中国科学 C 辑:生命科学 2009 年 第 39 卷 第 10 期: 933 ~ 939
www.scichina.com life.scichina.com
《中国科学》杂志社
SCIENCE IN CHINA PRESS

论 文

大熊猫 9 个 BAC 的测序、注释和进化分析

郑杨[①②③], 蔡晶[④], 李建文[③⑥], 李波[③], 林润茂[③], 田凤[③], 王晓玲[③], 王俊[③⑤*]

① 中国科学院北京基因组研究所, 北京 100029;
② 中国科学院研究生院, 北京 100049;
③ 深圳华大基因研究院, 深圳 518083;
④ 中国科学院昆明动物研究所, 中德马普进化基因组学小组, 遗传资源与进化国家重点实验室, 昆明 650223;
⑤ Department of Biology, University of Copenhagen, Copenhagen DK-2200, Denmark;
⑥ 华南理工大学生物科学与工程学院, 广州 510006
* 联系人, E-mail: wangj@genomics.org.cn

收稿日期: 2009-04-14; 接受日期: 2009-08-12

摘要 本文构建了相当于大熊猫 10 倍基因组覆盖度的 BAC 文库, 并选择了 9 个 BAC 进行基于第一代 Sanger 测序技术的测序和组装, 以获得完整序列. 这 9 个 BAC 的组装将为评估基于新一代 Illumina GA 测序技术的大熊猫全基因组测序及组装的准确性提供有效资源. 运用同源比对和从头预测的方法, 对 9 个 BAC, 共约 878 kb 的序列进行了基因和重复序列的注释以及进化分析. 一共预测到 12 个蛋白编码基因, 其中, 7 个基因匹配到同源基因的功能注释. 这 7 个基因平均大小约 41 kb, 编码区平均大小约 1.2 kb, 每个基因平均约含 6 个外显子. 同时预测到 7 个 tRNA 基因. 大约 27%的序列被注释为重复序列. 同时, 基于邻接法, 构建了包含人、小鼠、狗、猫以及大熊猫 5 个物种的物种进化树, 结果显示狗的基因与其他 4 个物种相比距大熊猫最近. 本实验结果提供了大熊猫 9 个 BAC 的详细序列及注释信息, 为对大熊猫的研究提供了数据资源.

关键词
大熊猫
BAC 文库
Sanger 测序
组装
注释
进化

大熊猫(*Ailuropoda melanoleuca*)是中国的国宝, 也是世界级濒危保护动物, 最新的统计数据表明, 野外现存大熊猫 2500~3000 只[1]. 大熊猫的饮食特异性、居住地的相互隔离(种群间交流少)、繁育能力低下等原因, 使得大熊猫面临严重的生存危机, 被认为到了物种进化的尽头. 近年来, 大量的生理学、生物化学、遗传多态性、生态学等方面的研究相继展开, 旨在为大熊猫的物种保护提供科学依据.

2006 年, 中国农业大学成功完成了大熊猫第一个 BAC 全基因组文库的构建, 达全基因组 7 倍覆盖度, 并用 FISH(荧光原位杂交)方法将 10 个基因定位到大熊猫的 8 对染色体上[2]. 2007 年, 浙江大学构建了大熊猫二类 MHC(组织相容性复合物)区域连续 650 kb 的克隆重叠群, 为大熊猫 MHC 区域基因的研究提供了很好的工具[3]. 同年, 四川大学完成了大熊猫基因组线粒体的测序、注释和分子进化分析, 结果显示, 大熊猫与熊科物种的亲缘关系最近[4]. 2008 年, 德国科学家利用 10 个熊科物种, 包括包含大熊猫在内的 8 个现存物种和 2 个刚灭绝的物种, 构建基于线粒体基因组的物种进化树, 很好的诠释了熊科物种的进化关系[5].

