

论生物基因组的起源和进化^{*}

张尚宏

(中山大学生物工程研究中心, 广州 510275)

摘要 通过分析核(类核)基因组、线粒体基因组和叶绿体基因组的大小与结构,发现这三种基因组都有“小基因组”和“大基因组”的形式及与之对应的结构特点。认为它们各自的祖先都是通过分别与“小基因组”和“大基因组”相应的两种途径进化成为现代核(类核)基因组、线粒体基因组和叶绿体基因组,发展了生物基因组起源和进化的统一模式。

关键词 核(类核),线粒体,叶绿体,基因组结构,起源,进化

分类号 Q 349

对基因组起源和进化的研究,是探索生命起源和进化的重要组成部分。根据现代基因组在细胞中的地位及功能,可以把它们分为自主性的核(类核)基因组、处于细胞器中半自主性的线粒体基因组和叶绿体基因组。从进化的角度看,这三种基因组都是源于一种原始基因组。因此,对现代各种基因组的结构进行深入分析,找出原始基因组的“遗迹”及其变化趋势,是研究基因组起源和进化的有效途径。本文正是在这方面进行一些开创性的探讨。

1 核(类核)基因组的结构

核基因组和类核基因组在细胞中的地位是等价的,在进化上也必须一起考虑。但从结构方面看,两者相差甚远,可以很自然地把它分开。原核生物类核基因组一般是环状,再形成类核高级结构。通常基因组大小只有数百万碱基对。类核基因组都是由单拷贝或低拷贝的 DNA 序列组成。基因的排列比较紧密,一般以多顺反子为转录单位,较少非编码序列,非编码的重复序列和内含子就更少了。至今为止,只在盐杆菌 (*Halobacterium*)、硫氧化菌 (*Sulfolobus*) 和脱硫球菌 (*Desulfurococcus*) 等原细菌的基因组中,以及在部分真细菌的基因组中发现少量种类各异的重复序列或内含子^[1~4]。至于真核生物的核基因组,一般是多线状,再形成染色体或间期核的高级结构,大小要比类核基因组大好几个数量级,且变化范围也很大,已知最小的多细胞动物核基因组约为 $\approx 10^5$ kb^[5],最小的高等植物核基因组也有 7×10^4 kb。在结构组成上,核基因组以普遍存在大量非编码序列为特点。其中重复序列一般可占基因组的 1/3,在有些植物核基因组中甚至可达 90%^[4]。内含子也占基因组相当重要的比例,真核基因基本上都是含有内含子的断裂基因。一个基因的长度可有多至 90% 是内含子^[6]。重复序列和内含子的含量直接影响了基因组的大小。

因此,在大小和构成上,类核基因组是“节俭型”的,而核基因组是“冗余型”的。节俭型

* 收稿日期: 1995-06-14 张尚宏,男,33岁,理学博士,副教授

与冗余型的差别在于基因组的大小以及重复序列和内含子的含量,所以也可把它们分别称为“小基因组”型和“大基因组”型。“小基因组”型以基因组较小,不含或只含少量重复序列和内含子为特点;“大基因组”型则以基因组巨大,含有大量的重复序列和内含子为特点。

2 线粒体基因组的结构

线粒体的功能在各类真核生物中都相同,但其基因组却变化很大。多细胞动物的线粒体基因组十分细小和致密,绝大部分的大小都在 15.7~ 19.5 kb 这个范围内(表 1)。基因组呈单环状,没有或很少间隔序列,基因中没有内含子,不存在非编码的重复序列^[5]。

相反,高等植物线粒体基因组是最大、最复杂的。基因组呈单环或多环状,在已研究过的种类中,大小范围从 186 kb 至 2 400 kb(表 1)。由于植物线粒体基因组所含的基因只比动物的多两倍左右,因此,有很多基因间隔,另外还有很多重复序列和内含子^[5]。

表 1 真核生物线粒体基因组的大小和构型^[5]

Tab. 1 Sizes and conformations of mitochondrial genomes in various kinds of eukaryotes

