第35卷 第5期 海 洋 与 湖 沼 Vol. 35, No. 5
2004年9月 OCEANOLOGIA ET LIMNOLOGIA SINICA Sep., 2004

中国对虾(*Fenneropenaeus chinensis*)基因组微卫星特征分析*

高 焕 刘 萍† 孟宪红† 王伟继† 孔 杰†1)

(中国水产科学研究院黄海水产研究所 青岛 266071;中国科学院海洋研究所 青岛 266071;
中国科学院研究生院 北京 100039)

†(中国水产科学研究院黄海水产研究所 青岛 266071)

提要 对中国对虾基因组随机测序,获得了总长度约为641000个碱基的基因组DNA序列,从中找到1362个重复序列。其中,两碱基重复类型的重复数目最多(985个),占重复序列总数目的72.32%;其次是三碱基(149个)和四碱基(102个),分别占重复序列总数目的10.94%和7.49%。另外,六碱基重复34个,单碱基重复50个,五碱基重复5个,分别占重复序列总数目的2.50%、3.67%、0.37%。在单碱基重复类型中,重复拷贝类别为A的重复数目最多;两碱基重复类型中,AT重复数目最多,其次是AC和AG;三碱基重复类型中以AAT重复拷贝类别最多,其次是AAG和ATC;四碱基重复类型中,AGAT重复数目最多;五碱基重复类型只发现了AGAGA、GAGGC、TCTTC和TTTCT四种重复拷贝类别;六碱基重复中以ATTATC重复数目最多。其中一些序列已经提交GeneBank注册,注册号为AY545898-AY545913。中国对虾基因组二碱基重复类型中以不完全(Imperfect)形式的微卫星序列为主,其中GC重复拷贝类别的重复数目很少。利用8对微卫星引物对60个个体遗传多样性分析,共获得了60个等位基因,因此认为微卫星技术在中国对虾基因组研究中具有较好的应用前景。

关键词 微卫星,中国对虾,基因组

中图分类号 Q75

微卫星(Microsatellites, MS)DNA由Miesfeld等(1981)首次发现,又称短串联重复(Short tandem repeat, STR)或简单序列重复(Simple sequence repeat, SSR),是指以1—6个核苷酸为基本重复单位的串联重复序列。它们广泛分布于各种生物的基因组中,尤其是真核生物基因组中,主要以2—3个碱基重复类型为主,如$(CA)_n$、$(GT)_n$、$(CAG)_n$等。其长度由重复单位的拷贝数决定,是一类呈高度多态的遗传标记,不仅可用于基因组遗传连锁图的构建以及基因的定位与克隆,而且可用于遗传性疾病的连锁分析和基因诊断、亲权分析或品种鉴定、农作物及动物育种、进化研究(Beheregaray *et al*, 2003)等领域,因而已经成为当前生物学研究中的一个热点。在海洋生物中,利用微卫星也已经开展了大量的工作,包括等位基因位点的多态性分析,利用微卫星构建连锁图谱和抗病相关标记的连锁分析等等,其中中国对虾的微卫星筛选工作已经展开(徐鹏等, 2001, 2003)。

中国对虾(*Fenneropenaeus chinensis*)属于甲壳纲、十足目、对虾总科、对虾科、明对虾属,是中国近海地区的特有物种,也是我国重要的经济虾类之一。通过对中国对虾基因组RAPD分析(石拓等, 1999, 2001; 刘萍等, 2000; 刘振辉等,

* 国家重点基础研究发展规划(973)资助项目,G1999012007号;国家高技术研究发展计划(863)资助项目,2003AA603021号。高 焕,博士研究生,E-mail:huanmr@163.com