中国科学院动物研究所的魏辅文教授领导的研究小组对 5 个山区(秦岭、岷山、邛崃、凉山以及相岭山系), 169 个大熊猫进行了线粒体 655 bp 的控制区

引用格式: 郑杨, 蔡晶, 李建文, 等. 大熊猫 9 个 BAC 的测序、注释和进化分析. 中国科学 C 辑: 生命科学, 2009, 39: 933—939

及 10 个微卫星位点的遗传多态性分析. 结果显示, 与其他熊科物种相比, 大熊猫有着不低于其平均水平的遗传多态性, 这意味着大熊猫在遗传结构上并没有走向进化的尽头. 因此, 魏辅文教授呼吁, 物种保护的策略应更多关注恢复和保护大熊猫野外生存空间, 以及保护大熊猫现存的遗传多态性, 促进独立山区之间的遗传物质的交换[6].

2009 年 1 月, 浙江大学解码了大熊猫连续 636 kb 的 MHC II 类区域基因组序列, 并开展了注释和进化分析, 为哺乳动物 MHC II 类基因的进化研究提供了线索[7]. 作为对大熊猫核基因组的进一步探索研究, 本研究组构建了大熊猫基因组 10 倍覆盖度的 BAC 文库, 选择了 9 个 BAC(总长 878 kb), 完成了基于第一代 Sanger 测序技术的测序和组装, 并开展了注释和进化分析, 提供了详细的蛋白质编码基因, 非编码 RNA 基因以及重复序列等的注释信息.

1 材料与方法

1.1 BAC 文库构建

用于文库构建的 DNA 样品来自成都大熊猫繁育基地的 3 岁的雌性大熊猫. BAC 文库的构建是依据 Osoegawa 等人[8]优化了的文库构建步骤完成的. 通过选择性裂解红细胞, 将提取的白细胞包埋于琼脂糖中, 然后用蛋白酶 K 提取高分子量核基因组 DNA, 提纯后包埋于琼脂糖中的核基因组 DNA 用限制性内切酶 B*am*H I (New England Biolab)进行部分酶切. 酶切后的大片段基因组 DNA 利用琼脂糖脉冲电泳进行分离, 选择 100~200 kb 大小的 DNA 进行透析纯化. 回收后的 DNA 与载体 pCC1BAC(Epicentre)环化连接, 最后通过电转化的方式将重组载体转化大肠杆菌感受态细胞(EPI300 高效大肠杆菌感受态细胞, Epicentre). 最终完成了包含 30 万个克隆的 BAC 文库的构建, 达大熊猫基因组 10 倍物理覆盖度.

1.2 BAC 克隆的选择

为了能更好地评估基于第二代 IlluminaGA 测序技术进行的大熊猫全基因组测序与组装的准确性, 选择了 9 个 BAC 进行基于第一代传统 Sanger 测序技术的测序与组装, 以获得 9 个 BAC 的完整序列. 主要依据以下筛选步骤: (1) 随机选择 790 个 BAC 克隆进行单末端测序, 产生 790 条序列, 分别代表每个 BAC, 每条序列长约 500 bp; (2) 用 RepeatMasker[9]对获得的序列进行注释, 丢弃被注释为重复序列的序列; (3) 把剩下的序列比对到 NCBI 的 NT 数据库(NCBI build 36), 丢弃不能比对到该数据库的序列; (4) 对经过前三步过滤后剩下的序列, 根据序列相似度进行聚类, 每一个类别中, 只留下最长的序列, 其他序列均被过滤; (5) 把剩下的序列比对到人的基因组上, 唯一比对上的序列被保留. 最后 要求选择的序列必须比对到人的不同染色体上.

1.3 测序和组装

选中的 9 个 BAC 采用鸟枪法策略, 建立高度随机, 插入片段大小为 1~2 kb 的亚克隆文库(pUC118 质粒). 采用高效, 大规模的末端测序, 对文库中的每一个亚克隆都进行双向测序, 得到成对的, 插入片段为 1~2 kb 的序列. 所有测序数据都由 ABI MegaBace 产生.

在利用 PHRED[10,11]和 PHRAP[12]软件进行组装之前, 首先利用自主开发的一系列软件进行数据粗处理, 过滤低质量数据. 序列补洞是利用引物步移和 PCR 相结合的方法完成的.

