74 种鸟类线粒体基因组碱基组成及特征分析*

高英凯[1]，苗永旺[1,2**]，苏小茜[1]，池振奋[3]，俞 赟[3]，姜 枫[1,3]
（1. 云南农业大学 动物科学技术学院，云南 昆明 650201；2. 云南大学，云南省生物资源保护与利用重点实验室，云南 昆明 650091；3. 中国科学院 北京基因组研究所，北京 101300）

摘要：在鸟类线粒体基因组中，不同物种的线粒体碱基组成和特性存在明显的差异。截至 2008 年 5 月，GenBank 细胞器基因组资源数据库共公布了 74 种鸟类线粒体全基因组数据。本研究利用已公布的鸟类线粒体基因组全序列分析其碱基组成特征。结果表明：（1）鸟类线粒体基因组密码子第 2 位的 GC 含量值波动范围十分狭窄，而密码子第 3 位的 GC 含量值波动范围很大。（2）密码子第 3 位碱基中 C 的含量波动范围较大，为 32.60% ~50.70%。（3）线粒体基因组 GC 含量主要由密码子第 3 位的碱基 C 和 T 的变化引起。（4）密码子第 3 位碱基的 GC 含量与线粒体基因组的 GC 含量的变化存在相关性。本次结果为今后深入研究鸟类线粒体基因组提供了一定的借鉴和参考资料。

关键词：鸟类；线粒体基因组；碱基组成；GC 含量；特征
中图分类号：Q 953.3 **文献标识码**：A **文章编号**：1004 - 390X（2009）01 - 0051 - 08

A Comprehensive Analysis on 74 Avian Mitochondrial Genome Base Compositions

GAO Ying-kai[1], MIAO Yong-wang[1,2], SU Xiao-xi[1], CHI Zhen-fen[3], YU Yun[3], JIANG Feng[1,3]
（1. College of Animal Science and Technology, Yunnan Agricultural University, Kunming 650201, China;
2. Laboratory for Conservation and Utilization of Bio-resources, Yunnan University, Kunming 650091, China;
3. Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China）

Abstract: The unusual feature of several species and heterogeneity of gene nucleotide content are common phenomena that exist in the avian mitochondrial genomes. In the present study, the base compositions of functional site over 74 avian mitochondrial genomes retrieved from GenBank Organelle Genome Resources on May 2008 were analyzed. The results showed as follow: (1) The GC-content at the third codon position typically showed the largest range of variation, while the GC-content at the second codon position was not subject to large fluctuation. (2) By ordering the species by increasing percentage of C, a very large variation in composition from 32.60% to 50.70% was observed. (3) Most major changes of GC-content in the whole genomes between species could be attributed to shifts between the proportions of C and T in the third position. (4) In addition, for each of the 74 species considered in this study, GC-content at the third codon position has been significant positive correlated with GC-content of the whole genome sequence. All of the results will provide fundamental information for further study on avian mitochondrial genome.

Key words: avian; mitochondrial genome; base composition; GC content; characteristic

收稿日期：2008 - 09 - 09 修回日期：2008 - 10 - 17
* 基金项目：国家自然科学基金项目（30660024）；云南省应用基础研究重点项目（2007C0003Z）；云南省应用基础研究计划面上项目（2006C0034M）；云南省教育厅科学研究基金项目（5Y0196B）。
作者简介：高英凯（1983 - ），男，在读硕士研究生，主要从事动物分子遗传学研究。
* * 通讯作者 Corresponding author：苗永旺（1964 - ），男，内蒙古通辽人，教授，主要从事家养动物遗传学教学与科研工作。E - mail：yongwangmiao999@ yahoo. com. cn

线粒体是真核细胞内重要的细胞器，是能量生成的场所。线粒体拥有自身的一套遗传控制系统，同时与细胞核基因组 DNA 相互协调[1]。与其它脊椎动物线粒体 DNA（mtDNA）一样，鸟类 mtDNA 是一种双链 DNA 分子，含有 13 个左右的蛋白编码基因，2 个核糖体 RNA 和 22 个左右的 tRNA 基因[2]。蛋白编码基因包括编码细胞色素 b（Cytb）和 ATP 酶 2 个亚基基因（ATPase6，ATPase8）、细胞色素 C 氧化酶的 3 个亚基基因（COX1，COX2，COX3）和尼克酰胺腺嘌呤二核苷酸氧化还原酶 7 个亚基（ND1，ND2，ND3，ND4，ND4L，ND5，ND6）的基因。其中 ND6 和 6 个 tRNA 基因（Glu，Gln，Ala，Asn，Cys 和 Tyr）由轻链编码，其余的皆为重链编码。但是与脊椎动物相比，鸟类 mtDNA 缺少一个形成轻链复制起始（O_L）的发夹结构，该发夹结构存在于 tRNA－Asn 和 tRNA－Cys 之间的一个称为 WANCY 的 tRNA 簇内部[3]。由于 mtDNA 存在进化速率较快、缺乏 DNA 重组和便于分离扩增等特点[2]，并且随着基因组研究技术的发展和测序成本的降低，线粒体基因组序列分析越来越广泛地运用于分子进化、系统发育和家养动物的亲缘分析等领域的研究。

