日本囊对虾(*Marsupenaeus japonicus*)基因组微卫星特征分析*

栾 生 孔 杰** 王清印 高 焕 王伟继 张庆文
中国水产科学研究院黄海水产研究所，农业部海洋渔业资源可持续利用重点开放实验室，青岛 266071

摘要 通过建立随机基因组文库和序列测序，对日本囊对虾基因组进行了较大规模的微卫星分布特征分析．在1606711 bp的随机基因组序列中，共找到1206个微卫星序列，其序列总长度约占测序序列总长度的5.9%．微卫星序列中，以两核苷酸重复的序列数目最多，约占微卫星总数的69.82%，三核苷酸重复次之，约占12.35%．两核苷酸重复中以AT重复最为丰富，约占两核苷酸重复序列总数的29.44%．微卫星在低拷贝区间(≤42)数量相对较大，约占比例为72.31%．重复单位拷贝数的变异能力分析表明，变异系数最大的前3种重复类型，均为4—6核苷酸，较长重复单位类型有着较强的变异能力．微卫星重复单位长度与其拷贝数的相关分析表明，二者呈负相关($r=-0.428$)，即随着重复单位长度的增加，其拷贝数在减少．两核苷酸重复4种类型和三核苷酸重复10种类型基序AT含量与重复序列数目的相关分析表明，随着重复类型基序AT含量的增加，其相应的序列数目增多．同时，微卫星各重复类型基序GC含量在0—70%间时，两端侧翼序列GC含量与之存在着正相关关系．这可能意味着微卫星两端的侧翼序列的碱基组成对微卫星的产生和进化有一定的影响．以上结果为物种间微卫星分布频率和丰度的比较、微卫星标记开发以及微卫星进化和功能的研究等工作提供基础．

关键词 微卫星 日本囊对虾 重复单位 重复类型 拷贝数

微卫星，又称作简单序列重复(simple repeat sequence，SSR)或者短串联重复(short tandem repeat，STR)，一般是指以1—6个核苷酸为重复单位的重复序列．微卫星作为一种共显性的分子标记，具有高度的多态性和丰富的信息量，广泛应用于群体遗传多样性分析[1,2]、遗传连锁图谱的构建[1,3]、数量性状连锁分析[4]、家系识别[5]和分子标记辅助育种[6]等研究领域．

随着物种大规模基因组测序工作的开展，已经在人类(*Homo sapiens*)、蚊子(*Anopheles gambiae*)、鸡(*Gallus gallus*)、斑马鱼(*Danio rerio*)和中国对虾(*Fenneropenaeus chinensis*)等脊椎和无脊椎动物，拟南芥(*Arabidopsis thaliana*)、水稻(*Oryza sativa*)和小麦(*Tritcium aestivum*)等植物基因组以及一些原核生物等30个以上的物种基因组上进行了微卫星分布特征分析[7—11]．研究结果表明，微卫星广泛分布于真核生物和原核生物基因组，不同物种基因组水平上微卫星的分布特征存在着较大差异．人、拟南芥和秀丽隐杆线虫(*Caenorhabditis elegans*)等基因组微卫星序列中，单核苷酸重复为优势类型[10]；果蝇(*Drosophila melanogaster*)、家蚕(*Bombyx mori*)和中国对虾中，两核苷酸重复为优势类型[9,10,12]；酵母(*Saccharomyces cerevisiae*)和丝状真菌(*Neurospora crassa*)中，三核苷酸重复为

2006-08-15收稿，2006-10-23收修改稿
* 国家“八六三”计划项目资助(批准号2005AA603210、2003AA603032)
** 通信作者，E-mail：kongjie@sina.com

优势类型[10,13]. 另外，物种基因组间微卫星密度、碱基组成间也存在着较大差异[14]. 一些研究认为物种本身碱基组成也是选择的结果，另外的研究认为这是由于不同的物种微卫星突变率不同的结果[14]. 尽管不同物种基因组间微卫星的频率、分布和突变率差异较大[15]，但是基于物种间基因组水平上微卫星分布特征的比较分析，有助于从进化学的角度上了解各重复类型分布特点及其功能. 譬如在大部分真核生物两核苷酸重复类型中，发现AT/TA重复是变异能力最强的重复类型[7]. 29个物种基因组序列两核苷酸重复类型的比较分析表明，陆地生物的优势重复类型为AT重复，海洋生物的优势重复类型为AG和CG类型[7].

尽管微卫星已经在遗传学领域得到了广泛应用，然而关于微卫星的起源以及重复单位拷贝数突变的产生机制，一直是研究的难点. 一些分析研究[16]认为，只有当重复单位数目达到一定的阈值，DNA复制滑动机制才会起作用. Zhu等[17]通过分析人的基因突变数据库，认为微卫星起源于DNA链上的随机突变. 因此，最初几个重复单位的产生，可能是由于碱基的插入或者替换产生的. 果蝇基因组测序序列AT含量与$(AT)_n$重复密度间的正相关关系，为这一理论提供间接的证据[18]. 重复单位拷贝数突变的产生机制主要包括复制滑动和重组两种机理. 复制滑动机理认为SSR位点的重复单位拷贝数变化是由于DNA复制时的链滑动失配错误引起的[19]. 重组机理则认为通过非均等交换或者基因转化，重组可能潜在的改变微卫星的长度[20,21].

日本囊对虾(*Marsupenaeus japonicus*)属于甲壳纲、十足目、对虾总科、对虾科，属暖水性种类，广泛分布于非洲东海岸、红海、印度、马来西亚、菲律宾、日本和朝鲜沿海等海域. 在中国，日本囊对虾主要分布于东海南部和南中国海，是中国东南沿海的主要经济虾类之一. 关于日本囊对虾微卫星的研究工作，仅见Moore等[22]通过限制性酶切、超声波随机打断和GAA富集法构建的基因组文库和cDNA文库，筛选出了12对微卫星引物，以及相应的微卫星群体遗传结构调查研究[23]. 基于基因组水平上的微卫星分布特征分析，未见相关报道. 由于日本囊对虾基因组较大，大约相当于人类基因组的80%[24]，从当前的研究需要和成本来看，进行完整的基因组测序并不具备可行性，因此本研究对进行微卫星引物开发的日本囊对虾随机基因组文库测序序列进行了微卫星分析. 这对于从基因组水平上了解日本囊对虾微卫星的分布特征、物种间微卫星分布频率和丰度的比较、微卫星进化和功能的研究以及进一步的微卫星标记开发等工作都具有重要的参考价值.