1) 通讯作者:孔 杰,研究员,E-mail:kongjie@sina.com

收稿日期:2004-02-16,收修改稿日期:2004-05-08

2000)、16S rRNA 基因序列多态性的分析(邱高峰等, 2000; 高天翔等, 2003),人们对中国对虾基因组的遗传多样性有了初步的认识。而从 SSR 方面来研究中国对虾基因组及其遗传多样性方面才刚刚起步,虽然已经从中国对虾的基因组中获得了一万多条的序列表达标签(Expressed Sequence Tag, EST)序列(Xiang *et al*, 2002),并从中发现了 229 个微卫星序列(徐鹏等, 2003),以及采用 PCR 的方法利用$(CT)_n$、$(AT)_n$微卫星序列设计引物从中国对虾小片段基因组中筛选微卫星 DNA(徐鹏等, 2001),但对中国对虾各类微卫星及其分布特征方面还知之甚少。作者分析了大量的基因组随机测序结果中微卫星的重复序列,旨在阐明全基因组水平上的微卫星重复序列的分布特征,以期深化对中国对虾基因组的认识,为开发微卫星的标记、开展深入的遗传标记辅助育种工作奠定理论基础。

1　材料与方法

1.1　材料来源

中国对虾(*Fenneropenaeus chinensis*)样品取自黄、渤海自然群体(捕自威海外海)。2001 年 2 月—2003 年 10 月,对其基因组进行随机测序,共获得 3699 个 DNA 克隆序列,经过序列装配,最终得到 1520 个互不重复的克隆序列。每个克隆序列长度从 300 到 1000 个左右碱基不等,代表着总长度约 641000 个碱基。

1.2　统计方法

1.2.1　重复序列 4 个统计术语的定义　为了更好的对重复序列进行有效的统计,对与重复序列特征相关的 4 个术语作以下说明:(1) 重复类型:指重复序列中每个重复单元由几个碱基组成,如单碱基重复、两碱基重复、三碱基重复等等,对于微卫星来讲,共有 6 种重复类型。(2) 重复数目:是指每种重复类型在基因组中的数目。(3) 重复拷贝数:是针对一个重复序列来讲,其核心序列重复的次数,如$(CT)_7$表示这个 CT 重复类型的重复拷贝数为7。(4) 重复拷贝类别:是对重复类型的细化,是指各重复类型中具体由哪些碱基组成,如两碱基重复类型中,CT 和 CG 分别属于不同的重复拷贝类别。

1.2.2　统计的标准　利用生物软件对 1520 个克隆序列中的重复序列进行查找。首先把测序的源文件经 DNASTAR (Version: 5.0) 去除载体序列和污染的序列后装配输出,共输出 1520 个克隆序列。然后利用微卫星查找软件 Tandem Repeats Finder (Version: 2.02) (Benson, 1999) 进行重复序列的初步查找,对于查找到的重复序列再进行人工细化分析。使用的标准如下:14 个或 14 个拷贝以上的单碱基重复序列,7 个或 7 个拷贝以上的 2 碱基重复序列,5 个或 5 个拷贝以上的 3 碱基重复序列,4 个或 4 个拷贝以上的 4 碱基重复序列,3 个或 3 个拷贝以上的 5—6 碱基重复序列。考虑到碱基互补配对和记数拷贝数起始碱基顺序的差异,将同类重复兼并为一种重复代表,1—3 碱基重复可以兼并为以下几种类型(表 1)。

表 1　1—3 碱基重复类型

Tab. 1　Types of 1—3 bp repeats

重复碱基类型		对应碱基
单碱基	A	A、T
	C	C、G
两碱基	AT	AT、TA
	AG	AG、GA、TC、CT
	AC	AC、CA、TG、GT
	GC	GC、CG
三碱基	AAT	AAT、ATA、TAA、TTA、TAT、ATT
	AAC	AAC、ACA、CAA、TTG、TGT、GTT
	AAG	AAG、AGA、GAA、TTC、TCT、CTT
	ATC	ATC、TCA、CAT、TAG、AGT、GTA
	ACG	ACG、CGA、GAC、TGC、GCT、CTG
	ACT	ACT、CTA、TAC、TGA、GAT、ATG
	AGC	AGC、GCA、CAG、TCG、CGT、GTC
	GCC	GCC、CCG、CGC、CGG、GGC、GCG
	AGG	AGG、GGA、GAG、TCC、CCT、CTC
	ACC	ACC、CCA、CAC、TGG、GGT、GTG