物种	大小 (kb)	构型 ¹⁾	物种	大小 (kb)	构型 ¹⁾
脊椎动物 昆虫			原生动物		
<i>Homo sapiens</i>	16.569	环状	<i>Plasmodium lophurae</i>	18.5	环状
<i>Mus musculus</i>	16.295	环状	<i>Tetrahymena pyriformis</i>	55	线状
<i>Gallus domesticus</i>	16.2	环状	<i>Trypanosoma brucei</i>	20~ 22	大环
<i>Drosophila neohydei</i>	15.7	环状	藻类		
<i>D. melanogaster</i>	18.7~ 19.5	环状	<i>Chlamydomonas reinhardtii</i>	15.8	线状
高等植物			<i>Euglena gracilis</i>	60	—
<i>Marchantia polymorpha</i>	186.608	环状	<i>Olisthodiscus luteus</i>	40	环状
<i>Brassica hirta</i>	208	环状	真菌		
<i>B. campestris</i>	218	三环	<i>Aspergillus nidulans</i>	32~ 33	环状
<i>Pisum sativum</i>	370	—	<i>Cochliobolus heterostrophus</i>	115	环状
<i>Cucumis sativus</i>	1 500	—	<i>Hansenula mrakii</i>	55	线状
<i>C. melo</i>	2 400	—	<i>Saccharomyces cerevisiae</i>	68~ 77.8	环状
<i>Zea mays</i>	570	多环	<i>Schizosaccharomyces pombe</i>	17.6~ 23	环状

1) 近年发现高等植物和真菌的线粒体基因组构型在大多数情况下是线状的

原生动物线粒体基因组的大小在 18.5~ 55 kb 之间,呈环状、线状或由大环与小环组成的网状结构(表 1)。动物目(Kinetoplastida)种类的动基体大环 DNA 除了有一段 15~ 17 kb 的转录区外,还有由长度不等的重复序列组成的多变区域^[5]。其他一些原生动物的线粒体基因组中也可有少量重复序列,不过还没有在研究过的基因组中发现内含子^[5]。

对藻类的线粒体基因组研究不多,已知莱茵衣藻(*Chlamydomonas reinhardtii*)的 15.8 kb 基因组的结构象多细胞动物的而不象高等植物的^[5]。

真菌的线粒体基因组大小在 17.6~ 115 kb 之间,绝大部分呈环状(表 1)。基因组的大小主要由基因间隔的长度、内含子及重复序列等决定。随着基因组变小,这几种非编码序列,特别是内含子也减少。但即使在最小的基因组中,也有个别基因具内含子^[5]。

概括起来,多细胞动物和高等植物线粒体基因组的差异与类核基因组和核基因组的显著差别类似。它们代表了线粒体基因组在大小、结构上的两种形式。因此也可以把它们分

别称为(线粒体)“小基因组”型(动物型)和(线粒体)“大基因组”型(植物型)。“小基因组”型以线粒体基因组细小、致密,基本上不含非编码序列(重复序列、内含子)为特点;“大基因组”型则以线粒体基因组较大,含有很多非编码序列为特点。至于原生动物、藻类和真菌的线粒体基因组,一般来说,基因组较小的偏向于动物型;基因组较大的则偏向于植物型^[5]。

3 叶绿体基因组的结构

在研究过的藻类、苔藓植物、蕨类植物以及种子植物中,叶绿体基因组绝大多数呈单环状,大小在 85~ 292 kb 之间(表 2)。若不考虑刺松藻(*Codium fragile*)、摩氏衣藻(*Chlamydomonas moewusii*)等几种藻类,则这一变化范围大大缩小。在研究过的高等植物中,叶绿体基因组的大小在 120~ 217 kb 之间,若再除去烟草(*Nicotiana accuminata*)、少根紫萍(*Spirodela oligorhiza*)和天竺葵(*Pelargonium hortorum*)这 3 种被子植物(它们的叶绿体基因组分别为 171 kb、180 kb 和 217 kb),则变化范围就只有 120~ 160 kb^[7]。

在这些叶绿体基因组中,基因的总数达 100 个以上^[7],从而对于一个 150 kb 左右的基因组来说,基因的排列还是相当紧凑的。在基因组中只发现少量的非编码重复序列^[7],其拷贝数的多少对基因组大小有一定的影响。此外,基因组中还可有几个至几十个内含子(与种类有关)。内含子在某一基因内存在与否是比较随意的,但仍具有物种特异性。