在中性进化和选择进化两种学说的争论中，决定 DNA 碱基组成的主要因素是最具有争议性的问题之一[4,5]。DNA 的鸟嘌呤和胞嘧啶的碱基含量（GC 含量）在不同物种的基因组中波动范围广阔[6~8]，甚至在同一物种的基因组不同区域也不相同[9]。这种基因组局部碱基不均衡的现象表现为不同区域的核苷酸碱基组成存在差异[10]。与密码子前两位的碱基组成相比，密码子第 3 位的碱基组成变异率更大[8]。而这种情况是否是选择或者中性变化的结果，目前尚无定论[4]。此外，GC 含量的差异也被认为与重复元件分布、甲基化模式和基因密度等基因组特征紧密相关[5]。

与其它许多脊椎动物不同的是，多数鸟类的线粒体 ND3 基因中都发现存在特定位点胞嘧啶插入的现象，这种现象在今颌超目（*Neognathae*）中的原鸡（*Gallus gallus*）、翻石鹬（*Arenaria interpres*）、和古颌超目（*Palaeognathae*）中的凤头[共鸟]（*Eudromia elegans*）[11~13]都被发现。造成这种现象的原因并未研究清楚[11]，通常认为额外插入的这个核苷酸是受到 RNA 编辑或者翻译位移（translational frameshifting）的作用。在 ND3 基因中的插入胞嘧啶是否是鸟类的共同特征之一，还是具有一定的种属的特异性，尚有待足够的鸟类线粒体基因组数据加以证明[14]。

本研究采用线性回归统计分析方法对现有已报道的 74 种鸟类线粒体基因组全序列进行比较分析，以期阐明鸟类不同物种的线粒体基因组全序列及其线粒体基因组中蛋白质编码序列、核糖体编码序列和密码子的 3 个不同位点的碱基组成变化特征和差异，为今后鸟类线粒体基因组的深入研究提供依据和参考。

1 材料与方法

1.1 数据来源和数据分类

截止 2008 年 5 月 1 日，GenBank 细胞器基因组资源数据库（Organelle Genome Resources，http：//www. ncbi. nlm. nih. gov/genomes/ORGANELLES/organelles. html）总共公布 74 条鸟类（见表 1）线粒体全基因组序列数据。登录 GenBank 使用 Batch Entrez Tools 将所有序列信息（Accession Number，AN）下载到本地数据库，使用 Bioperl 工具包和自编写的 Perl 程序将所下载的序列信息按照全长序列、蛋白质编码区、核糖体 RNA 区和密码子不同位置进行分类整理。

1.2 线粒体基因组 GC 含量和长度分析

将分类整理后的数据复制到 BioEdit 7.0 软件，使用 BioEdit 7.0 软件中的核苷酸组成分析（nucleotide composition）功能计算基因组中不同区域的序列碱基组成，[15]得出相应的 GC 含量和序列长度的值，配合 Excel 软件将获得的鸟类线粒体全基因组的 GC 含量和序列长度的值进行排序，使用 SPSS v15.0 统计软件计算它们的最大值、最小值、分布范围、平均值、标准差、变异系数、偏度系数和峰度系数等数值。

1.3 线粒体基因组 GC 参数相关性和不同密码子位置的碱基分布

基因组 GC 参数相关性分析使用 MEGA v4.0 软件分别计算密码子不同位点的 GC 含量及碱基组成，设置参数如下：“Nucleotide Sequence、Protein-Coding data、Select Genetic Code = Vertebrate Mitochondrial”。将所得出的结果使用 SigmaPlot 软件构建线性回归方程及图谱，同样使用 SPSS v15.0 统计软件进行最大值、最小值、分布范围、平均值的计算。研究线粒体基因组全序列、蛋白

质编码序列、核糖体编码序列、密码子的 3 个不同位点碱基组成特点。

1.4 ND3 基因胞嘧啶插入现象

根据所下载序列中的生物体（ORGANISM）信息，按照动物分类地位中的目级分类为单位，将所有 74 种鸟类进行划分。通过 Bioperl 工具包中的 Bio:Seq 模块截取 ND3 基因的注释信息逐个判断是否存在 ND3 基因胞嘧啶插入现象。[16]