1.4 重复序列和非编码 RNA 注释

重复序列的预测主要是基于比较常用的预测软件 RepeatMasker, 将 Repbase[13](Repbase 13.07)中所有物种的重复序列合并为一个重复序列库, 作为 ReaptMasker 的参考文库. 串联重复序列是用专门的预测软件 Tadem Repeat Finder[14]使用经验参数进行预测的.

对于非编码 RNA(ncRNA)基因的预测, 转运 RNA(tRNA)基因使用比较成熟的预测软件 tRNAscan-SE[15], 核糖体 RNA(rRNA)应用 blast 软件在人的 rRNA 数据集中(NCBI Entrez Nucleotide database, NCBI build 36)寻找同源的方法进行预测. 其余 ncRNA 基因的预测均应用基于 Rfam[16]数据库的 INFERNAL 软件(默认参数)进行预测.

1.5 蛋白质编码基因的预测和功能注释

蛋白质编码基因主要通过两种方法进行注释: (1) 蛋白质同源搜索的方法. 用 tblastn(Evalue≤1×10^{-5})将人和狗的蛋白序列(Ensembl Release 52)分别比

对到大熊猫的 9 个 BAC 上, 将比对到同一个蛋白的大熊猫 BAC 区域聚到一个类. 再用 GeneWise[17]对聚类的区域进行基因结构的详细注释, 包括外显子内含子边界的界定等; (2) 从头预测的方法选择了两款比较流行的公共预测软件 GENSCAN[18]和 AUGUSTUS[19], 均使用人的基因集训练得到的软件参数. 最终的基因集是通过对两种方法得到的基因集根据其在 BAC 上的位置进行聚类, 每个聚类中编码区最长的基因将作为该聚类的代表被留下.

基因功能的注释是通过利用 blastp 寻找 SwissProt (Release 56.1)和 TrEMBL(Release 39.1)数据库[20]中同源性最高的蛋白的注释来完成的.

1.6 系统发生树的构建

得到最终预测的基因集后, 分别在选择的另外 4 个物种中(人、狗、猫、小鼠)寻找与大熊猫基因同源性最高的直系同源基因(blastp, Evalue≤1×10^{-5}). 这 5 个哺乳动物物种的同源基因的蛋白序列被用于构建系统发生树, 使用 Treebest 软件, 邻接法, 以 p-distance (两条氨基酸序列比较时, 差异氨基酸个数占氨基酸序列总长的比例)代表进化距离.

Treebest 可以在 sourceforge 网站上免费下载 (http://treesoft.svn.sourceforge.net/viewrc/treesoft/).

2 结果

2.1 基本序列特征

本次测序共产出高质量数据 7605014 bp, 错误率低于 1‰, 覆盖 BAC 基因组 8.7 倍. 9 个 BAC 大小总和为 878324 bp, 约占大熊猫基因组的 0.3‰, 每个 BAC 的平均大小约 97592 bp. 组装结果的详细信息见表 1.

2.2 重复序列

大约 27%的 BAC 区域被预测为重复序列, 主要包含 5 种类型的重复序列, 包括长散在重复序列(LINE)、短散在重复序列(SINE)、长末端重复序列(LTR)、DNA 转座子(transposon)以及串联重复序列(Tandem Repeat)(表 2).

2.3 非编码 RNA 基因

非编码 RNA(ncRNA)基因不翻译成蛋白质, 而以转录 RNA 分子的形式来行使功能. 包括具有重要功能, 在体内非常丰富的 tRNA 和 rRNA. 另外还有核仁小 RNA(snoRNA)、小 RNA(microRNA)以及小干扰 RNA(siRNA)等. 在本次预测中, 发现了 7 个 tRNA 编码基因, 包括 6 个丝氨酸(Sec)和 1 个赖氨酸(Lys), 平均长度约 82 bp, 占 9 个 BAC 总长的 0.6‰. 另有预测到 32 个伪 tRNA 基因(一级序列和二级结构与 tRNA 基因家族一致性序列和结构相似程度较低的基因). 其他 ncRNA 基因都没有在本实验数据集中找到.

