

研究论文

蜜蜂全基因组中微卫星的丰度及其分布^{*}

魏朝明^{1**} 孔光耀¹ 廉振民^{1,2} 刘 慧¹ 范永文¹ 张 慧³

(1 陕西师范大学生命科学学院 西安 710062; 2 延安大学 延安 716000; 3. 陕西省微生物研究所 西安 710043)

Abundance and distribution of microsatellites in the entire *Apis mellifera* genome. WEI Zhao-Ming^{1**}, KONG Guang-Yao¹, LIAN Zhen-Min^{1,2}, LIU Hui¹, FAN Yong-Wen¹, ZHANG Hui³ (1. College of Life Sciences, Shaanxi Normal University, Xi'an 710062, China; 2 Yan'an University, Yan'an 716000, China; 3. Shaanxi Institute of Microbiology, Xi'an 710043, China)

Abstract In order to improve related studies on genetics of *Apis mellifera*, microsatellites(SSRs) of the entire *A. mellifera* genome with 1~6 nucleotide motifs were found out. Abundance and distribution of microsatellites across the *A. mellifera* genome were analyzed and compared among various chromosomes. The results showed that the frequency of SSRs is 1/0.804 kb in *A. mellifera* genome, and dinucleotide repeats(26.86%) motifs appeared to be the most abundant type, while the hexanucleotide, mononucleotide, trinucleotide, tetranucleotide and pentanucleotide repeats are 24.74%, 22.19%, 13.65%, 10.98% and 2.59%, respectively. Meanwhile, the A-rich and T-rich repeats are predominant, whereas G-rich and C-rich repeats are rare. The frequency of SSRs on chromosome 4, 3 and 1 is much higher than that on chromosome 11, 14 and 12.

Key words *Apis mellifera*, genome, microsatellite, abundance, distribution

摘 要 查找出蜜蜂基因组中由1~6个碱基重复单元组成的简单序列重复,分析蜜蜂基因组中微卫星的分布频率,并比较其在各染色体中的分布频率。微卫星在蜜蜂基因组中的分布频率为1/0.804 kb,其中二碱基重复序列占26.86%,是最丰富的重复单元,而六、一、三、四、五碱基重复单元序列分别占24.74%,22.19%,13.65%,10.98%,2.59%。同时发现富含A和T碱基的微卫星占主导地位,富含G和C碱基的微卫星数量较少。第4、3条染色体微卫星分布频率较高,而第11、14、12条染色体微卫星分布频率较低。

关键词 蜜蜂,基因组,微卫星,丰度,分布

微卫星又称简单序列重复(simple sequence repeats)、短串联重复(short tandem repeats)或简单序列长度多态性(simple sequence repeats polymorphism)^[1]。是指以1~6个碱基为基本重复单元的串联重复序列,其长度大多在100 bp以内,它们广泛存在于各类真核生物基因组中,原核生物基因组中也含有少量的微卫星。尽管微卫星在基因组中已经被人们广泛关注,但人们对其产生、进化和在基因组中的作用的研究都才刚刚开始。大多数研究者认为微卫星是在DNA复制和修复时,滑动链与互补链碱基错配

导致移码突变造成的^[2]。

微卫星分子标记是目前研究生物群体遗传结构和变异的极有价值的分子遗传标记。比RFLP、RAPD、AFLP表现出更高的灵敏性^[3]。目前,微卫星标记已被广泛应用到基因连锁与遗传图谱构建^[4]、遗传多样性研究^[5]、谱系和发育研究^[6]、疾病检测^[7]以及品种鉴定上^[8]。但目

^{*} 陕西省自然科学基金资助项目(2005C113)。

^{**}通讯作者, E-mail: weizhaoming@snnu.edu.cn

收稿日期: 2006-07-06, 修回日期: 2006-07-28, 接受日期: 2007-

前已知的微卫星位点很有限,这限制了微卫星标记的大量使用。近年来利用构建部分基因组文库和富集文库法筛选微卫星位点已经得到了一些微卫星位点。同时随着生物信息学的发展,利用公共数据库获得微卫星位点的方法也随之出现。