表 1　本研究中的鸟类物种

Tab. 1　Avian species under consideration in current study

登陆号 AN	物种 species	登陆号 AN	物种 species	登陆号 AN	物种 species
NC_000877	*Aythya americana*	NC_007598	*Nisaetus nipalensis*	NC_007897	*Taeniopygia guttata*
NC_004539	*Anser albifrons*	NC_007599	*Nisaetus alboniger*	NC_007975	*Cnemotriccus fuscatus*
NC_005933	*Anseranas semipalmata*	NC_008547	*Falco sparverius*	NC_010227	*Acrocephalus scirpaceus*
NC_007011	*Branta canadensis*	NC_008548	*Micrastur gilvicollis*	NC_010228	*Sylvia atricapilla*
NC_007691	*Cygnus columbianus*	NC_008550	*Pandion haliaetus*	NC_010229	*Sylvia crassirostris*
NC_009684	*Anas platyrhynchos*	NC_001323	*Gallus gallus*	NC_007979	*Phaethon rubricauda*
NC_008540	*Apus apus*	NC_003408	*Coturnix japonica*	NC_010089	*Phoenicopterus ruber roseus*
NC_002782	*Apteryx haastii*	NC_004575	*Coturnix chinensis*	NC_008546	*Dryocopus pileatus*
NC_002778	*Casuarius casuarius*	NC_006382	*Numida meleagris*	NC_008549	*Pteroglossus azara flavirostris*
NC_002784	*Dromaius novaehollandiae*	NC_007227	*Alectura lathami*	NC_008140	*Podiceps cristatus*
NC_003712	*Arenaria interpres*	NC_007235	*Gallus gallus spadiceus*	NC_010095	*Tachybaptus novaehollandiae*
NC_003713	*Haematopus ater*	NC_007236	*Gallus gallus gallus*	NC_007172	*Thalassarche melanophris*
NC_007006	*Larus dominicanus*	NC_007237	*Gallus gallus bankiva*	NC_007174	*Pterodroma brevirostris*
NC_007978	*Synthliboramphus antiquus*	NC_007238	*Gallus varius*	NC_005931	*Strigops habroptilus*
NC_002196	*Ciconia boyciana*	NC_007239	*Gallus lafayetii*	NC_009134	*Melopsittacus undulatus*
NC_002197	*Ciconia ciconia*	NC_007240	*Gallus sonneratii*	NC_000846	*Rhea americana*
NC_007628	*Cathartes aura*	NC_010195	*Meleagris gallopavo*	NC_002783	*Pterocnemia pennata*
NC_008132	*Nipponia nippon*	NC_007007	*Gavia stellata*	NC_004538	*Eudyptula minor*
NC_008551	*Ardea novaehollandiae*	NC_008139	*Gavia pacifica*	NC_008138	*Eudyptes chrysocome*
NC_009736	*Egretta eulophotes*	NC_010091	*Rhynochetos jubatus*	NC_005932	*Ninox novaeseelandiae*
NC_002672	*Dinornis giganteus*	NC_010092	*Porphyrio hochstetteri*	NC_002785	*Struthio camelus*
NC_002673	*Emeus crassus*	NC_000879	*Smithornis sharpei*	NC_002772	*Eudromia elegans*
NC_002779	*Anomalopteryx didiformis*	NC_000880	*Vidua chalybeata*	NC_002781	*Tinamus major*
NC_000878	*Falco peregrinus*	NC_002069	*Corvus frugilegus*	NC_010094	*Archilochus colubris*
NC_003128	*Buteo buteo*	NC_007883	*Menura novaehollandiae*	–	–

2　结果与分析

2.1 基因组 GC 含量

鸟类动物线粒体基因组的 GC 含量（GCAll）、蛋白质编码区 GC 含量（GCPro）、核糖体 RNA 区 GC 含量（GCrRNA）、密码子第 1 和 2 位 GC 含量（GC12）以及密码子第 3 位 GC 含量（GC3）的基本情况见表 2。GCAll 以及 GC3 分布情况见图 1。GCAll 平均含量与 GCPro，GCrRNA 相似，波动范围也相同。相对而言，GC3 值的波动范围大，最低的 GC 含量是雀形目中的暗褐霸鹟（*Cnemotriccus fuscatus*），仅为 38.0%（NC_007975，长度 17 171 bp），最高的 GC 含量是鴷形目的北美黑啄木鸟（*Dryocopus pileatus*）达到 57.9%（NC_008546，长度 16 832 bp），最低与最高 GC3 值之间的差值为 19.9%。GCAll 值最低的物种同样是暗褐霸鹟，为 42.3%（NC_007975，长度 17 171 bp）；而 GCAll 值最高的物种也同样是北美黑啄木鸟，为 49.5%（NC_008546，长度 16 832 bp），最低与最高 GC3 值之间的差值为 7.2%。表 3 显示线粒体基因组 GC 含量的物种分布，结果显示 87.8%（64/74）的 GC1 的值位于 47.5% ~ 52.5%；100%（74/74）的 GC2 的值位于 40.0% ~ 50.0%。说明相对其他几种 GC 含量而言，鸟类线粒体基因组 GC1 和 GC2 的值分布范围十分狭窄，主要集中在 5% 的范围之类。

表 2 基因组及不同研究区域的 GC 含量常规统计表

Tab. 2 Description statistics of GC contents at the different regions of Mt genomes %