四碱基及其以上重复的兼并原则与以上相同,因 4—6 碱基重复的类型较为复杂,而中国对虾基因组中其类型及各类型的拷贝数相对较少,这里不再一一列出。

2　结果

2.1　重复序列种类、数目和相应的百分比

在约 641000 个总碱基长度中共找到 1362 个

重复序列,平均每隔470个碱基就有一个重复序列。统计所有重复类型,以两碱基重复数目最多,为985个,占重复序列总数目的72.32%;其次是三碱基,149个,占10.94%;再次分别是四碱基102个,占7.49%,单碱基50个,占3.67%,六碱基34个,占2.50%,五碱基5个,占0.37%。另外也存在一定数目的六碱基以上的重复类型(表2)。

表2 不同重复类型的重复序列数目及其百分比

Tab.2 The number of repeat sequence and the percentage in different types of repeat

重复类型	单碱基	两碱基	三碱基	四碱基	五碱基	六碱基	其他重复
重复序列数目	50	985	149	102	5	34	37
总计				1362			
各重复类型百分比(%)	3.67	72.32	10.94	7.49	0.37	2.50	2.72

在同类型的重复序列中,各重复拷贝类别占的比例也各不相同。在单碱基重复类型中,以重复拷贝类别为A的重复数目最多,为43个,占单碱基重复序列总数目的86%。两碱基中,AT重复拷贝类别最多,为418个,占两碱基总重复序列数目的42.44%,其次是AC和AG,各为339(34.42%)和228(23.15%)个。没有发现核心序列为GC的重复拷贝类别。三碱基重复中,共发现7种重复拷贝类别,它们分别是AAT、AAG、ATC、AGG、AAC、ACT和ACC,其中,以AAT最多,其次是AAG和ATC(表3)。

表3 1—3碱基重复类型中重复拷贝类别及其在所属重复类型中的百分比

Tab.3 Type of repeats and the percentage in own types in 1—3 bp repeats

重复类型	单碱基		两碱基			三碱基						
	A	C	AT	AG	AC	AAT	AAG	ATC	AGG	AAC	ACT	ACC
重复数目	43	7	418	228	339	75	24	22	14	8	5	1
总计	50		985						149			
百分比(%)	86	14	42.44	23.15	34.42	50.34	16.11	14.77	9.40	5.37	3.36	0.07

在四碱基重复类型中,AGAT重复拷贝类别最多,共29个(重复拷贝数为4的10个,5的6个,6的5个,7的3个,8的1个,10的2个,13的1个,15的1个);其次为ACAT,12个(重复拷贝数10以下的为11个,拷贝数13的一个)。其余的依次分别为:AGAC,11个(重复拷贝数为4的3个,5的1个,7的1个,10的2个,20的2个,21的1个,23的1个);AAAT,9个(拷贝数都在10以下);AGGG,7个(拷贝数都在10以下);AAAG,7个(拷贝数都在10以下);AAAC,7个(拷贝数都在10以下);ACTC,4个(拷贝数在10以下的3个,12的1个);ACGC,4个(拷贝数在10以下的3个,12的1个);AAGG,3个(拷贝数分别为4、6、8);AGCG,2个(拷贝数分别为4、5);ACCT,2个(拷贝数分别为4、5);AGCA,1个(拷贝数4);ACTT,1个(拷贝数为4);AGTC,1个(拷贝数为11);ACCC,1个(拷贝数为5);ACGG,1个(拷贝数为12)。

五碱基重复类型中的重复拷贝类别和相应的重复数目都比较少,分别为:AGAGA,2个(拷贝数分别为3和4);GAGGC,1个(拷贝数为4);TCTTC,1个(拷贝数为4);TTTCT,1个(拷贝数为3)。

六碱基重复与五碱基重复相比,种类和数量较多,它们依次为:ATTATC,3个(拷贝数都是3);TTTTTC,2个(拷贝数分别是3和19);TCTCTT,2个(拷贝数分别是4和5);AGAGAA,1个(拷贝数为6);TATACA,1个(拷贝数为4);CACGCA,1个(拷贝数为5);GTGTGC,1个(拷贝数为4);ATACAT,1个(拷贝数为3);TTGAAT,1个(拷贝数为4);GAGAAA,1个(拷贝数为6);CCTCTC,1个(拷贝数为3);GAGGGA,1个(拷贝数为4);GCACAC,1个(拷贝数为3);TTCCTC,1个(拷贝数为