1 材料和方法

1.1 测序材料

用于基因组测序的日本囊对虾样本来自于广东汕头野生群体，取样时间为2004年9月. 2005年3月—2005年7月期间，对其部分基因组序列进行了测序. 测序个体数目为3个，测序方法为：提取基因组DNA，采用超声波随机打断，经过低熔点琼脂糖凝胶电泳回收400—1500 bp的片段，与PUC19质粒连接，将重组DNA转化到大肠杆菌DH5α中，利用蓝白斑筛选阳性克隆，进而建立日本囊对虾部分基因组文库. 对以上建立的部分基因组文库进行双向测序，共获得3395个DNA克隆序列，经过Seqman Ⅱ软件进行序列拼装，去除重复序列，合并侧翼序列相同，只是微卫星重复数变异的微卫星序列，最终获得2815个独立的克隆序列. 每个克隆序列的长度从400—1000 bp不等，所有克隆序列的总长度为1606711 bp.

1.2 统计方法

通过软件Tandem Repeats Finder (Version 3.21)[23]对拼装后的克隆序列进行分析，查找微卫星序列. Tandem Repeats Finder的查找参数如下：alignment parameters (match, mismatch, indel)=(2, 7, 7)，minimum alignment score to report repeat=50，maximum period size=1000. 利用本实验室编写的Excel宏程序对Tandem Repeats Finder的初步分析结果进行细化和汇总分析. 微卫星序列的最小长度为12 bp[7,8]. 根据碱基互补配对原则和阅读起始碱基顺序的差异，对每种长度类型的重复单位基序进行同类兼并，譬如两核苷酸重复可以兼并为以下4种类型，AC(AC/CA/GT/TG)、AG(AG/GA/CT/TC)、AT(AT/TA)、GC(GC/CG). 1—6核苷酸重复兼并后的重复类型数分别为2，4，

10, 33, 102, 350.

在本研究中，通过 SPSS 13 软件进行统计分析. 由于微卫星重复单位拷贝数、序列长度、不同长度重复单位类型微卫星序列数目等数据(单样本 Kolmogorov-Smirnov 方法进行正态性检验)偏离正态分布($P<0.001$)，因此两变量间的相关分析主要是通过 Spearman 等级方法，微卫星 6 种不同长度重复单位类型的拷贝数差异分析，采用非参数检验中的多个独立样本检验 Kruskal-Wallis H 方法. 显著性检验的水平定义为：$P\leqslant 0.05$，差异显著；$P\leqslant 0.01$，差异极显著.

为了衡量微卫星重复类型的变异能力大小，这里引入变异系数概念，变异系数的计算公式为

$$CV = \frac{S}{\bar{x}} \times 100\% .$$

其中 S 为标准差，$\bar{x}$ 为平均值. 变异系数可以消除单位和(或)平均数不同对两个或多个资料变异程度比较的影响，能够真实反映重复类型变异能力的大小，便于进行类型间的比较.

表 1 微卫星重复类型的频率和分布[a)]

重复类型	AT 含量	重复序列数目/个	占 SSR 总数的分数/%	累积长度/bp	占 SSR 总长度的分数/%	占测序总长度的分数/%	拷贝数范围/个	平均拷贝数/个	拷贝数变异系数/%	侧翼序列的平均 GC 含量
A/T	1.00	50	4.15	1772	1.87	0.11	25—64	36.44	28.89	0.37
小计		**50**	**4.15**	**1772**	**1.87**	**0.11**	**25—64**			
AT	1.00	355	29.44	28766	30.32	1.80	12.5—239	41.05	63.33	0.35
AC	0.50	269	22.31	18841	19.86	1.18	12.5—123	35.54	59.09	0.33
AG	0.50	216	17.91	18906	19.93	1.18	13—146	44.38	54.04	0.39
GC	0.00	2	0.17	81	0.09	0.01	16—25.5	20.75	32.37	0.52
小计		**842**	**69.82**	**66594**	**70.19**	**4.16**	**12.5—239**			
AAT	1.00	57	4.73	5525	5.82	0.35	8.3—116	32.65	81.61	0.30
ATC	0.67	26	2.16	1878	1.98	0.12	8.3—57.3	24.46	53.09	0.35
AGC	0.33	19	1.58	2296	2.42	0.14	9.3—81.7	40.73	50.73	0.46
AGG	0.33	16	1.33	1326	1.40	0.08	9.3—75	27.95	70.32	0.47
AAG	0.67	13	1.08	629	0.66	0.04	8.3—29.3	16.45	43.02	0.41
ACT	0.67	9	0.75	515	0.54	0.03	10.7—29.3	19.79	34.47	0.35
ACC	0.33	3	0.25	150	0.16	0.01	11.3—27	17.00	51.11	0.58
ACG	0.33	3	0.25	376	0.40	0.02	32.7—51	42.67	21.70	
AAC	0.67	2	0.17	65	0.07	0.00	8.7—13.7	11.20	31.57	0.38
GCC	0.00	1	0.08	24	0.03	0.00	8.3—8.3	8.30	0.00	0.37
小计		**149**	**12.35**	**12784**	**13.47**	**0.80**	**8.3—116**			
AGAT	0.75	30	2.49	1666	1.76	0.10	6.3—51.3	14.24	62.65	0.35
ATAC	0.75	15	1.24	1018	1.07	0.06	6.3—49.8	17.31	67.26	0.37
AAAG	0.75	7	0.58	340	0.36	0.02	6.8—30.8	12.40	67.22	0.36
ACTC	0.50	5	0.41	309	0.33	0.02	7—29	15.72	68.33	0.45
AGGG	0.25	4	0.33	285	0.30	0.02	10.3—35.5	17.72	67.13	0.46
AGAC	0.50	4	0.33	339	0.36	0.02	11.3—36.3	21.35	56.19	0.45
GCAC	0.25	4	0.33	267	0.28	0.02	7.8—36.3	16.97	79.24	0.41
AAAT	1.00	3	0.25	245	0.26	0.02	7.8—39.3	21.30	76.17	0.26
TATA	1.00	3	0.25	561	0.59	0.04	27.8—58.8	46.80	35.56	0.26
GGGT	0.25	2	0.17	184	0.19	0.01	8.8—38.8	23.80	89.13	0.48
GTTT	0.75	2	0.17	55	0.06	0.00	6.3—8	7.15	16.81	0.39
TTAG	0.75	2	0.17	85	0.09	0.01	7.3—14.5	10.90	46.71	0.28
TCTC	0.50	2	0.17	274	0.29	0.02	23.8—45.8	34.80	44.70	0.37
GCAG	0.25	1	0.08	85	0.09	0.01	23.3—23.3	23.30	0.00	0.53
ATTC	0.75	1	0.08	24	0.03	0.00	6.3—6.3	6.30	0.00	0.31
CATC	0.50	1	0.08	30	0.03	0.00	7.8—7.8	7.80	0.00	0.42
TGCA	0.50	1	0.08	35	0.04	0.00	9—9	9.00	0.00	0.58