参数 items	GCALL	GCPro	GCrRNA	GC1	GC2	GC3	GC12
最小值 minimum	42. 30	42. 60	43. 00	47. 90	40. 90	38. 00	44. 80
最大值 maximum	49. 50	50. 70	50. 30	53. 50	42. 80	57. 90	47. 80
分布范围 range	7. 20	8. 10	7. 30	5. 60	1. 90	19. 90	3. 00
平均值 mean	45. 53	46. 52	46. 21	50. 88	41. 85	46. 83	46. 37
中值 median	45. 50	46. 65	46. 30	50. 75	41. 80	46. 75	46. 35
标准差 Std. Deviation	1. 4015	1. 6416	1. 3174	1. 2032	0. 3589	3. 8749	0. 656
变异系数 C. V.	1. 1702	1. 1901	1. 1698	1. 1169	1. 0465	1. 5237	1. 067
方差 variance	1. 964	2. 695	1. 736	1. 448	0. 129	15. 015	0. 43
偏度系数 skewness	0. 105	0. 026	0. 169	0. 038	0. 195	0. 226	-0. 192
峰度系数 kurtosis	0. 454	0. 129	1. 094	-0. 288	0. 191	0. 611	-0. 197

注：GCALL：线粒体基因组 GC 含量；GCPro：蛋白质编码区 GC 含量；GCrRNA：核糖体 RNA 区 GC 含量；GC1：密码子第 1 位 GC 含量；GC2：密码子第 2 位 GC 含量；GC12：第 1 和 2 位 GC 含量；GC3：密码子第 3 位 GC 含量，下同。

Note：GCALL：GC complete mitochondrial genome；GCPro：GC content in the protein encoding region；GCrRNA：GC content of RNA encoding region；GC1：GC content in the first position of codons；GC2：GC content in the second position of codons；GC12：GC content in the first and second position of codons；GC3：GC content in the third position of codons，the same as below.

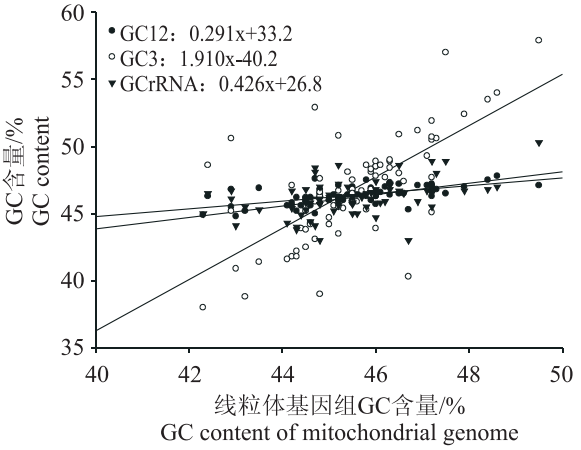


图 1 GCALL与GCrRNA、GC12和GC3的相关性

Fig. 1 Correlation of GCALL with GCrRNA, GC12 and GC3

表 3 线粒体基因组 GC 含量在不同区域的分布

Tab. 3 Distribution of GC contents of Mt genomes in different region

范围 /% contents range	GCALL	GCPro	GC1	GC2	GC3
37. 5 ~ 40. 0	0	0	0	0	3
40. 0 ~ 42. 5	2	0	0	69	7
42. 5 ~ 45. 0	23	11	0	5	10
45. 0 ~ 47. 5	44	46	0	0	21
47. 5 ~ 50. 0	5	15	15	0	19
50. 0 ~ 52. 5	0	2	50	0	9
52. 5 ~ 55. 0	0	0	9	0	3
55. 0 ~ 57. 5	0	0	0	0	1
57. 5 ~ 60. 0	0	0	0	0	1

2. 2 基因组 GC 参数相关性分析

一般而言，线粒体蛋白编码区（13 个蛋白编码基因区域之和）与整个线粒体基因组的 GC 含量最为相似。因为线粒体基因组数据大部分由蛋白编码区组成，所以线粒体蛋白编码区的 GC 含量反映了基因组的 GC 含量。GCAll 与 GCrRNA、GC12 和 GC3 的相关性及线性回归关系如图 1 所示。在回归公式中，GC12 的斜率为 0. 291（r = 0. 385，$P < 0.01$），GCrRNA 的斜率为 0. 426（r = 0. 205，$P < 0.01$），说明 GC12、GCrRNA 对 GCAll 变化的影响相关性较弱，且其 GC 含量的波动相对较小。而 GC3 的斜率大于 1，为 1. 910（r = 0. 478，$P < 0.01$），说明 GC3 与 GCALL 之间的相关性。

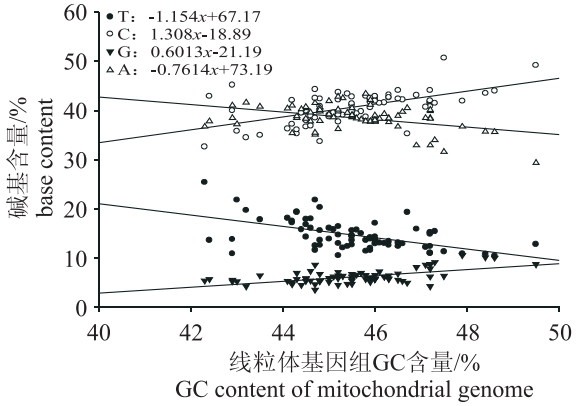


图 2 GCALL与密码子第3位各碱基相关性

Fig. 2 Correlation of GCALL with the base composition of the third codon position