13);AGGGTC,1 个(拷贝数为 9);ATCATT,1 个(拷贝数为 16);CTCCTA,1 个(拷贝数为 3);CAATAA,1 个(拷贝数为 3);GCGTGT,1 个(拷贝数为 5);TCATTA,1 个(拷贝数为 5);TCTCTG,1 个(拷贝数为 3);CACTCT,1 个(拷贝数为 6);ATAGAT,1 个(拷贝数为 4);GGGGTA,1 个(拷贝数为 6);AGAGAA,1 个(拷贝数为 6);CTTCCT,1 个(拷贝数为 10);AGATGA,1 个(拷贝数为 4)。

2.2　各种重复拷贝数的分布

根据微卫星重复类型和重复拷贝数的统计数据,统计各重复类型中重复拷贝数的分布(表 4—表 7)。

表 4　单碱基重复拷贝数的分布

Tab. 4　Distribution of the copy number in mononucleotide repeats

拷贝数范围	14—19	20—29	30—39	40—49
A	1	29	11	2
C	1	3	1	2
小计	2	32	12	4
总计		50		
百分比(%)	4	64	24	8

表 5　两碱基重复拷贝数的分布

Tab. 5　Distribution of the copy number in dinucleotide repeats

拷贝数范围	7—9	10—19	20—29	30—39	40—49	50—59	60—69	70—79	80—89	90—99	≥100
AT	34	147	140	48	23	8	7	3	2	2	4
AG	28	58	42	42	26	15	5	5	1	1	5
AC	37	136	75	49	16	10	8	2	2		4
GC	0	0	0	0	0	0	0	0	0	0	0
小计	99	341	257	139	65	33	20	10	5	3	13
总计					985						
百分比(%)	10.05	34.62	26.09	14.11	6.60	3.35	2.03	1.02	0.51	0.30	1.32

表 6　三碱基重复拷贝数的分布

Tab. 6　Distribution of the copy number in trinucleotide repeats

拷贝数范围	5—9	10—19	20—29	30—39	40—49	50—59	60—69	70—79	80—89
AAT	31	29	10	1	3				1
AAC	4	2	1		1				
AAG	8	11	2	2	1				
ATC	8	8	5	1					
ACT	4			1					
AGG	9	3	2						
ACC	1								
小计	65	53	20	5	5				1
总计					149				
百分比(%)	43.62	35.57	13.42	3.36	3.36				0.67

表7 四、五、六碱基重复拷贝数的分布

Tab.7 Distribution of the copy number in tetra-, penta- and hexanucleotide repeats

拷贝数范围	3—9	10—19	20—29	30—39
四碱基	86	10	6	
五碱基	5			
六碱基	29	5		

从表4—表7中可以看出,单碱基重复拷贝数主要分布在20—29个之间,占单碱基重复类型的64%,两碱基主要分布在10—29之间,占两碱基重复类型的60.71%(34.62% + 26.09%),三碱基主要分布在5—19之间,占三碱基重复类型的79.19%(43.62% + 35.57%),四、五、六碱基重复拷贝数主要分布在10个以下范围内。同时,两碱基重复的频率分布呈现一种正态分布的趋势(图1)。

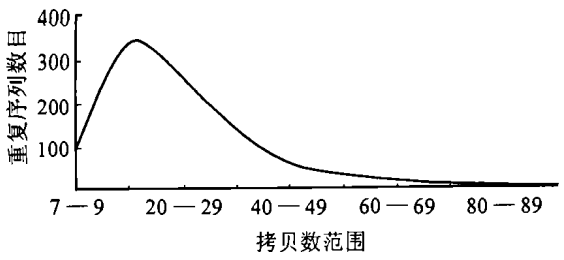

图1 两碱基重复中不同拷贝数所对应的重复序列数目

Fig.1 Number of repeat sequences corresponding to the different copy number in dinucleotide repeats