续表

重复类型	AT 含量	重复序列数目/个	占 SSR 总数的分数/%	累积长度/bp	占 SSR 总长度的分数/%	占测序总长度的分数/%	拷贝数范围/个	平均拷贝数/个	拷贝数变异系数/%	侧翼序列的平均 GC 含量
AGTC	0.50	1	0.08	37	0.04	0.00	9.5—9.5	9.50	0.00	0.54
TTGC	0.50	1	0.08	30	0.03	0.00	7.8—7.8	7.80	0.00	0.56
CTGG	0.25	1	0.08	336	0.35	0.02	84.3—84.3	84.30	0.00	
小计		**90**	**7.46**	**6205**	**6.54**	**0.39**	**6.3—84.3**			
ACCTA	0.60	2	0.17	384	0.40	0.02	15.4—61.4	38.40	84.71	0.46
AAAAT	1.00	2	0.17	65	0.07	0.00	6—7.8	6.90	18.45	0.33
ATATT	1.00	2	0.17	98	0.10	0.01	7.8—11.8	9.80	28.86	0.39
AAAAG	0.80	1	0.08	29	0.03	0.00	5.8—5.8	5.80	0.00	0.45
GGCGA	0.20	1	0.08	52	0.05	0.00	10.6—10.6	10.60	0.00	0.54
GGGGT	0.20	1	0.08	51	0.05	0.00	9.8—9.8	9.80	0.00	
AGGGA	0.40	1	0.08	57	0.06	0.00	12.6—12.6	12.60	0.00	0.59
CCCTC	0.20	1	0.08	111	0.12	0.01	22.4—22.4	22.40	0.00	0.56
TCCAC	0.40	1	0.08	34	0.04	0.00	7—7	7.00	0.00	0.59
CTCTC	0.40	1	0.08	57	0.06	0.00	11.6—11.6	11.60	0.00	0.61
GTATA	0.80	1	0.08	182	0.19	0.01	36.6—36.6	36.60	0.00	0.36
小计		**14**	**1.16**	**1120**	**1.18**	**0.07**	**6—61**			
ATCATT	0.83	16	1.33	1679	1.77	0.10	5.7—33.2	17.79	49.34	0.28
ACGCAC	0.33	6	0.50	558	0.59	0.03	6.8—27.3	15.78	61.94	0.51
ACACAT	0.67	4	0.33	801	0.84	0.05	8.5—100.2	33.62	132.20	0.32
TCTTTC	0.67	3	0.25	671	0.71	0.04	9.2—61	36.97	70.61	0.36
TCCCTC	0.33	3	0.25	299	0.32	0.02	9.8—22	16.37	37.60	0.48
TCTCTG	0.50	2	0.17	130	0.14	0.01	9.7—12.2	10.95	16.14	0.30
AAAAAT	1.00	2	0.17	75	0.08	0.00	5.8—6.8	6.30	11.22	0.39
GGGTGG	0.17	2	0.17	77	0.08	0.00	6.2—6.5	6.35	3.34	0.41
GGGAGG	0.17	2	0.17	67	0.07	0.00	4.8—6.5	5.65	21.28	0.36
ATATAC	0.83	2	0.17	140	0.15	0.01	8.5—14.7	11.60	37.79	0.40
ATATAA	1.00	1	0.08	182	0.19	0.01	30.3—30.3	30.30	0.00	0.21
TACCAT	0.67	1	0.08	67	0.07	0.00	11.7—11.7	11.70	0.00	0.47
CTCTGG	0.33	1	0.08	255	0.27	0.02	42.7—42.7	42.70	0.00	0.47
ACCCTC	0.33	1	0.08	77	0.08	0.00	13—13	13.00	0.00	0.40
CTCACT	0.50	1	0.08	101	0.11	0.01	17.3—17.3	17.30	0.00	0.45
AGGGGA	0.33	1	0.08	69	0.07	0.00	11.7—11.7	11.70	0.00	0.52
CACAAA	0.67	1	0.08	170	0.18	0.01	29.5—29.5	29.50	0.00	0.40
TAGAGA	0.67	1	0.08	97	0.10	0.01	16—16	16.00	0.00	
AAATAG	0.83	1	0.08	33	0.03	0.00	5.8—5.8	5.80	0.00	0.44
CTTATC	0.67	1	0.08	26	0.03	0.00	4.5—4.5	4.50	0.00	0.31
GGAGGT	0.33	1	0.08	49	0.05	0.00	8.3—8.3	8.30	0.00	0.41
CACACA	0.50	1	0.08	107	0.11	0.01	18—18	18.00	0.00	
ACTACC	0.50	1	0.08	56	0.06	0.00	9.5—9.5	9.50	0.00	0.58
CCTATA	0.67	1	0.08	201	0.21	0.01	34—34	34.00	0.00	0.03
TCTTCC	0.50	1	0.08	106	0.11	0.01	17.8—17.8	17.80	0.00	0.45
GGAGAT	0.50	1	0.08	30	0.03	0.00	5—5	5.00	0.00	0.38
CATCAC	0.50	1	0.08	67	0.07	0.00	11.3—11.3	11.30	0.00	0.38
TCTCTC	0.50	1	0.08	180	0.19	0.01	30.5—30.5	30.50	0.00	0.38
TATAGA	0.83	1	0.08	38	0.04	0.00	6.5—6.5	6.50	0.00	0.32
小计		**61**	**5.06**	**6408**	**6.75**	**0.40**	**4.5—100**			
总计		**1207**		**94883**						

a）分析序列的总长度为 1606711bp

2 结果

在长度约为1600000 bp的日本囊对虾基因组序列中，共筛选到了1206个微卫星序列，其序列总长度为约占测序序列总长度的5.9%，即平均每1 kb核苷酸序列中包含59 bp的重复序列.