2.3　不同密码子位置的碱基分布

表 4 显示不同密码子位置碱基含量的波动范围。4 种碱基在第 1，2，3 位密码子的波动范围分别为 2.70% ~3.80%，0.60% ~2.10%，7.00% ~18.10%；在鸟类线粒体基因组中，相对于其它两个位置的密码子，第 3 位密码子的波动范围较大，而第 3 位密码子碱基 C 的含量在所有密码子的不同碱基中波动范围最大（18.10%）。由表 4 可知，在密码子第 1 位和第 2 位的碱基含量变化较小，而第 3 位碱基含量变化较大。因此，对密码子第 3 位各碱基与整个线粒体基因组 GC 含量进行回归分析，图 2 显示密码子第 3 位碱基 C 和碱基 T 的斜率分别为 1.308（$r = 0.328$，$P < 0.01$）和 −1.154（$r = 0.292$，$P < 0.01$），而密码子第 3 位碱基 G 和碱基 A 的斜率分别为 0.601（$r = 0.351$，$P < 0.01$）和 −0.7614（$r = 0.182$，$P < 0.01$），碱基 C 和碱基 T 的绝对斜率远大于碱基 G 和碱基 A。因为密码子第 1 位与第 2 位的碱基相对恒定，在基因组结构紧凑的鸟类线粒体基因组中，占有绝大部分区域的线粒体蛋白编码区的 GC 含量反映了基因组的 GC 含量，所以线粒体基因组中的碱基变化主要由密码子第 3 位 C 和 T 碱基变化造成。

表 4　不同密码子位碱基含量波动范围

Tab.4　Variability of base contents at 3 codon positions　%

碱基 base	密码子第 1 位 the first codon position			
	最小值 minimum	最大值 maximum	平均值 mean	范围 range
腺嘌呤 adenine	27.0	29.7	28.5	2.7
胸腺嘧啶 thymine	19.2	23.0	20.62	3.8
胞嘧啶 cytosine	27.1	30.3	28.69	3.2
鸟嘌呤 guanine	20.8	24.1	22.19	3.3
碱基 base	密码子第 2 位 the second codon position			
	最小值 minimum	最大值 maximum	平均值 mean	范围 range
腺嘌呤 adenine	17.7	18.6	18.09	0.9
胸腺嘧啶 thymine	39.1	40.9	40.06	1.8
胞嘧啶 cytosine	28.0	30.1	29.12	2.1
鸟嘌呤 guanine	12.4	13.0	12.74	0.6
碱基 base	密码子第 3 位 the third codon position			
	最小值 minimum	最大值 maximum	平均值 mean	范围 range
腺嘌呤 adenine	29.3	43.4	38.53	14.1
胸腺嘧啶 thymine	10.4	25.4	14.64	15.0
胞嘧啶 cytosine	32.6	50.7	40.66	18.1
鸟嘌呤 guanine	3.5	10.5	6.18	7.0

2.4　ND3 基因胞嘧啶插入现象

目前在多数的鸟类都发现 ND3 基因插入一个胞嘧啶核苷酸。表 5 显示本研究中各个目中 ND3 基因存在胞嘧啶插入现象的物种数目。在古颌超目中，所有的物种都存在 ND3 基因存在胞嘧啶插入现象，而在今颌超目中，雨燕目（Apodiformes）、鸻形目（Charadriiformes）、鹳形目（Ciconiiformes）、隼形目（Falconiformes）、鸡形目（Galliformes）、鹈形目（Pelecaniformes）和雀形目（Passeriformes）都存在不携带胞嘧啶插入的物种。而且在已测定线粒体基因组全序列的 9 个雀形目物种中都不存在胞嘧啶的插入现象。