2.3 两碱基重复序列的评价及分类

根据Weber(1990)针对人类$(CA)_n$中CA重复序列排列方式的不同,两碱基重复的微卫星排列方式可以划分为三种类型:完全(Perfect)、不完全(Imperfect)和复合型(Compound)。完全微卫星是指核心序列以不间断的重复方式首尾相连而成,且其周围(三个碱基以内)不存在其他类型的重复序列;不完全的微卫星是指两个或两个以上的同种重复序列被三个或三个以下的非重复碱基所间隔;复合型微卫星是指一种重复序列和其他种重复序列由三个碱基以下的非重复序列所间隔(包括直接相连接)所组成的重复序列类型。据此划分标准,作者也对中国对虾基因组中的两碱基重复序列进行了划分,见表8。

在此分析的基础上,作者向GeneBank提交注册了其中主要以Perfect形式存在的部分重复序列(登录号为AY545898-AY545913)。其他的一些序列,将在研究其位点多态性信息后再向GeneBank提交注册。

表8 两碱基重复中完全、不完全和复合型微卫星比例

Tab.8 Proportion of perfect, imperfect and compound microsatellites

重复拷贝类别	评价		
	Perfect	Imperfect	Compound
AT	130	263	
AG	80	138	50
AC	158	166	
小计	368	567	50
总计		985	
百分比(%)	37.36	57.56	5.08

3 讨论与结论

3.1 中国对虾微卫星序列分布特征

在中国对虾基因组单碱基重复类型中,A重复拷贝类别最多,这与在其他物种中研究的结果相一致(Katti *et al*, 2001; Tóth *et al*, 2000)。两碱基重复类型中,以AT重复拷贝类别最为丰富,这与有胚植物和真菌类生物相同(Tóth *et al*, 2000),而与人类(Weber, 1990)和其他脊椎动物及秀丽隐杆线虫(*C. elegans*)不同(Tóth *et al*, 2000)。三碱基重复类型中,共发现7种重复拷贝类别,其中以AAT最多,其次是AAG和ATC。人类中的21和22号染色体含有更多的AAT和AAC,果蝇中AGC重复最多,拟南芥和秀丽隐杆线虫含有较多的AAG重复,而酵母中含有更多的AAT、AAG、ATG和AGC重复(Katti *et al*, 2001)。四碱基重复类型中,AAAY(Y代表除A以外的任何碱基)重复拷贝类别在灵长类和啮齿类中最丰

富(Tóth *et al*, 2000),在中国对虾中,这种重复拷贝类别的数量虽然总体上也很多(AAAT, 9个; AAAG, 7个; AAAC, 7个; 共23个),但AGAT重复拷贝类别更丰富(29个)。五碱基重复类型的种类和数量都比较少,其生物学意义有待进一步的研究。六碱基重复类型中发现了29种重复拷贝类别,但各种重复拷贝类别的拷贝数都很少。在其他生物中对此研究的也较少,而且重复拷贝类别也不完全相同(Tóth *et al*, 2000)。由此可见,不同生物中各种重复类型中的重复拷贝类别和其重复数目是不同的,这是否与生物进化的程度有关? 它们的不同对不同生物的基因表达调控有什么影响? 这些都还需要进一步的研究。

作者还根据Weber(1990)对$(AC)_n$微卫星排列的分类标准,研究了中国对虾两碱基重复类型的分布特征,结果表明,不完全(Imperfect)类型最多,这与徐鹏等(2001)的研究结果相反。徐鹏等(2001)采用人工设计合成的$(CT)_7$、$(AT)_7$重复片断作引物,利用PCR法筛选中国对虾小片断部分基因组文库来获得微卫星序列后认为,Perfect的微卫星占大多数,而作者的统计结果是Imperfect占了57.56%。这可能是由于分析方法不同所产生的差异。

