

真核生物内含子数据库统计规律研究^{*}

何 森¹, 李继东¹, 张尚宏²

(1. 中山大学生命科学学院, 广东 广州 510275;

2. 中山大学基因工程教育部重点实验室/生物工程研究中心, 广东 广州 510275)

摘 要: 根据 GenBank 的序列数据, 构建了真核生物内含子数据库 (EID)。对 EID 统计规律的研究表明, 数据库共有 103 848 个基因, 478 484 个内含子, 582 332 个外显子, 平均每个基因有 4.61 个内含子, 5.61 个外显子, 内含子长度为 40~120 个核苷酸的最多。对人、大鼠、小鼠、鸡、果蝇、线虫、拟南芥、玉米和裂殖酵母等 9 种模式生物的数据的统计分析表明, 在真核生物中, 并不是生物越高等, 基因中的内含子数或外显子数就越大。进一步, 对各种模式生物的基因组大小与内含子比例及内含子密度的关系、内含子相位、内含子剪接位点等特征进行了统计研究。

关键词: 真核生物; 内含子; 数据库; 特征统计

中图分类号: Q754 **文献标识码:** A **文章编号:** 0529-6579 (2005) 06-0079-04

内含子是非编码序列的重要组成部分之一, 普遍存在于真核基因中, 而且往往占了基因长度的很大一部分, 甚至可达 90%^[1]。此外, 内含子还存在于细胞器基因组中以及一些古细菌和真细菌的基因组中^[2]。内含子的起源、进化及功能的问题, 是基因和基因组研究中的一个重大理论问题。对此, 一直有众多的争论。

目前, 一类观点认为内含子是基因和基因组进化的副产物, 并偏向于认为它们是后起源的, 即在功能基因出现之后或在真核生物出现之后由基因组中的已有序列衍生^[1]。另一类观点则认为, 内含子在原始基因和基因组的起源与进化中起着重要的作用, 在现代基因和基因组的整体调控方面也有着重要的功能, 它们在原始基因组中即存在, 因而是先起源的^[1,2]。

在内含子的功能方面, 已有一些例子显示其具有功能多样性的特征, 即内含子可充当启动子^[3]、增强子^[4]、反式作用因子^[5]、参与 RNA 编辑^[6]等。此外, 有不少核仁小分子 RNA (snoRNA) 是由内含子编码^[7]。这些发现对深入探讨内含子的起源与功能无疑都有一定的启示。

生物信息学使我们可以系统地掌握内含子的信息, 并对其进行统计分析, 而这十分有助于促进与内含子有关的探索^[8]。本研究在构建了一个富含准确真核生物内含子信息的数据库^[9]的基础上, 进一

步对内含子特征变量, 如内含子的数量、长度分布、密度、相位、剪接位点等进行了统计分析, 研究数据库中内含子的变化趋势, 并以此来探讨内含子的起源与进化。

1 真核生物内含子数据库的构建及统计方法

1.1 数据来源

真核生物内含子数据库 (EID) 的数据来自从 GenBank (Release 125) 下载并提取的序列信息。下载的数据包括无脊椎动物、灵长类、啮齿类、其他哺乳动物、其他脊椎动物和植物。

1.2 真核生物内含子数据库结构

EID 的设计包括一个主数据库和多个子数据库。主数据库包括 3 个分别代表 DNA EID、Protein-EID、Header EID 的 FASTA 格式文本文件库 dEID、pEID、hEID。子数据库有核 mRNA 内含子数据库、细胞器内含子数据库, 以及灵长类、啮齿类、其他哺乳类、其他脊椎动物、无脊椎动物、植物等的内含子数据库。

1.3 统计方法

根据各数据库的共同文件格式, 应用 PERL 语言分别编写了相应的程序读取文本文件, 统计数据库中如基因数、内含子数和外显子数等基本数据和内含子的特征变量。统计结果的可视化主要应用了

* 收稿日期: 2004-12-09

基金项目: 国家自然科学基金资助项目 (30270752); 广东省自然科学基金资助项目 (031616)

作者简介: 何森 (1963 年生), 男, 副教授, 通讯联系人: 张尚宏, E-mail: lsbrcl1@zsu.edu.cn

PERL 的 GD: GRAPH 模块。

2 EID 数据库的统计结果

2.1 基本统计结果

数据库有 103 848 个基因，合共 478 484 个内含子和 582 332 个外显子，平均每个基因有 4.61 个内含子和 5.61 个外显子。不同物种或同一物种的不同基因所包含的内含子数差别很大，如人的第七类胶原质基因含有 117 个内含子。