表 5　鸟类 ND3 基因胞嘧啶插入现象分布

Tab.5　Comparison of extra cytosine insertion in ND3 gene for avian taxa

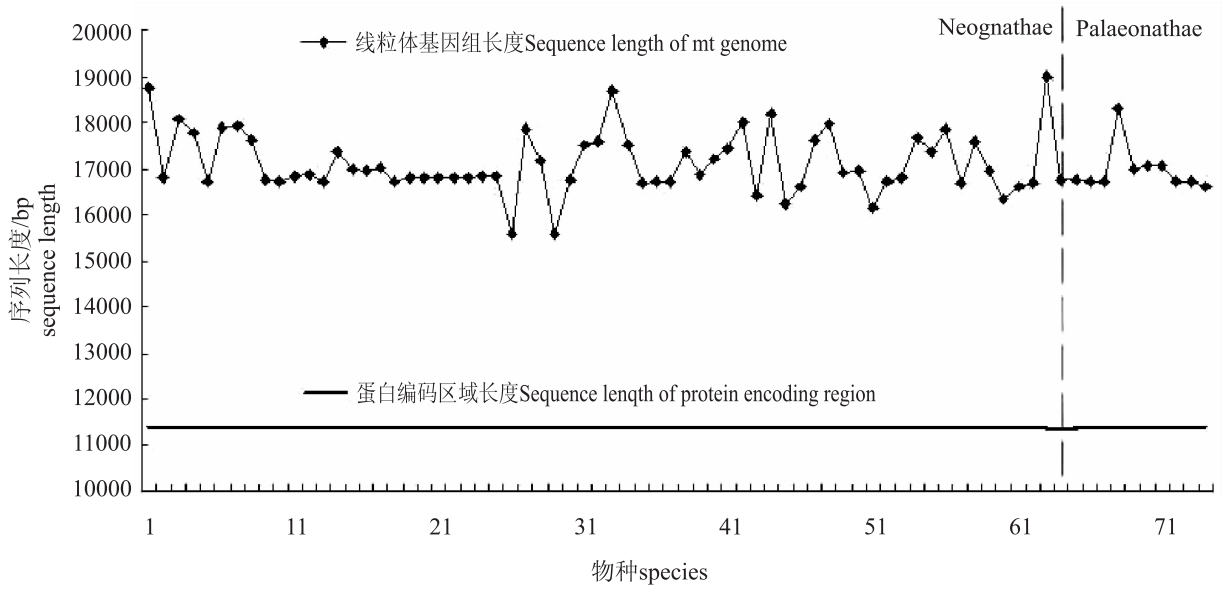
	目 Order	No. Spe	C Insertion	
			+	−
今颌超目 Neognathae	雁形目 Anseriformes	6	6	0
	雨燕目 Apodiformes	1	0	1
	鸻形目 Charadriiformes	4	3	1
	鹳形目 Ciconiiformes	6	4	2
	隼形目 Falconiformes	7	4	3
	鸡形目 Galliformes	12	11	1
	潜鸟目 Gaviiformes	2	2	0
	鹤形目 Gruiformes	2	2	0
	雀形目 Passeriformes	9	0	9
	鹈形目 Pelecaniformes	1	0	1
	红鹳目 Phoenicopteriformes	1	1	0
	鴷形目 Piciformes	2	2	0
新额超目 Neognathae	鹱鹬目 Podicipediformes	2	2	0
	鹪形目 Procellariiformes	2	2	0
	鹦形目 Psittaciformes	2	2	0
	企鹅目 Sphenisciformes	2	2	0
	鸮形目 Strigiformes	1	1	0
	蜂鸟目 Trochiliformes	1	1	0
古颌超目 Palaeognathae	几维目 Apterygiformes	1	1	0
	鹤鸵目 Casuariiformes	2	2	0
	恐鸟目 Dinornithiformes	3	3	0
	美洲鸵目 Rheiformes	2	2	0
	鸵形目 Struthioniformes	1	1	0
	［共鸟］型目 Tinamiformes	2	2	0

注：No. Spe：本研究中该目中的物种数目；C Insertion（+）：存在胞嘧啶插入现象的物种数目；C Insertion（−）：不存在胞嘧啶插入现象的物种数目；

Note：No. Spe：the number of total species；C Insertion（+）：the number of species with extra base；C Insertion（−）：the number of species without extra base

2.5 基因组和蛋白编码区序列长度

图 3 为 74 种鸟类线粒体基因组和蛋白编码区域的序列长度分布示意图，横坐标随着蛋白编码区域的序列长度逐渐升高。横坐标 1 ~63 为今颌超目中的物种，64 ~74 为古颌超目中的物种。由图 3 显示，线粒体基因组全序列的长度波动性较大，分布范围由 15 574 bp（太平洋潜鸟 *Gavia pacifica*）到 18 967 bp（黑眉信天翁 *Thalassarche melanophris*）。而蛋白编码区域的序列长度非常恒定，长度最短的物种为小美洲鸵（*Pterocnemia pennata*，11 328 bp），长度最长的物种为日本鹌鹑（*Coturnix japonica*）、黑眉信天翁（*Thalassarche melanophris*）和美洲潜鸭（*Aythya Americana*），长度皆为 11 373 bp。虽然这三个物种在蛋白编码区域的长度一致，但是其线粒体基因组序列长度分别为 18 967 bp（黑眉信天翁）、16 697 bp（日本鹌鹑）和 16 616 bp（美洲潜鸭）。说明线粒体基因组中蛋白编码区域与线粒体基因组的总长度不存在关联。



注：横坐标1~63为今颌超目中的物种，64~74为古颌超目中的物种。
Note: No.1~63 denotes *Neognathae*; No.64~74 denotes *Palaeognathae*.