2.1 微卫星各重复类型分布情况

如表1所示，从微卫星序列数目的角度分析，以两核苷酸重复的序列数目最多(842个)，占69.82%；其次是三核苷酸重复(149个)，占12.35%；以五核苷酸重复(14个)所占比例最低(1.16%)．重复序列长度的分析结果与此类似，6种重复类型序列长度占全部微卫星序列长度百分比，以两核苷酸重复(39.36%)最高、三核苷酸重复(7.56%)次之，五核苷酸重复最低(0.66%)．

如表1所示，从微卫星序列数目的角度分析，以两核苷酸重复的序列数目最多(842个)，占69.82%；其次是三核苷酸重复(149个)，占12.35%；以五核苷酸重复(14个)所占比例最低(1.16%)．重复序列长度的分析结果与此类似，6种重复类型序列长度占全部微卫星序列长度百分比，以两核苷酸重复(39.36%)最高、三核苷酸重复(7.56%)次之，五核苷酸重复最低(0.66%)．

单核苷酸重复2种类型中，只发现了A/T重复，未见有G/C重复．两核苷酸重复4种类型中，以AT重复的序列数量最多(355个)，占两核苷酸重复数目的42.16%；CG重复的序列数目最少(2个)，占0.24%．三核苷酸重复10种型中，以AAT重复的序列数目最多(57个)，其次分别是ATC(26个)，以GCC重复的序列数目最少(1个)．四核苷酸重复33种类型中，发现了20种，以ATAG重复的序列数目最多(30)个；其次是ATAC(15个)、AAAG(7个)、ACTC(7个)；其余类型重复序列的数目范围在1—5个．五核苷酸重复102种兼并类型，只发现了11种，并且重复序列的数目很少，范围在1—2个．六核苷酸重复350种兼并类型，只发现了29种，其中以AATGAT重复的序列数目最多，为16个；其次是ACACGC类型，重复序列的数目为6个；其余重复类型序列数目较少，范围1—4个.

2.2 微卫星重复单位拷贝数分布及变异能力分析

微卫星各兼并类型的拷贝数、平均拷贝数和拷贝数范围统计结果见表1．1206个简单重复序列中，重复单位拷贝数范围以两核苷酸重复拷贝数范围最广，为12.5—239；其次是三核苷酸重复，范围在6.3—84.3；再次分别是六核苷酸重复(4.5—100)，四核苷酸重复(6.3—84.3)，五核苷酸重复(6—61)和单核苷酸重复(25—64)．微卫星6种不同长度重复单位类型的拷贝数秩和检验表明：重复类型拷贝数间存在着显著差异($P<0.001$)．6种长度重复单位类型微卫星平均拷贝数分别为36.44，40.09，28.99，17.58，16.18，17.69．其中以两核苷酸重复的平均拷贝数目为最高，五核苷酸重复为最低．重复单位长度和其平均拷贝数间存在着负相关关系($r=-0.771$)，即随着重复单位长度的增加，平均拷贝数减少，但是二者的相关性检验并不显著($P=0.072$)．进一步分析表明，重复单位长度与其拷贝数间也存在着显著负相关($r=-0.428$，$P<0.001$)．

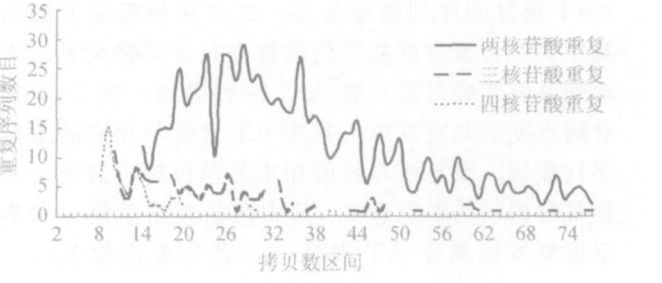

图1 不同重复类型拷贝数区间微卫星序列数目分布

不同拷贝数区间，各重复类型重复序列数量的分布情况见图1．考虑到单核苷酸，五核苷酸、六核苷酸在微卫星序列中所占的百分比很小，这里只对两核甘酸、三核苷酸和四核苷酸重复类型进行分析．从图1中可以看出，两核甘酸拷贝数区间重复序列分布呈现为偏正态分布，微卫星序列在低拷贝区间重复序列的数目相对较多．以42拷贝为基准，小于等于42个拷贝的6种长度重复单位类型的微卫星序列数目分别占其总数目的66.63%，79.19%，93.33%．

变异系数是衡量观测值变异程度的一个统计量，变异系数越大，这种重复类型的变异能力越

大. 微卫星各种类型重复单位拷贝数的变异系数分析结果见表 1. 两核苷酸重复类型中变异能力最大的类型为 AT 重复，其拷贝数变异系数最大(63.33). 三核苷酸重复中变异能力最大的重复类型为 AAT(拷贝数变异系数为 81.61). 1—3 核苷酸重复除 GGC 类型外，其余重复类型拷贝数变异系数均大于 0.4—6 核苷酸重复类型中变异能力最大的重复类型分别为 GGGT(89.13)，ACCTA(84.71)，ACACAT(132.20). 微卫星 6 种长度重复单位类型的平均变异系数平均值分别为 28.89，52.21，43.76，38.86，12.00，15.22，其中两核苷酸重复类型的平均变异系数最高，五核苷酸重复类型的变异系数最小，这说明两核苷酸类型是基因组微卫星最活跃的重复类型.