表 2 高等植物和藻类叶绿体基因组的大小和构型^[7]

Tab. 2 Sizes and conformations of chloroplast genomes of higher plants and algae

物 种	大小 (kb)	构型	物 种	大小 (kb)	构型
被子植物 (200多种, 90多属, 30多科)	120~ 217	环状	藻类		
裸子植物			<i>Godium fragile</i>	85	环状
<i>Ginkgo biloba</i>	158	环状	<i>Chlamydomonas reinhardtii</i>	195	环状
<i>Pinus</i> (2种)	120	环状	<i>C. moewusii</i>	292	环状
蕨类植物			<i>Vaucheria sessilis</i>	125	环状
<i>Osmunda cinnamomea</i>	144	环状	<i>Olisthodiscus luteus</i>	154	环状
<i>Asplenium nidus</i>	150	环状	<i>Euglena gracilis</i>	130~ 152	环状
<i>Pteris vittata</i>	150	环状	<i>Cyanophora paradoxa</i>	127	环状
苔藓植物			<i>Dictyota dichotoma</i>	123	环状
<i>Marchantia polymorpha</i>	121. 024	环状	<i>Plylaiella littoralis</i>	140	多环
<i>Sphaerocarpus donnellii</i>	125	环状	<i>Acetabularia mediterranea</i>	约 2 000	—
			<i>A. cliftonia</i>	约 2 000	—

叶绿体基因组在大多数种类中都比较均一,不过伞藻属(*Acetabularia*)种类的却与众不同,它们的大小约为 2 000 kb(表 2)。所含的基因和一般的叶绿体基因组差不多,但却含有较多的非编码序列,包括重复序列^[4]。它们无疑是代表叶绿体基因组的另一种形式。考虑到有很多藻类的叶绿体基因组还未被研究过,很可能会再发现一些类似伞藻属的情况。

因此,叶绿体基因组也有两种可从大小及相应的结构特点加以区别的形式,也可以把它们分别称为(叶绿体)“小基因组”型和(叶绿体)“大基因组”型。“小基因组”型包括大部分藻类和高等植物的叶绿体基因组,以基因组较小,基因排列较紧密,非编码序列(重复序列、内含子)虽然存在,但数量较少为特点;“大基因组”型则以伞藻属的叶绿体基因组为代表,以基因组较大,含有很多非编码序列为特点。

4 原始基因组的构成

通过以上分析,不难看出:核(类核)基因组、线粒体基因组和叶绿体基因组都分别存在着“小基因组”型和“大基因组”型,而且三种基因组的“小基因组”型的结构特征十分相似,三种基因组的“大基因组”型的结构特征也同样相似,这说明三种基因组都有一些共同的进化过程,至于这三种基因组的具体进化途径,则取决于原始基因组的构成。

对原始基因组的构成已有不少研究,越来越多的事实支持原始基因组含有重复序列和内含子的观点,其中比较重要的是:①生物大分子的自发形成是一个非随机的过程,从而其序列种类是有限的^[6],再加上竞争排斥作用,就使得生命起源时期或生物进化初期能自复制的原始生物大分子的序列多样性很低^[8],这样,重复序列就可以通过原始生物大分子之间的化学连结而产生^[6]。②在研究得比较详细的现代基因组中,发现很多所谓的单拷贝序列其实是分化了的重复序列^[4]或化石重复序列^[6],从而重复序列对现代基因组的构成所起的作用就更大了。③对现代 DNA 序列的分析表明,寡聚核苷酸的重复序列是原始编码序列,现代编码序列是从周期性过渡到混沌性的^[4]。④对大量核苷酸序列的统计学分析发现,核苷酸序列中存在一种难以解释的自相似或具分形性质的远距相关现象^[9],而且这种现象只在具内含子的基因序列及非转录的调控序列中发现,在 cDNA 序列和不具内含子的基因序列中却不存在^[9],这一现象可以很好地解释为原始基因组中的重复序列在现代基因组中的进化遗迹,同时也暗示着重复序列与内含子序列及调控序列在起源上的关系(至于外显子序列,由于它们是高度特化的,所以有关与重复序列在起源上的可能关系的遗迹已消失)^[4]。⑤已有一些迹象表明,内含子与外显子之间的连接序列(具特异重复模式)的起源是与重复序列有关的,是原始重复序列的一种遗迹^[4],从而可以认为内含子(断裂基因)是在生物进化早期直接由原始重复序列进化产生的。⑥相应于原核生物基因组系与相应于真核生物的基因组系在生物进化的早期就开始分化了,而有分子证据显示内含子在原核生物与真核生物分化之前就存在^[4]。⑦现代基因组中还有一些内含子具自剪接及催化特性,说明它们是一种原始结构^[4]。⑧作为原始基因组含有重复序列和内含子的重要非直接证据,在原核生物基因组、线粒体基因组和叶绿体基因组中都发现重复序列和内含子的存在^[1-3,5,7]。