图 3 线粒体基因组和蛋白编码区域的序列长度分布示意图
Fig. 3 Distribution of sequence length of whole Mt genomes and protein coding region

3 讨论

本研究结果表明线粒体基因组的 GC 含量和编码区的 GC 含量的变化主要由 GC3 的变化引起。在细菌、植物叶绿体和节肢动物中也存在这种现象[17~19]。虽然密码子第 3 位的碱基差别较大，但是整个线粒体基因组的 GC 含量却差别不大。与内含子数目较少和基因组结构同样紧凑的细菌基因组（GC 含量范围 25% ~75%）比较[20]，线粒体基因组 GC 含量的波动范围却十分狭窄。

通常认为，第 1 及第 2 位置密码子面临的选择压力较大，而第 3 位置密码子及非编码区所受的选择压力较小，这种差异是造成基因组碱基分布不均衡的主要原因[21]。而第 2 位密码子的各种碱基在不同物种的碱基组成中相对比较恒定（0.60% ~2.10%），一般认为，如果碱基含量波动范围作为衡量净化选择的指标，则第 2 位密码子受到的净化选择的强度远大于第 1 和第 3 位。而由于第 3 位密码子的摆动性，其各种碱基波动范围的大于另外两个位置的碱基。但是密码子第 3 位的碱基 G 与其它 3 种碱基不同，始终保持在比较低的水平（小于 10%），这种现象在哺乳动物中也被发现[22]。

鸟类线粒体基因组在不同物种间呈现出异常的碱基组成。本研究通过分析 74 个鸟类动物线粒体基因碱基组成，结果显示，造成线粒体的碱基波动的主要原因是密码子第 3 位 C 和 T 两个碱基的变化。SCHMITZ 等首先在灵长类动物中发现这种 C 和 T 之间的变化，并认为其产生的变异率升高现象，最终对氨基酸的使用造成影响[23]。此

外，对 69 个哺乳动物线粒体基因组的研究中，同样观察到类似 C 和 T 变化的现象[22]。在第 3 位密码子位置上，碱基 C 的含量由 32.6% 上升到 50.7%，碱基 T 的含量伴随着 C 的升高而降低。造成这种基因组局部碱基不均衡的重要原因是由于密码子的摆动性、氨基酸对碱基的限制及相应的密码子使用频率偏好造成[17]。SUEOKA 认为第 3 位密码子达到特别高的水平的现象意味着突变压力对基因组内的这些核苷酸能够产生非常强的影响。不同位置上的碱基组成存在关联的现象在不同的研究中都有发现[24,25]，并且利用线性回归分析的方法检测到在不同物种间（或者同一基因组中不同区域）第 1 位和第 2 位密码子存在相对中性情况[26]。关于第 3 位密码子的变化是否中性一直存在争议。1988 年 SUEOKA 根据分子中性进化理论提出方向性突变压力学说（Directional Mutation Pressure Theory）[25]。该学说认为突变是决定 DNA 的 GC 含量的主要因素而不是选择。方向性突变压力学说认为 GC 含量在第 3 位密码子的变化是中性的。但是有学者提出不同的观点认为第 3 位密码子存在翻译选择效应，翻译选择效应对于密码子的 GC 含量的变化直接造成影响[21]。包括原核和真核生物的多数物种存在同义密码子选择现象是为了更好实现高效、准确的翻译。

在许多鸟类 ND3 基因的特定位点都发现存在一个插入的胞嘧啶，通常认为，这个额外的核苷酸在翻译过程中是不参与表达的，而是被 RNA 的自我编辑后剪切掉，使得基因功能得以恢复，从而避免了移码突变导致的 ND3 基因转录提前终止[11]。原鸡（*Gallus gallus*，NC _ 001323）是 GenBank 细胞器基因组资源数据库公布的 12 个鸡形目物种中，唯一被发现不存在胞嘧啶插入的物种，但是在其他的不同作者的工作中[11,28]，证实原鸡中亦存在此插入碱基[14]。在基于线粒体基因组全序列构建的鸟类系统发生树中[29]，未有胞嘧啶插入的物种聚集形成一个分支，而存在胞嘧啶插入的雁形目、鸡形目和古颌超目中的物种形成另外的分支。推测 ND3 基因不存在胞嘧啶插入的现象可能是鸟类快速辐射进化中线粒体基因组内遗留物或者是自然选择的结果造成。

[参考文献]

[1] GARESSE R, VALLEJO C G. Animal mitochondrial biogenesis and function: a regulatory cross-talk between two genomes [J]. Gene, 2001, 263 (1-2): 1-16.

[2] BOORE, J. L. Animal mitochondrial genomes [J]. Nucleic Acids Research, 1999, 27: 1767-1780.

[3] 李庆伟，马飞. 鸟类分子进化与分子系统学 [M]. 北京：科学出版社，2007.

[4] NECSULEA A, LOBRY JR. Revisiting the directional mutation pressure theory: the analysis of a particular genomic structure in Leishmania major [J]. Gene, 2006, 385: 28-40.

