M., Gitelman S. E., Miller W. L., A promoter within intron 35 of the human C4A gene initiates abundant adrenal-specific transcription of a 1 kb RNA: location of a cryptic CYP21 promoter element? Hum. Mol. Genet., 1995, 4(11): 2109–2116

Vinogradov A. E., Intron-genome size relationship on a large evolutionary scale. J. Mol. Evol., 1999, 49(3): 376–384

Zhang S. H., The origin and evolution of repeated sequences and introns. Speculat. Sci .Technol., 1998, 21(1): 7–16