2.3 微卫星重复单位基序 **AT** 含量与重复序列数目、重复序列长度和重复单位拷贝数间的相关分析

从以上的分析中可以发现，单核苷酸重复序列中，只发现了 A/T 序列，两核苷酸重复中，AT/TA 重复的序列数量最多，三核苷酸重复中，也以 AAT 重复的序列数量最多. 这在某种程度上揭示基序 AT 含量与重复序列数量存在着某种关系. 考虑到单核苷酸重复只有 A/T 一种类型，对 2—6 核苷酸重复类型重复单位基序 AT 含量与相应的重复序列数目、重复序列长度和重复单位拷贝数进行了相关分析，结果见表 2. 其中两核苷酸重复 4 种类型重复单位基序 AT 含量与三者显著正相关($r=0.959$，$P=0.041$；$r=0.977$，$P=0.023$；$r=0.977$，$P=0.023$)，即重复单位基序 AT 含量越高，其重复序列的数目、长度和重复单位的拷贝数分别表现为越多、越长和越高. 三核苷酸 10 种类型 AT 含量与重复序列数目间的相关系数为 0.681，显著相关. 五核苷酸 11 种类型 AT 含量与重复序列数目间的相关系数为 0.663，达到了显著水平. 四核苷酸和六核苷酸兼并类型基序 AT 含量与相应的重复序列数目间相关系数低且不显著.

2.4 微卫星侧翼序列分析

为了分析微卫星序列的可能起源，对微卫星重复单位基序 GC 含量与两端侧翼序列 GC 含量进行比对分析. 对微卫星两端的侧翼序列各取 100 个碱基，总共筛选到了 476 个符合标准的微卫星序列，约占微卫星序列总数量的 39.47%. 对这部分微卫星序列的侧翼序列的 GC 含量进行计算，结果见表 1. 两核苷酸重复 4 种类型 5′和 3′端侧翼序列 GC 含量间无显著差异(数据表 1 中未列出). 3—6 核苷酸重复中的一些重复类型两端侧翼序列 GC 含量间存在较大差异，如三核苷酸重复中的 AGG 类型，四核苷酸重复中的 AGGG 类型，五核苷酸重复中的 AAAAT 类型以及六核苷酸重复中的 ACACAT 类型，其两端侧翼序列 GC 含量间的差值都超过 5%(数据表 1 中未列出). 6 种长度重复单位类型的微卫星重复类型按照 GC 含量分类，两端侧翼序列 GC 含量取上游序列和下游的平均值，分析结果见图 2. 从图 2 中可以看出，二核苷酸重复，随着重复单位基序 GC 含量的降低，相应的侧翼序列 GC 含量也在减少. 3—6 核苷酸重复，重复单位 GC 含量在 0—70%范围内，侧翼序列 GC 含量变化趋势也与两核苷酸相吻合. 两者的相关分析表明，2—6 核苷酸重复单位基序 GC 含量与侧翼序列 GC 含量的相关系数分别为 1，0.4，0.8，0.9 和 0.657，均为正相关.

表 **2** 重复单位基序 **AT** 含量与相应的重复序列数目、长度和重复单位拷贝数间的相关分析[a)]

重复类型		两核苷酸重复	三核苷酸重复	四核苷酸重复	五核苷酸重复	六核苷酸重复
微卫星数量	相关系数	0.959*	0.681*	0.276	0.663*	0.154
	P 值	0.041	0.03	0.240	0.006	0.426
微卫星累积长度	相关系数	0.977*	0.599	0.287	0.166	0.2
	P 值	0.023	0.067	0.219	0.626	0.298
微卫星拷贝数	相关系数	0.977*	0.600	0.288	0.169	0.201
	P 值	0.023	0.067	0.219	0.618	0.297

a) * 表示相关系数的显著性检验为显著

3 讨论

3.1 日本囊对虾基因组微卫星各种重复类型的分布特征

结果分析表明，在日本囊对虾基因组微卫星中，两核苷酸重复为优势重复类型；在重复序列数目(69.82%)、长度(70.19%)和重复单位拷贝数(78.90%)上所占的比例均为最高(表 1). 两核苷酸 4 种重复类型的平均变异系数最大，均大于 0，因此也是基因组微卫星中最活跃的重复类型. 这与果

蝇、家蚕和中国对虾等物种基因组微卫星的优势重复类型相一致[9, 10, 12]，与酵母、丝状真菌(三核苷酸重复)[10, 13]，人、拟南芥和秀丽隐杆线虫(单核苷酸重复)[10, 25]等物种间存在着差异. 进一步分析表明，两核苷酸重复 4 种类型中，AT 重复序列数目所占比例最大，约占两核苷酸重复序列数目的 42.15%，占全部微卫星序列数目的为 29.44%. 拟南芥、酵母、家蚕、中国对虾等基因组两核苷酸重复序列中，也以 AT 重复频率最高[9, 10, 25]. 郭平久等[7]分析表明，陆地生物的优势重复类型为 AT 重复，两种海洋生物海鞘(*Ciona intestinalis*)和斑马鱼的优势重复类型为 AG 和 CG 类型，陆地和海洋生物间存在着显著差别. 同属于海洋生物，本研究结果和中国对虾的分析结果与此并不一致. 究其原因，海鞘和斑马鱼作为海洋生物，属于脊索动物门，而日本囊对虾和中国对虾同属于无脊椎动物节肢动物门甲壳纲对虾属，因此进化上的差别可能是造成它们之间存在差异的主要原因.