[5] GALTIER N, PIGANEAU G, MOUCHIROUD D, et al.. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis [J]. Genetics, 2001, 159: 907-911.

[6] BELOZERSKY AN, SOIRIN AS. A correlation between the compositions of deoxyribonucleic and ribonucleic acids [J]. Nature, 1958, 182 (4628): 111-112.

[7] DAUBIN V, LERAT E, PERRIERE G. The source of laterally transferred genes in bacterial genomes [J]. Genome Biology, 2003, 4 (9): R57.

[8] DAUBIN V, PERRIERE G. G+C3 structuring along the genome: a common feature in prokaryotes [J]. Molecular Biology and Evolution, 2003, 20 (4): 471-483.

[9] BERNARDI G. Codon usage and genome composition [J]. Molecular Evolution, 1985, 22 (4): 363-365.

[10] SUEOKA N. Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A=T and G=C [J]. Molecular Evolution, 1999, 49 (1): 49-62.

[11] MINDELL DP, SORENSON MD, DIMCHEFF DE. An extra nucleotide is not translated in mitochondrial ND3 of some birds and turtles [J]. Molecular Biology and Evolution, 1998, 15 (11): 1568-1571.

[12] PATON T, HADDRATH O, BAKER AJ. Complete mitochondrial DNA genome sequences show that modern birds are not descended from transitional shorebirds [J]. Philosophical Transactions of the Royal Society B: Biological Sciences, 2002, 269 (1493): 839-846.

[13] HADDRATH O, BAKER AJ. Complete mitochondrial DNA genome sequences of extinct birds: ratite phylogenetics and the vicariance biogeography hypothesis [J]. Philosophical Transactions of the Royal Society B: Biological Sciences, 2001, 268 (1470): 939-945.

[14] 孙毅，马飞，肖冰，等. 鸮形目两种鸟类线粒体基因组全序列测定与比较研究 [J]. 中国科学（C 辑），2004, 34 (6): 527-536.

[15] HALL, T. A. BIOEDIT. a user-friendly biological se-

quence alignment editor and analysis program for Windows 95/98/NT [M]. Nucleic Acids Symposium Series, 1999.

[16] STAJICH J E, BLOCK D, BOULEZ K, et al.. The Bioperl toolkit: Perl modules for the life sciences [J]. Genome Research, 2002, 12: 1611 - 1618.

[17] 武伟, 刘洪斌, 张泽, 等. 节肢动物线粒体基因组碱基组成特征分析 [J]. 生物信息学, 2007, 5 (3): 102 - 105.

[18] KUSUMI J, TACHIDA H. Compositional properties of green-plant plastid genomes [J]. Molecular Evolution, 2005, 60 (4): 417 - 425.

[19] MORAN NA, MICROBIAL MINIMALISM. Genome reduction in bacterial pathogens [J]. Cell, 2002, 108 (5): 583 - 586.

[20] LOBRY JR, SUEOKA N. Asymmetric directional mutation pressures in bacteria [J]. Genome Biology, 2002, 3 (10): 1 - 14.

[21] 钟东, 赵贵军, 张振书, 等. 基因组内碱基分布整体均衡与局部不均衡的研究进展 [J]. 遗传, 2002, 24 (3): 351 - 355.

[22] GIBSON A, GOWRI - SHANKAR V, HIGGS PG, et al.. A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods [J]. Molecular Biology and Evolution, 2005, 22 (2): 251 - 264.

[23] SCHMITZ J, OHME M, ZISCHLER H. The complete mitochondrial sequence of Tarsius bancanus: evidence for an extensive nucleotide compositional plasticity of primate mitochondrial DNA [J]. Molecular Biology and Evolution, 2002, 19 (4): 544 - 553.

[24] D'ONOFRIO G, BERNARDI G. A universal compositional correlation among codon positions [J]. Gene, 1992, 110 (1): 81 - 88.

[25] SUEOKA N. Directional mutation pressure and neutral molecular evolution [M]. Proceedings of the National Academy of Sciences of the United States of America, 1988, 85 (8): 2653 - 2657.

[26] SUEOKA N, KAWANISHI Y. DNA G + C content of the third codon position and codon usage biases of human genes [J]. Gene, 2000, 261 (1): 53 - 62.

[27] PALACIOS C, WERNEGREEN JJ. A strong effect of AT mutational bias on amino acid usage in Buchnera is mitigated at high-expression genes [J]. Molecular Biology and Evolution, 2002, 19 (9): 1575 - 1584.

[28] NISHIBORI M, HAYASHI T, TSUDZUKI M, et al.. Complete sequence of the Japanese quail (Coturnix japonica) mitochondrial genome and its genetic relationship with related species [J]. Animal Genetics, 2001, 32 (6): 380 - 385.

[29] SLACK KE, DELSUC F, MCLENACHAN PA, et al.. Resolving the root of the avian mitogenomic tree by breaking up long branches [J]. Molecular Phylogenetics and Evolution, 2007, 42 (1): 1 - 13.