3.2 两碱基重复的分布特征及GC重复过少的分析

在人类和果蝇基因组中,微卫星的研究表明,AC两碱基重复拷贝类别的频率最高,其次是AT和AG,而与此不同的是,酵母中AT重复拷贝类别占着绝对的优势(Katti *et al*, 2001)。可能不同的物种各种占优势的两碱基重复拷贝类别是不同的。表3的统计结果表明,中国对虾中AT重复拷贝类别的频率最高,占两碱基重复总数的42.44%,其次是AC,占两碱基重复数的34.42%。这与已报道的甲壳动物DNA中微卫星重复序列的研究结果有一定差别,Xu等(1999)研究斑节对虾(*Penaeus monodon*)基因组中两碱基重复类型时得到的初步结果是$(CT)_n$(即AG)最多,其次才是$(AT)_n$。GC两碱基的重复拷贝类别在所有已经研究过的生物基因组中的含量都很少(Katti *et al*, 2001; Tóth *et al*, 2000),而至今在中国对虾基因组的研究中除了发现一个$(GCG)_3$的重复(Genebank登录号: AF295791)(徐鹏等, 2001)外,还未发现完全由GC重复拷贝类别组成的重复序列的存在。Schorderet等(1992)研究了6种脊椎动物基因组后,对此的解释是:由于基因组DNA中的CpG的甲基化,使之成为一个突变的热点,因为甲基化的胞苷酸C很容易经过脱氨基作用转变成胸腺嘧啶T,而少量的GC又是维持DNA热力学稳定性所必须的。这样的结果是GC重复减少,同时突变后的序列TG(即AC类型)相应增加,这可以一定程度上解释人类基因组中AC重复最多的现象。新近的研究表明,CpG的甲基化可以抑制玉米基因组中种子储藏蛋白基因Zeins的表达,但这种基因表现出很大的变异性,这种变异就是与胞苷酸残基C脱氨基作用转变成胸腺嘧啶T相关(Lund *et al*, 2003)。中国对虾基因组中GC含量如此稀少,作者认为至少与此有一定的关系,因为与之相对应的突变类型AC重复的量仅次于AT重复。另外,统计结果中GC重复过少也可能与GC重复的测序工作比较困难有关。真正的原因还需要进一步研究。

3.3 微卫星技术在中国对虾中的应用前景

微卫星技术虽然建立的时间不长,但已经在包括人类到细菌的基因组研究中得到了广泛的应用,并取得了许多重要的结果。究其原因在于微卫星DNA在生物基因组中具有丰富的长度多态性信息,在群体间和群体内变异范围大,杂合性高、种类多、分布广以及重组率低、容易筛选等优点,因而在甲壳动物虾类的研究中也得到了广泛的应用。在6个微卫星位点上研究白虾(*Litopenaeus setiferus*)不同地理群体的遗传多样性表明,每个位点平均的期望杂合度为0.68(Ball *et al*, 2003),而类似的在另外6个微卫星位点上对另一种白虾(*L. schmitti*),进行筛选的结果表明,这些位点具有更高的多样性(杂合度0.863),平均每个位点的等位基因数为37.8个(Maggioni *et al*, 2003)。在与中国对虾亲缘关系较近的斑节对虾(*Penaeus monodon*)的种群遗传多样性的分析中,5个微卫星位点(CUPmo18、Di25、Di27、CSCUPmo1、CSCUPmo2)上也同样具有较高的杂合度(平均为0.78)(Supungul *et al*, 2000)。综上所述,甲壳类中虾类的微卫星位点具有丰富的多态性信息,这从一定角度也展现了中国对虾基因组中微卫星位点的应用前景。由图1可知,在总的重复序列中占绝对优势的两碱基重复拷贝数主要分布在10—29重复数目的范围内,即中国对虾基因组中的微卫星长度主要分布在20—60个碱基的长度范围内,且以不完全(Imperfect)的形式占优势