蜜蜂是一种很有经济价值的农业昆虫,能产蜜和传粉。也是研究免疫,过敏反应,精神疾病和 X 染色体疾病等人类遗传疾病的模式动物^[16]。同时许多学者对蜜蜂的社会生活和行为特点很感兴趣。自从在蜜蜂的基因组中发现微卫星以来,已被广泛应用于蜜蜂群体遗传学、社会行为及遗传育种研究中。目前已有学者对蜜蜂 EST 中的微卫星进行了分析^[9],但迄今还没有关于蜜蜂全基因组中微卫星的丰度和分布的研究报告。通过分析微卫星在蜜蜂全基因组中的分布能够为进一步研究蜜蜂微卫星提供框架,并且通过与其他物种微卫星丰度和分布比较,研究微卫星在物种间的差异,为进一步揭示其产生机制、进化和功能提供资料。本研究主要通过查找和分析全基因组中以 1~6 个碱基为重复单元的微卫星,分析微卫星的丰度及其在基因组不同区域的分布规律。

1 研究方法

1.1 蜜蜂基因组序列

蜜蜂全基因组序列由 Genbank <http://www.ncbi.nlm.nih.gov/> 下载成 FASTA 格式文件,共包括 16 条染色体共约 181.354 Mb。

1.2 数据分析

在 <http://www.maizemap.org/> 下载了一个 SSRfinder 程序,该程序由一系列的 Perl5 脚本组成。通过 SSRfinder 程序,可以查找蜜蜂基因组中的所有的完美微卫星。结果以文本格式输出。输出的微卫星资料包括 DNA 的 2 条互补链,如 AT 包含 TA,CGA 包含 GCT。同时要求微卫星长度不小于 12 bp^[9,10,12],即要求单碱基 12 次重复以上、二碱基 6 次重复以上、三碱基 4 次重复以上、四碱基 3 次重复以上、五碱基和六碱基 2 次重复以上。选择微卫星的基本原理就是

前一个微卫星被单一碱基隔开后即开始查找下一个微卫星。

随后利用我们自己设计的一系列由 Perl5 脚本组成的程序对通过 SSRfinder 程序查找所得的微卫星进行整理,对不同类型的微卫星进行分类并统计其数量,用于最后的数据分析比较。

2 数据分析结果

2.1 微卫星在基因组及其在各染色体上的分布频率

在蜜蜂基因组 181.354 Mb 的碱基中,总共搜索到微卫星 225 556 个微卫星,微卫星出现的频率是 1/0.804 kb,比以往的研究微卫星频率的结果高^[11]。16 条染色体中第 4,1,3 条微卫星分布频率较高,分别为 1/0.670,1/0.681,1/0.689 kb。第 11,14,12 条中微卫星频率较低,分别为 1/0.873,1/0.872,1/1.864 kb。

在 6 种基本重复单元的微卫星重复中,以二碱基重复序列丰度最高(1/2.993 kb),随后依次是六碱基重复序列(1/3.250 kb)、单碱基重复序列(1/3.795 kb)、三碱基重复序列(1/5.892 kb)、四碱基重复序列(1/7.325 kb)、五碱基重复序列(1/31.048 kb)。分别占有微卫星的百分比依次为:26.86,24.74,22.19,13.65,10.98,2.59%。(图 1)在 16 条染色体中,除了第 7 条染色体中六碱基重复序列(2 887 个)比二碱基重复序列(2 829)稍多外,其余 15 条染色体上 6 种基本重复单元的微卫星的数量均与以上顺序一致。

2.2 不同基本重复单元微卫星的各种类型分布频率分析

单碱基重复序列中,(A)_n和(T)_n占 97%,而(C)_n和(G)_n仅占 3%,这与前人关于单碱基重复序列的研究结果一致。二碱基重复序列中,根据碱基互补原则,共分成 4 种类型,(AG)_n, (AT)_n, (CA)_n, (CG)_n。其中(AG)_n出现的频率最高(1/5.946 kb),其次是(AT)_n和(CA)_n,而(CG)_n在蜜蜂基因组中出现的频率很低。在三碱基重复序列中 TAA 出现频率最高(1/28.393

kb), CTT, TCC, ATA, GAG, TCT 出现的频率也都较高, 而 CGC, CAC 等重复序列较少, 甚至有的染色体都不出现(图 1)。

在四碱基重复序列中, GAAA 出现频率最高(1/32.652 kb), 随后依次是 TAAA, ATAA, TAAT, AGGG, AGAA 等, 其余的则出现的频率较小(图 1)。而在五碱基重复序列和六碱基重复

序列中, GAAAA (197. 292 kb) 和 GAAAAA (1 / 49. 686 kb) 占主导地位, 其他类型的微卫星重复序列如 TAAAA, TAAAAA, TAAATA, ATTAAA, ATAAAA, AGAAAA, GAAAGA 也都占有较大的比例, 很多其它的排列组合在基因组中出现的频率很小, 在某些染色体上甚至不出现。