图 1 表示了数据库中外显子和内含子的分布模式。外显子长度为 90~120 个核苷酸的最多。这个统计结果与 Dorit 等人用更早 GenBank 版本统计的结果基本一致^[10]。内含子长度则为 40~120 个核苷酸的最多。最长的内含子在人第八号染色体上的锌指蛋白基因 (AF178030) 中发现，达 152 kb；最短的内含子只有 18 个核苷酸，在一种纤毛虫中发现。与外显子长度分布相比，内含子长度分布图中有 2 个峰，分别为 50 和 90 个核苷酸。

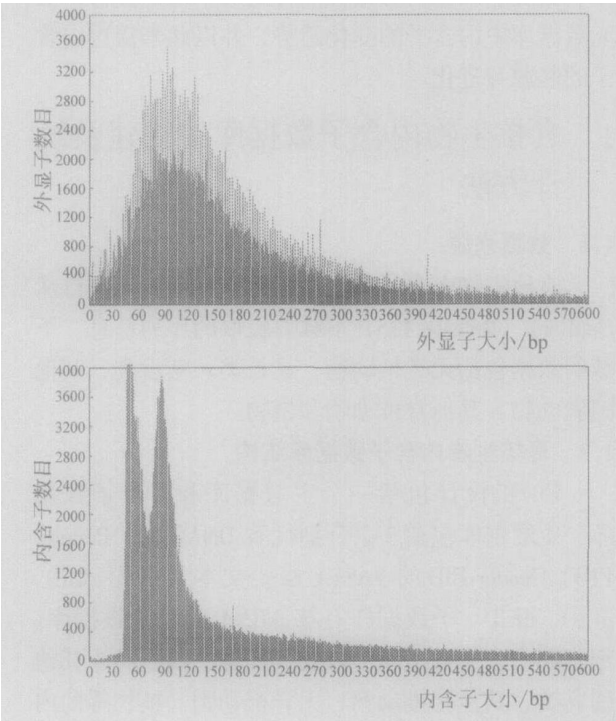


图 1 主数据库的外显子和内含子大小分布图
Fig 1 Distribution of exon size and intron size in EID

此外，建立了 9 种模式生物（人 *Homo sapiens*，大鼠 *Rattus norvegicus*，小鼠 *Mus musculus*，鸡 *Gallus gallus*，果蝇 *Drosophila melanogaster*，线虫 *Caenorhabditis elegans*，拟南芥 *Arabidopsis thaliana*，玉米 *Zea mays* 和粟酒裂殖酵母 *Schizosaccharomyces pombe*）的内含子数据库。用统计程序分别计算各模式生物的基本数据，统计结果见表 1。

在真核生物中，并不是生物越高等，每个基因所包含的内含子数或外显子数就越大。这两项指标的数值除了在裂殖酵母中较小外，在其余物种中并不随着生物的不同而产生显著变化。人每个基因含有的内含子和外显子最多，分别平均有 5.77 个和 6.77 个。

表 1 内含子数据库中 9 种模式生物的统计结果
Tab 1 Statistical results of nine model organisms in EID

数据库	基因数	内含子数	外显子数	内含子数	外显子数
				/基因	/基因
主数据库	103848	478484	582332	4.61	5.61
人	16524	95379	111903	5.77	6.77
大鼠	881	4021	4902	4.56	5.56
小鼠	3296	17310	20606	5.25	6.25
鸡	442	2133	2575	4.83	5.83
果蝇	10961	36730	48054	3.35	4.38
线虫	21064	116548	139216	5.53	6.61
拟南芥	24596	129650	154428	5.27	6.18
玉米	163	741	910	4.55	5.58
裂殖酵母	156	366	524	2.35	3.36

从内含子长度分布图上可以看出，小于 20 bp 的很少，一般短的内含子大约为 20~30 bp。而外显子则没有这个规律，最短的外显子只有几个 bp。这很可能是由于 20~30 个 bp 的序列才能保证剪接中内含子的识别并提供剪接信号。