综合众多的事实,可得出原始基因组由重复序列和具内含子的原始基因构成的结论,这一观点不但比重复序列和内含子后起源的观点能更好地解释这些非编码序列在各种基因组中的存在模式及潜在功能^[4],而且还能把重复序列和内含子的起源与进化联系在一起。

5 生物基因组的起源和进化

既然一系列事实表明原始基因组是含有重复序列和内含子,而且断裂基因(内含子)的起源是与重复序列有关的,那么,最原始的基因组就很有可能完全由重复序列组成^[6,8],即生物基因组是起源于重复序列,在这样的一个基础上,进化产生原始基因组^[4,6]。

核(类核)基因组是比较直接地从原始基因组进化而来,而从原始基因组进化成线粒体基因组和叶绿体基因组还要经历一个内共生过程,但不管怎样,由于处于基因组进化的原始阶段,进入内共生状态初期的线粒体和叶绿体各自祖先(真细菌类)的基因组是仍然含有重复序列和内含子的^[4,5],起码,对含有内含子这一点已有不少共识^[7],因此,核(类核)基因组、线粒体基因组和叶绿体基因组各自的祖先在结构特征上是相似的,而对现代三种基因

组的结构分析发现,它们都存在着“小基因组”型和“大基因组”型以及相对应的结构特点.因此,这三种基因组都有相应于“小基因组”型和“大基因组”型的两种进化途径.

进化途径是基于进化可能性的.在进化过程中,对生命始终存在着效率方面的选择压力.如果没有其他限制,生物对这种选择压力就可有两种应答方式:或者变得具更高效率、更快速繁殖,使这种选择压力得以加强,从而使向高效率的进化成为一种趋向;或者变得更复杂、具更多功能,从而减少了效率方面的选择压力,使进化可以继续向复杂性和多功能方向发展.这样,基因组大小的进化可能性就成为两种进化途径:或者向高效的“小基因组”发展,或者向复杂的“大基因组”发展.如果效率选择压力对不同的生物(或生物结构)原本就有差异(正如动物与植物线粒体的情况^[5]),则同样会造成相似的进化结果.因此,原始基因组若向“小基因组”方向进化,是由于效率选择压力大,而且有空间大小的约束.这种进化使基因组趋向于只含必需序列,非必需的重复序列和内含子在进化过程中就被丢掉了.若向“大基因组”方向进化,是由于效率选择压力小.这种进化使基因组中的重复序列和内含子可以与必需基因共存及发展,并起着结构上与进化上的作用^[4,6],同时也使基因组大小的变化范围可以很大.这两种进化趋向正是各种生物基因组共有的进化途径,即分别是“小基因组”途径和“大基因组”途径.“小基因组”进化途径以丢失、残留重复序列和内含子为特点,包括类核基因组、动物型线粒体基因组以及大部分藻类和高等植物的叶绿体基因组的进化;“大基因组”进化途径以保留、发展重复序列和内含子为特点,包括核基因组、植物型线粒体基因组和以伞藻属为代表的叶绿体基因组的进化.这样,通过这两种进化途径就把现代三种基因组的结构特点的产生予以说明,并把它们的进化统一起来了(表 3).