(57.56%),这说明在长期进化中微卫星位点积累的变异相对较大,因而这些微卫星序列应该具有较为丰富的多态信息。作者利用8对微卫星引物对黄渤海群体、朝鲜半岛西海岸群体和朝鲜半岛南海岸群体3个野生群体进行了微卫星遗传标记分析,结果表明:8对引物对60尾中国对虾进行PCR扩增,共获得了60个等位基因,其中有1个微卫星位点检测到5个等位基因;5个微卫星位点检测到6个等位基因;1个微卫星位点检测到11个等位基因;1个微卫星位点检测到14个等位基因(结果待发表)。从作者对中国对虾微卫星的统计结果看,在约641000个总碱基长度中共找到了1362个重复序列,平均每隔470个碱基就有一个重复序列的存在,如果按每个基因平均长度1000个bp左右计算,那么平均每个基因内或附近都可能有2—3个微卫星重复序列的存在,因此,微卫星技术的相关研究无论是在中国对虾遗传连锁图谱的构建,还是在抗病、抗逆相关基因的连锁分析、家系分析和亲权鉴定方面都有广阔的应用前景。

参 考 文 献

石 拓,孔 杰,刘 萍等,1999. 中国对虾遗传多样性的RAPD分析——朝鲜半岛西海岸群体的DNA多态性. 海洋与湖沼,30(6):609—615

石 拓,庄志猛,孔 杰等,2001. 中国对虾遗传多样性的RAPD分析. 自然科学进展,11(4):360—364

刘振辉,孔 杰,石 拓等,2000. 中国对虾两个不同地理种群遗传结构的RAPD分析. 应用与环境生物学报,6(5):440—443

刘 萍,孔 杰,石 拓等,2000. 中国对虾黄、渤海沿岸地理群的RAPD分析. 海洋学报,22(5):88—93

邱高峰,常林瑞,徐巧婷等,2000. 中国对虾16S rRNA基因序列多态性的研究. 动物学研究,21(2):35—40

高天翔,李 健,王清印等,2003. 中国对虾线粒体16S rRNA基因序列分析. 中国水产科学,10(5):359—364

徐 鹏,周令华,相建海,2001. 中国对虾微卫星DNA的筛选. 海洋与湖沼,32(3):255—259

徐 鹏,周令华,田丽萍等,2003. 从中国对虾ESTs中筛选微卫星标记的研究. 水产学报,27(3):213—218

Ball A O, Chapman R W, 2003. Population genetic analysis of white shrimp, *Litopenaeus setiferus*, using microsatellite genetic markers. Mol Ecol, 12(9): 2319— 2330

Beheregaray L B, Ciofi C, Geist D *et al*, 2003. Genes record a prehistoric volcano eruption in the Galapagos. Science, 302(5642): 75

Benson G, 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res, 27: 573—580

Katti M V, Ranjekar P K, Gupta V S, 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Molecular Biology and Evolution, 18: 1161—1167

Lund G, Lauria M, Guldberg P *et al*, 2003. Duplication-dependent CG suppression of the seed storage protein genes of maize. Genetics, 165: 835—848

Maggioni R, Rogers A D, Maclean N, 2003. Population structure of *Litopenaeus schmitti* (Decapoda: Penaeidae) from the Brazilian coast identified using six polymorphic microsatellite loci. Mol Ecol, 12(12): 3213—3217

Miesfeld R, Krystal M, Arnheim N, 1981. A member for a new repeated sequence family which is conserved throughout eucaryotic evolution is found between the human δ and β globin genes. Neucleic Acid Res, 9: 5931—5947

Schorderet D F, Gartler S M, 1992. Analysis of CpG suppression in methylated and nonmethylated species. Proc Natl Acad Sci USA, 89: 957—961

Supungul P, Sootanan P, Klinbunga S *et al*, 2000. Microsatellite polymorphism and the population structure of the black tiger shrimp (*Penaeus monodon*) in Thailand. Mar Biotechnol (NY), 2(4): 339—347

Tóth G, Gáspári Z, Jurka J, 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Reseach, 10(7): 967—981

Weber J L, 1990. Informativeness of human (dC—dA)n. (dG—dT)n polymorphisms. Genomics, 7(4): 524—530

Xiang J H, Wang B, Liu B *et al*, 2002. Over 10000 expressed sequence tags from *Penaeus chinensis*. Plant, Animal & Microbe Genomes X Conference, 16