2.2 基因组大小和内含子关系的统计分析

利用基因组大小及有关子数据库的数据，研究各模式生物的基因组大小与内含子比例（每 kb CDS 所含的内含子总 kb 数）及内含子密度（每 kb CDS 所含的内含子个数）的关系（表 2）。

表 2 基因组大小和内含子的关系
Tab 2 Relationship between genome sizes and introns

物种	数据库核	内含子总	外显子总	基因组	内含子	内含子
(子数据库)	苷酸总量/kb	长度/kb	长度/kb	大小 Mb	比例	密度
人	166413	148769	17653	3165	8.43	5.40
大鼠	2405	1658	747	2700	2.22	5.38
小鼠	13879	10550	3329	3000	3.17	5.20
鸡	1137	715	422	1200	1.69	5.05
果蝇	37405	19489	17916	180	1.09	2.05
线虫	59247	29951	29296	97	1.02	3.98
拟南芥	56130	21975	34155	120	0.64	3.80
玉米	384	203	181	3300	1.12	4.09
裂殖酵母	251	31	220	12.6	0.14	1.66

人的内含子比例达 8.43，小鼠和大鼠次之，鸡的内含子比例又次于以上 3 种哺乳动物；两种无脊椎动物的内含子比例相似且较小，两种模式植物的内含子比例也较小，裂殖酵母的内含子比例最小，只有 0.14。至于内含子密度，脊椎动物普遍

都比较大, 无脊椎动物相对较小; 植物的内含子比例虽然较小, 但内含子密度却不小; 酵母的内含子密度与内含子比例一样, 都是最小。

可以看出, 数据库中核苷酸总量与内含子总长度的相关程度很高, 这是因为数据库中都是基因序列, 所以序列越多, 内含子总长度就越长。基因组大小与内含子密度及内含子比例的相关性没有那么好, 但仍然显示出随着基因组增大, 内含子密度及内含子比例均增大的趋势。另一方面, 似乎内含子密度在具较小基因组的真核生物中变化较大, 而内含子比例则在具较大基因组的真核生物中变化较大。

2.3 内含子相位统计分析

内含子相位是指内含子在基因内相对于密码子的 3 个核苷酸的位置。从表 3 看出, 每一种模式生物的内含子相位不是平均分布的, 相位为 0 的内含子约占所有相位的一半, 明显多于相位 1 和 2 的内含子。该结论与 Fedorov 等利用较早版本的 GenBank 数据统计的结果基本一致^[11]。

表 3 内含子相位统计

Tah 3 Statistical results of intron phases

数据库	相位 0/%	相位 1/%	相位 2/%
主数据库	49.1	24.3	26.6
人	2.2	25.1	22.7
大鼠	54.1	20.4	25.5
小鼠	49.4	25.3	25.3
鸡	51.2	25.1	23.7
果蝇	46.2	25.7	27.9
线虫	51.4	23.5	25.1
拟南芥	3.3	24.2	22.5
玉米	50.8	23.4	25.8
裂殖酵母	45.4	26.4	28.2

“内含子后起源”假说认为, 内含子是在基因组进化中插入到基因内的。因此, 3 种相位的内含子的分布应是随机的, 即各自所占的比例应差不多。而“内含子先起源”假说认为, 最原始的内含子和外显子都是源自早期的微型基因^[12], 因此原始内含子的相位都是 0。其他相位的内含子则是在后来的基因组进化中才产生。实际观察到的情况正是“内含子先起源”观点所预期的。

2.4 内含子剪接位点统计分析

在剪接体内含子中, 剪接位点通常都是 GT...AG (典型剪接位点)。剪接体内含子还有另一种剪接位点: AT...AC^[13]。GT...AG、GC...AG (很可能衍生于 GT...AG) 和 AT...AC 具一定的分布规律。各种剪接位点在主数据库中所占的比例见表 4。

表 4 主数据库中不同剪接位点的统计¹⁾

Tah 4 Distribution of different splice sites of the main database

剪接位点	内含子数	所占比例/%
GT...AG	465086	97.200
NN...NN	10126	2.116
GC...AG	2140	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

1) N 表示未知核苷酸

此外, 对各子数据库进行统计, 得出 9 种模式生物中的内含子典型剪接位点的比例都在 95% 以上。除了典型剪接位点外, 未知的剪接位点 (NN...NN、GT...NN、NN...AG) 占了不小的比例, 这说明与剪接位点有关的注释确实存在不少误差。