2.4 蛋白质编码基因和进化分析

9 个 BAC 共找到 12 个较可信的基因, 其中 7 基因匹配到同源基因的功能注释(表 3). 表格中预测所

表 1 9 个 BACs 的基本序列特征

BAC	测序总数据量/bp	BAC 大小/bp	深度	Scaffolds 个数	Contigs 个数	GC 含量
gpbaaa	788053	87808	9.0	1	1	0.34
gpbaab	876110	94868	9.2	1	1	0.45
gpbaac	937117	104552	9.0	1	1	0.40
gpbaad	768955	85777	9.0	3	3	0.44
gpbaae	759415	101483	7.5	1	2	0.40
gpbaaf	763742	93250	8.2	4	6	0.38
gpbaag	1031290	117931	8.7	1	1	0.40
gpbaah	888508	94933	9.4	1	3	0.39
gpbaak	791824	97722	8.1	1	7	0.51
平均	845002	97592	8.7	1.6	2.8	0.41
总计	7605014	878324		14	25	

得的大部分基因与它们的直系同源基因相比并不完整. 这些基因的不完整性主要是由于两个原因导致的: (1) 这些基因落在了 BAC 的边缘, 使得部分基因区域没有被 BAC 覆盖到(如 *SOHLH2*, 落在 BAC gpbaae 的边缘); (2) 基因太大以致无法被 100 kb BAC 完全覆盖(如 *PARD3B* 在人中的直系同源基因大小达 1 Mb).

这 7 个基因, 基因平均大小约 41 kb, 编码区平均长度约 1.2 kb, 每个基因平均约包含 6 个外显子. 其对应的蛋白, 与人和狗中的相应的直系同源蛋白均有着较高的相似度(Identity). 另外, 有 5 个基因没有匹配到相应的同源蛋白(表 4).

表 2　大熊猫 BAC 基因组重复序列组成与人和狗基因组中重复序列组成的比较 a)

	大熊猫		人*	狗*
	总长度/bp	BACs (%)	(% 基因组)	(% 基因组)
LINE	125770	14.32	21	16.49
SINE	49422	5.63	13	9.12
LTR	27870	3.17	8	3.25
DNA	23270	2.65	3	1.88
TR#	13212	1.50	1.27	1.51
其他	343	0.04	0.14	0
未知	41	0.005	0.01	0.01
总计	239887	27.32	46.42	32.26

a) LINE:长散在重复序列, SINE: 短散在重复序列, LTR: 长末端重复序列, DNA: DNA 转座子, TR#: 串联重复序列; * 数据来自狗基因组序列的分析[21]

表 3　有功能注释的基因集的基本信息

基因编号	基因大小/bp	编码区大小/bp	外显子个数	预测方法 a)	人的直系同源基因 a)	蛋白序列相似度 b) (% 人/狗)
1	51305	2421	9	GeneWise	*GNAS*	84.98/83.54
2	22722	1515	10	GeneWise	*CNTN6*	92.15/94.62
3	55222	1407	11	GeneWise	*CALCR*	85.71/88.66
4	70380	1053	6	GeneWise	*PARD3B*	88.93/92.49
5	16359	726	2	GeneWise	*DCAMKL1*	100/100
6	66161	642	3	GeneWise	*MDFIC*	85.53/93.60
7	6299	642	4	GeneWise	*SOHLH2*	74.19/93.09
平均	41207	1201	6			

a) 一个基因可能被多种预测方法所注释, 表格中所列的预测方法是指最终选择的决定该基因结构的预测方法. 这个表格中所列出的基因都是由 GeneWise 最终决定其结构信息的. 同时, 大部分的 GeneWise 的注释结果, 有 GENSCNA 或 AUGUSTUS 注释结果的支持.