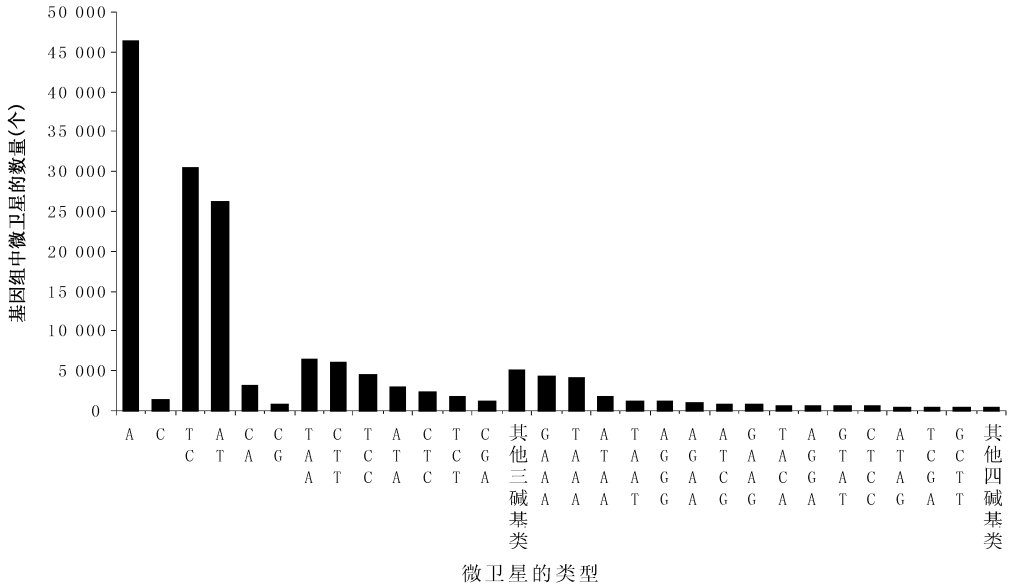


图1 整个基因组中一、二、三、四碱基重复的各种微卫星类型的数量比较

3 讨论

本研究在蜜蜂全基因组 181.354 Mb 中共搜索到微卫星 225 556 个, 分析得到蜜蜂基因组中微卫星位点的分布频率为 $1/0.804\text{ kb}$, 比以往研究者的研究其他真核动物微卫星位点的分布频率 ($1/4 \sim 6\text{ kb}$) 结果要高^[11]。但比李斌等分析蜜蜂 EST 数据库的得到的微卫星位点分布频率 ($1/0.52\text{ kb}$) 要低一些^[9], 但差异不是很大。分析原因, 我们认为主要有以下几点原因: (1) 不同研究者对微卫星的定义不同, 这也是目前研究微卫星很有争议的地方, 本研究采用了近来被大多数研究者认可的以微卫星长度不小于 12 bp 为判定标准^[9, 10, 12], 因而结果具有普遍适用意义。(2) 近期通过富集文库法得到微卫星位点的技术发展迅速, 使 Genbank 中的微卫星快速增长, 这是以往研究者所不能预测到的。

(3)再者,研究者普遍认为微卫星是由碱基错配导致移码突变造成序列重复,而本研究分析的整个基因组中微卫星的分布,大量的内含子序列被包含在其中,内含子为非编码序列,受到的选择压力较小,容易积累变异,因而错配几率较高,也是为微卫星频率增加的原因。

Varshney 等通过研究多种谷类植物 EST 中微卫星的分布^[11], 不计算单碱基重复序列, 发现三碱基重复序列占主导地位。并认为这是 DNA 复制时偏爱三碱基重复的微卫星, 是因为三碱基是密码子的组成形式, 而减少或增加一个碱基会增加错配引起移码突变的机会。而李斌等研究蚊子全基因组中微卫星分布发现, 六碱基重复序列占主导地位, 其次是二碱基重复序列^[12]。本研究分析发现蜜蜂全基因组中二碱基重复序列占主导地位, 其次是六碱基重复序列。由此可见微卫星类型在不同物种间分布

有差异。通过基因组研究物种间不同类型微卫星丰度,亦没有得到一个稳定的结果,分析原因可能是因为目前研究基因组都是通过数据库得到基因组序列信息,但不同物种的基因组大小不同,碱基组成含量也不同,因而微卫星丰度也不同。基因组越小,单碱基重复序列越多,随基因组增大,其他类型微卫星逐渐占主导地位,而GC含量越高的序列中微卫星也越少^[13]。有研究者发现,在人类基因组中一些三碱基重复序列的增加与人类的一些遗传疾病如脆X综合征的发生有关系^[14],因此不同重复单元微卫星的增加也有可能在不同物种的某些特殊功能有着密切联系。