Xu Z, Dhar A K, Wyrzykowski J *et al*, 1999. Identification of abundant and informative microsatellites from shrimp (*Penaeus monodon*) genome. Anim Genet, 30(2): 150—156

ANALYSIS OF MICROSATELLITE SEQUENCES IN CHINESE SHRIMP *FENNEROPENAEUS CHINENSIS* GENOME

GAO Huan, LIU Ping†, MENG Xian-Hong†, WANG Wei-Ji†, KONG Jie†

(*Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao*, 266071;
Institute of Oceanology, Chinese Academy of Sciences, Qingdao, 266071;
Graduate School, Chinese Academy of Sciences, Beijing, 100039)

†(*Yellow Sea Fisheries Research Institute, Chinese Academy of Fishery Sciences, Qingdao*, 266071)

Abstract By sequencing randomly, 3699 clones of sequences in the genome of *Fenneropenaeus chinensis* were obtained. Then, using software DNASTAR (Version 5.0) to assembly all of the clones, 1520 clones independent of each other, were made in which the length of DNA sequences is about 641,000 bp in total. With the help of the bio-soft Tandem Repeats Finder (Version 2.02), 1362 microsatellite repeat sequences are found in the sequences. Criterions to distinguish the repeat sequences of mono-, di-, tri, tetra-, pentra-, hexanucleotide are that the copy numbers of the motif composed of the mono-, di-, tri, tetra-, pentra-, hexanucleotide are ≥ 14, 7, 5, 4, 3 and 3, respectively. In the 1362 repeat sequences, the numbers of the dinucleotide repeats are 985, and most (72.32%) among all of the repeat sequences; the second are the trinucleotide repeats, 149 (10.94%); the third is the tetranucleotide repeats, 102 (7.49%); the forth is the mononucleotide repeats, 50 (3.67%); the fifth is the hexanucleotide repeats, 34 (2.50%); the sixth is the petranucleotide repeats, 5 (0.37%). Numbers of repeat sequences that composed of the motif of A are 43, accounting for 86%, and most among the mononucleotide repeats. In dinucleotide repeat, the numbers of AT repeats are 418, the most, accounting for 42.44%; and the second and third are AC and AG repeats, 339 (34.42%) and 228 (23.15%) respectively. Seven classes of repeat sequences that respectively composed of the motif AAT, AAG, ATC, AGG, AAC, ACT and ACC, are found in the trinucleotide repeats, in which the numbers of AAT repeats are 75, the most; the second are AAG(24); the others are ATC(22), AGG(14), AAC(8), ACT(5) and ACC(1) in turn. AGAT and ATTATC repeats are the most ones in tetranucleotide and hexanucleotide respectively. Both classes and copy number of repeat units are few in pentranucleotide; and there are together four classes: AGAGA, GAGGC, TCTTC and TTTCT. Some of the above sequences are referred to the GeneBank, and the numbers of accession are AY545898—AY545913. The reason of fewer GC dinucleotide repeats are also discussed in the article. Two possible answers are that: one is methylation of C in CpG islands resulting in the mutation of C-T; and another is that it is difficult to sequence the GC repeat sequences.

Distributions of copy numbers in different types of repeat sequences are as follows: copy numbers of mononucleotide repeats are mainly between 20 and 29, accounting for 64%; copy numbers of dinucleotide are mainly between 10 and 29, accounting for 60.71%; copy numbers of trinucleotide repeats are mainly between 5 and 19, accounting for 79.19%; copy numbers of tetre-, pentra- and hexanucleotide repeats together are mainly between 3 and 10. In general, the lengths of microsatellite repeat sequences are mainly between 20 to 60 bp. Among the sequences, the numbers of imperfect sequences are predominance. Based on the above point, it is believed that the nucleotide mutation of microsatellite locations are accumulated largely in a long term of evolution; and there would be abundant polymorphism in these locations. In fact, we get 60 alleles in 8 microsatellite locations, using 8 pairs of microsatellite primers to amplify the genome of 60 individuals by PCR technology. Therefore, it would be very practical to use microsatellite to study the genome of *F. chinensis*.

Key words Microsatellites, *Fenneropenaeus chinensis*, Genome