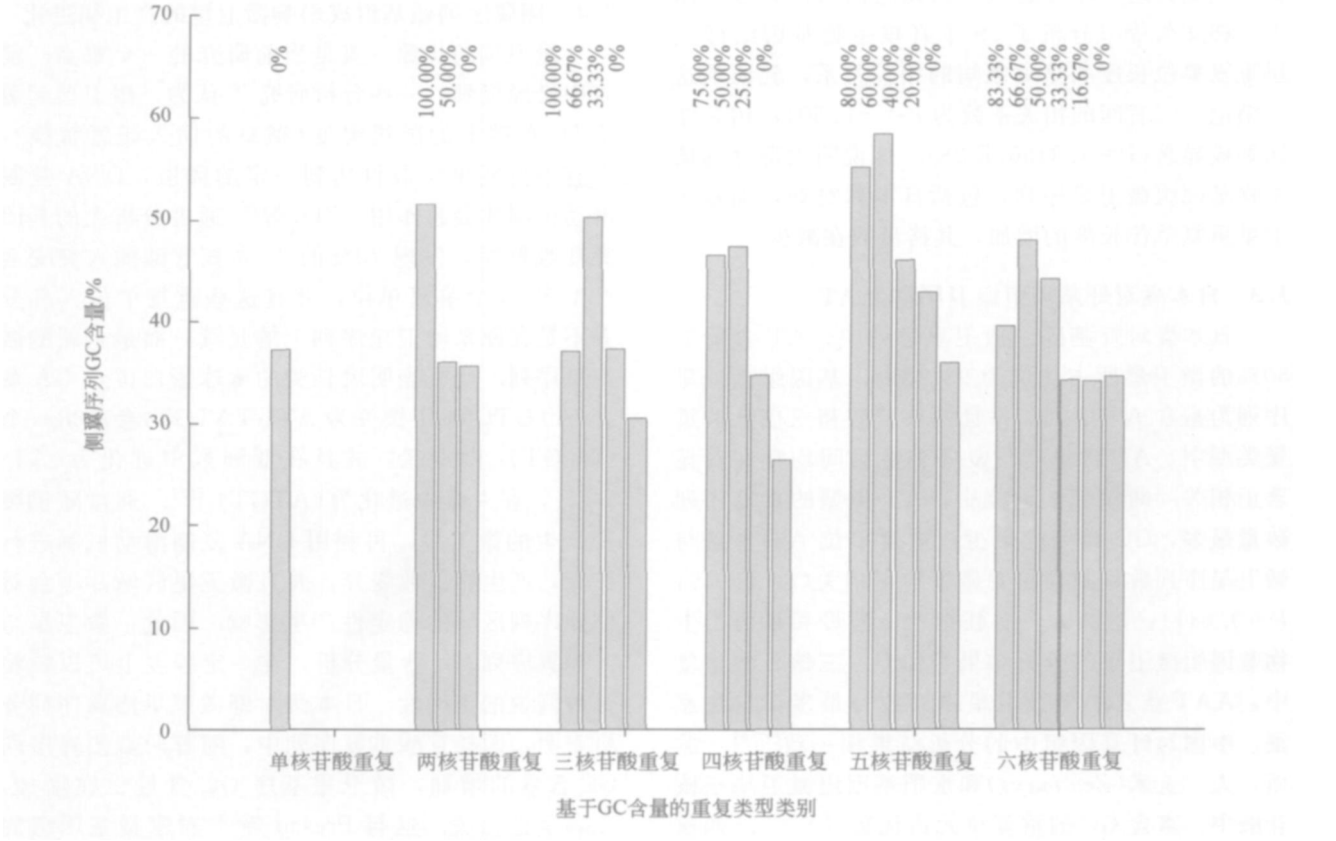

图 **2** 微卫星基序与其侧翼序列 **GC** 含量

本研究中一个值得注意的现象是，1—6 核苷酸重复中，五核苷酸重复序列的数目和拷贝数是最低的. 对家蚕、果蝇、蚊子、小家鼠(*Mus musculus*)、斑马鱼基因组编码区微卫星分布进行分析，发现五核苷酸重复序列占其微卫星总数的百分比都在 2%以下[26]. Gao 等[9]对中国对虾基因组微卫星分布进行分析，也以五核苷酸微卫星序列的数目为最少，并发现重复单位长度为质数 7，11，13，17 的重复序列数目都低于其两边的重复序列数目. 本研究中也发现类似的质数分布现象，重复单位长度为 7，11，13，17，19 的重复序列数目分别为 8，12，15，14，10，均低于两边的重复序列目. 关于这种分布规律的确切原因，目前无从得知，但是可能与重复序列的起源和进化机制，尤其是微卫星和小卫星类型之间的进化关系有关联[9].

3.2 微卫星重复单位的长度与拷贝数间的进化关系

分析结果表明，微卫星重复单位的长度与其相应的平均拷贝数间存在着负相关($r=-0.771$)，每条微卫星序列的重复单位长度与其拷贝数的相关系数为-0.428，也是负相关，这说明日本囊对虾微卫星随着重复单位长度的增加，其拷贝数在减少. 这与 Samadi 等[27]和郭文久等[7]的研究结论相一致. Samadi 等[27]的模拟分析研究认为重复单位长度越长，经受的选择压力越大，因此它们的拷贝数就越少. 郭文久等[7]分析了 29 个真核生物基因组微卫星重复单位长度和拷贝数间的相关关系，验证了这一结论. 二者间的相关系数为$r=-0.304$，相关性达到极显著($P=1.31967E-28$). 这说明大部分真核生物基因组微卫星序列，包括日本囊对虾，随着微卫星重复单位长度的增加，其拷贝数在减少.

3.3 日本囊对虾基因组微卫星富含 **AT**

日本囊对虾基因组微卫星序列中，AT 含量≥50%的微卫星所占比例为 93.45%，基因组微卫星序列为富含 AT 序列. 并且两核苷酸和三核苷酸重复类型中，AT 含量与重复序列数目间均存在着显著正相关. 两核苷酸重复中，AT 类型的重复序列数量最多，GC 类型的最少，重复单位 AT 含量与微卫星序列数目间存在着显著的正相关($r=0.959$, $P=0.041$)，这与人、有胚植物、酵母和真菌类生物基因组微卫星的分析结果类似[8]. 三核苷酸重复中，AAT 重复类型微卫星序列数目最多，与在家蚕、中国对虾基因组中的分析结果相一致[9,12]. 然而，人、玉米(*Zea mays*)和水稻基因组微卫星三核苷酸中，富含 GC 的重复单元占优势[26,28—30]. 四核苷酸、五核苷酸和六核苷酸重复类型序列中，重复单位中含有 G 或 C 碱基类型的重复序列数目较多. 4—6 核苷酸重复序列中，分别以 ATAG，TTAGG，ATCATT 类型的重复序列数目为最多. 这种分布，说明随着重复单位长度的增加，含有 G 或 C 碱基的重复类型会更加稳定[12].