表 3 3种基因组的结构共同点及共有的两种进化途径

Tab. 3 Common structural features and the two evolutionary scenarios of the three kinds of genomes

基因组种类	“小基因组”型	“大基因组”型	基因组进化	“小基因组”途径	“大基因组”途径
核(类核)基因组	原核生物	真核生物	原始 [核(类核)、线粒体、叶绿体] 基因组	在同一种基因组中属小型,基因排列紧密,重复序列和内含子在进化过程中丢失,因此不含或只有较少这些成分	在同一种基因组中属大型,基因排列分散,在进化过程中保留并发展重复序列和内含子
线粒体基因组	后动物,具较小线粒体基因组的低等真核生物	高等植物,具较大线粒体基因组的低等真核生物	现代 [核(类核)、线粒体、叶绿体] 基因组		
叶绿体基因组	大部分藻类和高等植物	以伞藻属为代表			

既然原始的核(类核)基因组、线粒体基因组和叶绿体基因组都有重复序列和内含子,那么在这三种基因组的“小基因组”进化途径中,应该都存在着去除这些成分的机制.实际上,内含子在某一基因内存在与否具随意性^[4,8],DNA剪接可通过整合酶催化的逆反应而实现^[4].这些事实说明这种机制在进化中是存在的,它可能包括剪接、逆转录、基因转换等.

6 结 语

从核(类核)基因组、线粒体基因组和叶绿体基因组的结构特点可以看出它们都具有两种进化途径,这说明生物进化在基因组水平上是有一定的方向性的.两种进化途径的存在首先是由于进化的可能性所允许,其次是选择压力的差异导致各种基因组的两极分化.沿“小基因组”途径进化的生物与沿“大基因组”途径进化的生物不但在有关基因组的结构组成上十分不同,而且在整体上也差别很大,以至它们的系统分类位置一般相差甚远,如核(类核)基因组进化途径不同的原核生物与真核生物、线粒体基因组进化途径不同的动物与植

物。不过,叶绿体基因组沿“大基因组”途径进化的伞藻却与沿“小基因组”途径进化的其他绿藻在现有分类系统上同属一个门,甚至同一个纲。但实际上,伞藻这一类群除了叶绿体基因组的独特性外,还有不少其他特异性状,如 65%~80% 的叶绿体不含 DNA^[4],在形态上也有独特的巨大主干细胞。因此,有必要重新考虑伞藻类的系统分类地位(另文论述)。

参 考 文 献

- 1 Daniels D L, Plunkett G III, Burland V, et al. Analysis of the *Escherichia coli* genome DNA sequence of the region from 84.5 to 86.5 minutes. *Science*, 1992, 257: 771~778
- 2 李靖炎. 真核细胞起源研究的进展. 见: 细胞生物学进展(第一卷). 北京: 高等教育出版社, 1989. 185~209
- 3 Ferat J L, Michel F. Group II. Self-splicing introns in bacteria. *Nature*, 1993, 364: 358~361
- 4 Zhang S H. The origin and evolution of repeated sequences and introns. *Speculat Sci Technol*, 1996, 19
- 5 张尚宏. 动物与植物线粒体基因组结构的差异——两种进化途径. *动物学研究*, 1995, 16(2): 132~145
- 6 张尚宏. 生物大分子与生命的起源和进化及其哲学思考. 见: 遗传学基础理论问题讨论文集. 北京: 北京师范大学出版社, 1993. 1~25
- 7 Sugiura M. The chloroplast chromosomes in land plants. *Annu Rev Cell Biol*, 1989, 5: 51~70
- 8 张尚宏. 逻辑斯缔方程在研究原始生物大分子动态及生命起源问题中的应用. *生态学报*, 1994, 14(1): 90~94
- 9 Peng C K, Buldyrev S V, Goldberger A L, et al. Long-range correlations in nucleotide sequences. *Nature*, 1992, 356: 168~170

On the Origin and Evolution of Organic Genomes

Zhang Shanghong*

Abstract The sizes and structural features of nuclear(nucleoid) genomes, mitochondrial genomes and chloroplast genomes were analyzed, revealing that these three kinds of genomes all have the“small genome” type and the“large genome” type with corresponding structural features. It is concluded that the three kinds of genomes would have evolved according to the“small genome” and the“large genome” evolutionary scenarios. A general pattern for the origin and evolution of organic genomes has been developed.

Keywords nucleus(nucleoid), mitochondrion, chloroplast, genome evolution

* Biotechnology Research Center, Zhongshan University, Guangzhou 510275