3 讨论

从本研究的结果看来, 无论是高等生物还是低等生物, 外显子编码的氨基酸为 30~40 个的最多; 内含子相位为 0 的也最多。这在一定程度上是支持内含子先起源假说。当然, 这并不排除基因组进化中也有内含子是后起源的。

结果还显示, 随着真核生物的基因组增大, 内含子含量也有增多的趋势。这是基因组进化的“大基因组”途径所预期的^[2]。在这种进化途径中, 除了内含子含量增多, 其他非编码序列, 如重复序列也增多^[2]。有关的分析还表明, 随着基因组增大, 基因外的非编码序列增多得更快^[4]。

结合内含子起源与进化问题, 开发更深入的分析软件和新算法, 将能使数据库的分析更具生物学意义。一方面, 把内含子的相位统计分析与外显子内含子连接区域的保守性联系起来, 以探讨内含子与外显子边界序列在结构和功能上的相关性及其起源。另一方面, 把与内含子相关的重复序列信息、内含子内的基因信息和保守区域信息等挖掘出来, 将有利于深入研究内含子与基因组其他序列的进化关系, 如内含子与重复序列的可能进化关系^[2]。

参考文献:

[1] DOOLITTLE W F. The origin and function of intervening sequences in DNA: a review [J]. Amer Nat, 1987, 130 (6): 915—928.
[2] ZHANG S H. The origin and evolution of repeated sequences and introns [J]. Speculat Sci Technol, 1998, 21 (1): 7—

16.

[3] TEE M K, BABALOIA G O, AZA—BLANC P, et al. 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? [J]. Hum Mol Genet, 1995, 4 (11): 2109—2116.

[4] LOU H, GAGEL R F, BERGET S M. An intron enhancer recognized by splicing factors activates polyadenylation [J]. Genes Dev, 1996, 10 (2): 208—219.

[5] MATTIK J S. Introns: evolution and function [J]. Curr Opin Genet Dev, 1994, 4 (6): 823—831.

[6] HERBERT A. RNA editing, introns and evolution [J]. Trends Genet, 1996, 12 (1): 6—9.

[7] MAXWELL E S, FOURNIER M J. The small nucleolar RNA [J]. Annu Rev Biochem, 1995, 35: 897—934.

[8] SAKHARKAR M K, KANGUEANE P, WOON T W, et al. IE—Kb: intron exon knowledge base [J]. Bioinformatics, 2000, 16 (12): 1151—1152.

[9] 何淼, 李继东, 张尚宏. 真核生物内含子数据库的构建 [J]. 中山大学学报: 自然科学版, 2004, 43 (增刊): 50—55.

[10] DORIT R L, SCHOENBACH L, GILBERT W. How big is the universe of exons? [J]. Science, 1990, 250 (4986): 1377—1382.

[11] FEDOROV A, MERICAN A F, GILBERT W. Large—scale comparison of intron positions among animal, plant and fungal genes [J]. Proc Natl Acad Sci USA, 2002, 99 (25): 16128—16133.

[12] GILBERT W. The exon theory of genes [J]. Cold Spring Harb Symp Quant Biol, 1987, 52: 901—905.

[13] HALL S L, PADGETT R A. Requirement of U12 snRNA for in vivo splicing of a minor class of eukaryotic nuclear pre—mRNA introns [J]. Science, 1996, 271 (5256): 1716—1718.

[14] VINOGRADOV A E. Intron—genome size relationship on a large evolutionary scale [J]. J Mol Evol, 1999, 49 (3): 376—384.

Studies on Characteristic Statistics of Eukaryotic Intron Database (EID)

HE Miao¹, LI Ji dong¹, ZHANG Shang hong²

(1. School of Life Sciences, Sun Yat sen University, Guangzhou 510275, China;

2. The Key Laboratory of Gene Engineering of Ministry of Education //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 of the statistical characteristics of EID show that there are 103848 genes, 478484 introns and 582332 exons, with averagely 4.61 introns and 5.61 exons per gene. Introns with 40~120 nt in length are abundant in the database. Results of the statistical analyses of the data from 9 model species show 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, relationship between genome size and intron proportion or intron density have been studied.

Key words: eukaryote; intron; database; characteristic statistics