分析微卫星序列富含 AT 的原因：首先，日本囊对虾基因组富含 AT，其 AT 含量约为 60.63%，基因组较高的 AT 含量是微卫星序列富含 AT 的一个基础. 家蚕基因组微卫星分析结果与此类似，基因组中 AT 含量约为 60%，其微卫星序列为富含 AT 序列；其次，微卫星序列 AT 含量高，Tm 值降低，DNA 链容易解开，通过 DNA 复制滑动机制和重组机制，产生富含 AT 重复类型的机率更高. 再次，AT 含量丰富可能与 GC 重复类型的稳定性有关系.[31] 通过对 6 种脊椎动物基因组 DNA 研究，认为 CpG 是一个突变的热点. 由于 CpG 的甲基化，胞苷酸 C 很容易经过脱氨基作用转变为胸腺嘧啶 T. 这可能是本研究中 CG 重复少的一个可能原因.

3.4 侧翼序列碱基组成影响微卫星的产生和进化

微卫星的起源一直是当前研究的一个难点. 前言中已经提到，一些分析研究[16]认为. 微卫星起源于 DNA 链上的随机突变(碱基的插入或者替换)，只有当重复单位数目达到一定的阈值，DNA 复制滑动机制才会起作用. Zhu 等[17]通过分析人的基因突变数据库，发现 70%的 2—4 核苷酸插入突变会产生 2—5 个重复单位，并且这些重复单位大部分并不是在原来微卫星序列上的延续，而是全新的微卫星序列. 研究表明灵长类的 η 球蛋白位点 GA 突变(ATGTG<u>T</u>GT 突变为 ATGT<u>A</u>TGT)会产生一个$(ATGT)_2$微卫星，并且在亚洲猴中进化为$(ATGT)_4$，在人类中进化为$(ATGT)_5$[32]. 通过随机突变产生的微卫星，再利用 DNA 复制滑动机制进行扩充，产生拷贝数变异，并且微卫星区域并不会对侧翼序列区域的稳定性产生影响. 因此，微卫星与其侧翼序列 GC 含量分析，在一定程度上可以检验这种假说的正确性. 日本囊对虾微卫星侧翼序列分析表明，两核苷酸重复序列中，随着两端侧翼序列 GC 含量的增加，微卫星基序 GC 含量也在增加，二者呈正相关. 这与 Prasad 等[12]对家蚕基因组的分析结果相一致，在家蚕基因组 3—6 核苷酸重复序列中，基序 GC 含量在 0—70%时，其侧翼序列 GC 含量也表现出正相关的趋势. 这表明，侧翼序列的碱基组成，对微卫星的产生和进化有一定的影响. 如果侧翼序列 GC 含量高，那么产生富含 GC 重复的微卫星的机率要大于其他重复类型. Bachtrog 等[18]对果蝇基因组含有$(AT)_n$的序列进行星分析，测序序列 AT 含量与$(AT)_n$密度间存在着极显著正相关($P<0.001$)，即 AT 含量越高的序列，序列中$(AT)_n$的密度越大，这表明微卫星很有可能起

源于序列本身. Brohede 等[33] 分析了驼鹿(*Alces alces*)、驯鹿(*Rangifer tarandus*)、梅花鹿(*Cervus dama*)、狍(*Capreolus capreolus*)和猪(*Sus scrofa*)等物种的$(CA)_n$微卫星及其侧翼序列，计算表明二者的碱基突变率分别为0.0112，0.0133，卡方检验显示没有显著的差异，说明微卫星及其侧翼序列面临着相同的选择压力. 这为本文侧翼序列与其微卫星序列碱基含量的正相关提供了可能的合理解释，为微卫星随机突变起源理论提供了可能的证据.

参 考 文 献

1 Hadonou AM, Sargent DJ, Wilson F, et al. Development of microsatellite markers in Fragaria, their use in genetic diversity analysis, and their potential for genetic linkage mapping. Genome, 2004, 47(3): 429—438

2 Romero C, Pedryc A, Munoz V, et al. Genetic diversity of different apricot geographical groups determined by SSR markers. Genome, 2003, 46(2): 244—252

3 Staten R, Schully SD, Noor MA. A microsatellite linkage map of *Drosophila mojavensis*. BMC Genetics, 2004, 5(1): 1—12

4 Sakurai K, Horiuchi Y, Ikeda H, et al. A novel susceptibility locus for moyamoya disease on chromosome 8q23. Journal of Human Genetics, 2004, 49(5): 278—281

5 Selvamani MJ, Degnan SM, Degnan BM. Microsatellite genotyping of individual abalone larvae: Parentage assignment in aquaculture. Marine Biotechnology, 2001, 3(5): 478—485

6 王 曦, 李红霞, 许尚忠, 等. 中国西门塔尔牛产奶性状与微卫星标记相关分析. 畜牧兽医学报, 2004, 35(4): 372—374

7 郭文久. 微卫星在基因组上的分布与功能及其计算方法初步研究. 四川农业大学. 博士学位论文, 2004

8 Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: Survey and analysis. Genome Research, 2000. 10(7): 967—81

9 Gao H, Kong J. The microsatellite and minisatellites in the genome of *Fenneropenaeus chinensis*. DNA Sequence, 2005, 16(6): 426—436

10 Katti MV, Ranjekar PK, Gupta VS. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Molecular Biology and Evolution, 2001, 18(7): 1161—1167

11 Cruz F, Pérez M, Presa P. Distribution and abundance of microsatellites in the genome of bivalves. Gene, 2005, 346: 241—247

12 Prasad MD, Muthulakshmi M, Madhu M, et al. Survey and Analysis of Microsatellites in the Silkworm, *Bombyx mori*: Frequency, Distribution, Mutations, Marker Potential and their Conservation in Heterologous Species. Genetics, 2005, 169(1): 197—214

13 Li CY, Li JB, Zhou XG, et al. Frequency and distribution of microsatellites in the genome of Filamentous Fungus, *Neurospora crassa*. Agricultural Sciences in China, 2005, 4(2): 118—124

14 Oliveira EJ, Pâdua JG, Zucchi MI, et al. Origin, evolution and genome distribution of microsatellites. Genetics and Molecular Biology, 2006, 29(2): 294—307