同时,在微卫星各种重复单元中,以富含A(或T)的重复单元占主导地位,如(A)_n和(T)_n占单碱基重复序列的97%,二碱基重复序列中(AT)_n的数量是(GC)_n的32.7倍。三碱基重复序列中TAA, GAA占主导地位, GAAA, TAAA, ATAA, TAAT, GAAAA, TAAAA, GAAAA, TAAAA, TAAATA等都是各不同基本重复单元中占主导地位的重复类型。由以上分析结果可以清楚地看出微卫星偏爱富含A或T的序列,因而富含G或C的序列微卫星出现的频率较低。碱基排列的组成决定着基因所编码的蛋白质,因而不同碱基组成的微卫星也会编码不同的蛋白质,致使微卫星在基因组中行使着不同的功能,对富含A或T碱基微卫星的进一步的研究将推进人们对微卫星功能的深入了解。这有待以后的科学实验来证实。

对蜜蜂基因组中各染色体上微卫星的分布频率分析发现,微卫星在各染色体间的分布频率略有差异,以第4, 1, 3条微卫星分布频率较高,分别为1 0.670, 1 0.681, 1 0.689 kb。第11, 14, 12条中微卫星频率较低,分别为1 0.873, 1 0.872, 1 0.864 kb。而不同基本重复单元的微卫星在各条染色体中的分布较为一致,基本都是二碱基重复序列占主导,随后依次是六碱基重复序列、单碱基重复序列、三碱基重复序列、

四碱基重复序列,五碱基重复序列。由此可见蜜蜂基因组各染色体上微卫星的分布主要差异在微卫星的分布频率,分布频率的不同一方面说明微卫星不是随机出现的,进一步的探讨其产生机制和功能将是下一步的研究方向。另一方面说明蜜蜂各个染色体间存在结构和功能的差异。蜜蜂的性别决定问题一直是遗传学家有争议的问题,遗传学家也提出了各种假说^[15, 17]。利用微卫星位点多态性或定位基因,将有利于加深对蜜蜂染色体功能的研究,同时也将为解决染色体是不是在蜜蜂性别决定起作用以及哪些具体基因决定了蜜蜂性别等问题提供理论依据。

参 考 文 献

- 1 Chin E. C. L., Senior M. L., Shu H., Smith J. S. C. *Genome*, 1996, **39**: 866~873.
- 2 Heale S. M., Petes T. D. *Cell*, 1995, **83**(4): 539~545.
- 3 Powell W., Morgante M., Andre C., Hanafey M., Vogel J., et al. *Mol. Breed.*, 1996, **2**: 225~238.
- 4 Miyao A., Zhong H. S., Morna L. Yano M., Yamamoto K., et al. *DNA Research*, 1996, **3**: 233~238.
- 5 Nomi A. T., Bradley D. G., Cunningham E. P. *Aquaculture*, 1999, **180**: 247~264.
- 6 Nicolas B. *Genetics*, 1998, **148**(4): 1 893~1 906.
- 7 Hee S. K., Byung L. L. *Cancer Letters*, 2001, **164**: 61~68.
- 8 Michael L. B., Guillermo O. *Mol. Ecol. Notes*, 2003 Primer note: 1~3.
- 9 李斌, 夏庆友, 鲁成, 周泽扬. 遗传学报, 2004, **31**(10): 1 089~1 094.
- 10 Subramanian S., Mishra R. K., Singh L. *Gen. Biol.*, 2003, **4**: 13.
- 11 Varshney R. K., Thiel T., Stein N., Langridge P., Graner A. *Cell. Mol. Biol. Lett.*, 2002, **7**: 537~546.
- 12 Yu Q. Y., Li B., Li G. R., Fang S. M., Yan H., et al. 生物化学与生物物理进展, 2005, **32**(5): 435~441.
- 13 Lin S., Notley-McRobb L., Lim M., Carter D. A. *Fung. Gen. Biol.*, 2004, **41**: 1 025~1 036.
- 14 Christian S. *Curr. Biol.*, 1998, **8**(4): 132~134.
- 15 曾志将, 饶波, 谢宪兵, 颜伟玉. 昆虫知识, 2003, **40**(3): 208~211.
- 16 黄智勇. 昆虫知识, 2006, **44**(1): 5~9.
- 17 王丽华. 昆虫知识, 2006, **44**(1): 25~29.