15 Ross CL, Dyer KA, Erez T, et al. Rapid divergence of microsatellite abundance among species of Drosophila. Molecular Biology and Evolution, 2003, 20(7): 1143—1157

16 Rose O, Falush D. A threshold size for microsatellite expansion. Molecular Biology and Evolution, 1998, 15(5): 613—615

17 Zhu Y, Strassmann JE, Queller DC. Insertions, substitutions, and the origin of microsatellites. Genetical Research, 2000, 76(3): 227—236

18 Bachtrog D, Weiss S, Zangerl B, et al. Distribution of dinucleotide microsatellites in the Drosophila melanogaster genome. Molecular Biology and Evolution, 1999, 16(5): 602—610

19 Eisen J. Mechanistic basis for microsatellite instability. In: Goldstein DB, Schlotterer C eds. Microsatellite: Evolution and Applications. Oxford University Press, 1999

20 Gendrel CG, Boulet A, Dutreix M. $(CA/GT)_n$ microsatellites affect homologous recombination during yeast meiosis. Genes and Development, 2000, 14(10): 1261—1268

21 Richard GF, Paques F. Mini-and microsatellite expansions: The recombination connection. EMBO Reports, 2000, 1(2): 122—126

22 Moore SS, Whan V, Davis GP, et al. The development and application of genetic markers for the kuruma prawn *Penaeus japonicus*. Aquaculture, 1999, 173(1—4): 19—32

23 Sugaya T, Ikeda M, Taniguchi N. Relatedness structure estimated by microsatellite DNA markers and mitochondrial DNA polymerase chain reaction-restriction fragment length polymorphism analyses in the wild population of kuruma prawn *Paenaeus japonicus*. Fisheries Science, 2002, 68(4): 793—802

24 Gregory TR. Genome size evolution in animals. In: The Evolution of the Genome. San Diego: Elsevier, 2005. 3—87

25 Li C, Han B. Diversity of simple sequence repeats in *arabidopsis thaliana* and rice. Acta Botanica Sinica, 2004, 46(5): 603—609

26 Li B, Xia QY, Lu C, et al. Analysis on frequency and density of microsatellites in coding sequence of several eukaryotic genomes. Genomics, Proteomics & Bioinformatics, 2004, 2(1): 24—31

27 Samadi S, Artiguebielle E, Estoup A, et al. Density and variability of dinucleotide microsatellites in the parthenogenetic polyploid snail *Melanoides tuberculata*. Molecular Ecology, 1998, 7(9): 1233—1236

28 Jurka J, Pethiyagoda C. Simple repetitive DNA sequences from primates: Compilation and analysis. Journal of Molecular Evolution, 1995, 40(2): 120—126

29 Chin EC, Senior ML, Shu H, et al. Maize simple repetitive DNA sequences: Abundance and allelic variation. Genome, 1996, 39(5): 866—873
30 Temnykh S, Declerck G, Lukashova A, et al. Computational and experimental analysis of microsatellites in rice(*Oryza sativa L*): Frequency, length variation, transposon associations, and genetic marker potential. Genome Research, 2001, 11(9): 1441—1452
31 Schorderet DF, Gartler SM. Analysis of CpG suppression in methylated and nonmethylated species. Proceedings of the National Academy of Sciences of the United States of America, 1992, 89(3): 957—961
32 Messier W, Li SH, Stewart CB. The birth of microsatellites. Nature, 1996, 381(6582): 483—483
33 Brohede J, Ellegren H. Microsatellite evolution: polarity of substitutions repeats and neutrality of flanking sequences. Proceedings of the Royal Society of London. Series B, Containing papers of a Biological Character. Royal Society(Great Britain), 1999, 266(1421): 825—833

国家自然科学基金重大国际(地区)合作项目“建筑节能设计的基础科学问题研究”取得重要进展

2007年2月9日，国家自然科学基金委员会在西安组织专家，对重大国际(地区)合作项目“建筑节能设计的基础科学问题研究”进行了结题验收，这是内地与香港地区科学家在本领域首次开展的合作研究项目．专家组认真听取了课题负责人对研究项目完成情况的汇报，并审阅了项目组提供的验收材料，一致认为，项目组按期全面完成了研究计划，研究工作取得了突出成果．

随着经济的快速发展和城镇化步伐的加快，我国建筑物的采暖、通风与空调能耗飞速增长．为此，国家将建筑节能作为基本国策，并颁布相关政策法规保障建筑节能工作的贯彻实施．在建筑方案设计阶段利用计算机技术进行室内热环境和能耗的全年逐时模拟分析，控制和降低建筑能耗的有效而快捷的方法已越来越被业界认可．拥有能够真实反映我国各地气候特征并适合动态模拟分析的逐时气象参数，是实现建筑节能设计的基础．在国家自然科学基金重大国际(地区)合作项目的支持下，3年多来，以西安建筑科技大学刘加平教授为负责人，由香港城市大学、中国建筑科学研究院共同组成的项目组围绕“建筑节能设计气象参数的确定”进行了深入研究和分析，取得了突出的、创新性的成果：

(1) 项目组针对我国建筑节能设计的需要，对我国1971—2000年原始气象数据，通过运用统计分析、数值分析等方法，开展了地域太阳辐射模型、建筑气候与太阳辐射区划、太阳能采暖设计与自然通风降温设计分区、人体热舒适与气候适应性等关键科学问题的研究，完成了我国194个地面台站逐时典型气象年(TMY)基础数据库的建设，建立了适用于我国建筑热工与节能设计、采暖通风与空调设计、建筑气候设计及太阳能建筑设计的建筑气象扩展数据库，为建筑节能设计及相关的国家规范和标准的编制及推行提供了数据支持．

(2) 依据新的统计数据，重新划分了全国太阳辐射能分区，建立了适应我国原始气象数据的太阳辐射相关模型．

(3) 完成了全国太阳能采暖与自然通风降温设计的区划指标和设计分区，初步建立了我国不同地区人体热中性温度与地域气候的适应模型．

(4) 初步建立了我国第一个建筑气候与节能基础数据研究中心．

(供稿：茹继平　王　逸)