b)蛋白序列相似度是基于 blastp 的局部比对结果

表 4　没有功能注释的基因集的基本信息

基因编号	基因大小/bp	编码区大小/bp	外显子个数	预测方法 a)
8 b)	29209	2295	9	GENSCAN
9	10178	723	5	GENSCAN
10	10665	690	4	GENSCAN
11	5476	594	4	GENSCAN
12	8370	465	3	GENSCAN
平均	12780	953	5	

a) 一个基因可能被多种预测方法所注释, 表格中所列的预测方法是指最终选择的决定相应基因结构的预测方法; b) 这个基因被注释为 LINE-1 逆转录酶同源基因, 它在人和狗中最近的同源基因是没有被注释的可能基因. 这些基因是通过 GENSCAN 或 AUGUSTUS 的从头预测软件注释的, 没有基于同源搜索的 GeneWise 的注释支持. 它们在现存蛋白数据库中没有找到同源, 没有功能注释结果

根据最后基因集, 利用大熊猫、人、狗、猫、小鼠 5 个物种的直系同源基因, 构建基于蛋白序列的物种进化树(图 1), 结果显示, 5 个物种中, 大熊猫基因与狗的基因最近.

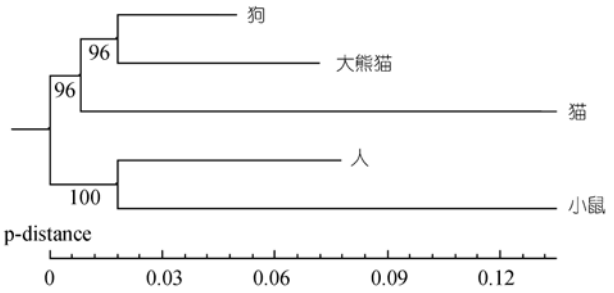

图 1 系统发生树

系统发生树的进化距离是指两条序列之间不同的氨基酸个数占整个氨基酸序列长度的比例(p-distance)

3 讨论

3.1 重复序列

大熊猫 9 个 BAC 的 27%的重复序列组成与人(46%)和狗(32%)的重复序列组成相比偏小, 一部分原因可能是因为本实验 BAC 选择的偏向性, 一部分原因可能是因为大熊猫物种特有的重复序列还不能被 RepeatMasker 所识别.

3.2 预测的蛋白质编码基因

在预测到的基因集中, 一些基因有着非常有意思的功能, 比如 *CALCR*, *SOHLH2* 和 *CNTN6*.

CALCR 是预测基因集中较大的基因, 编码降血钙素受体, 是一种 G 蛋白偶联受体, 与降血钙素结合, 激活腺苷酸环化酶, 包含 7 个跨膜结构域[22]. 研究表明, *CALCR* 的基因突变与骨密度(BMD)相关, 而骨密度低是骨质疏松性骨折的危险信号因子[23]. 大熊猫的身体非常的胖, 而且行动非常缓慢. 深度挖掘该基因在大熊猫个体之间的多态性将有可能为上述问题提供遗传学角度的解释.

SOHLH2 编码一个与精子和卵子发生相关的蛋白, 含有螺旋环螺旋的蛋白质结构域, 这个结构域是转录调控因子蛋白家族的基本特征[24]. 大熊猫面临受孕难、生育难、幼仔存活率低等问题的困扰. 对生殖相关基因的研究, 可能从遗传学的角度为上述现象产生的分子机制提供科学依据.

CNTN6 编码一种接触蛋白(Contactin-6), 在神经系统发育过程中介导细胞表面相互作用, 同时参与动作协调的控制过程[25]. 动作协调是日常生活中最基本的方面, 以大熊猫为例, 吃竹子, 在湖边喝水等看似简单的动作, 都需要肌肉, 肢体还有复杂的神经电路等各方面来协调完成. *CNTN6* 基因很可能参与到主导大熊猫可爱行为的复杂基因调控网络中.

在本实验提供的详细序列和注释信息的基础上, 继续开展基因的表达调控等功能研究, 将有助于人们从遗传学上更好地了解大熊猫这个物种. 物种进化树的结果显示, 5 个物种中, 狗的基因与大熊猫最近, 而小鼠的基因与大熊猫最远. 这个结果与之前的研究结果一致[4].

由于 9 个 BAC 的近似随机选择性, 本实验数据可以作为大熊猫的全基因组序列特征的一个参考. 基于第 2 代 Illumina GA 测序技术的大熊猫全基因组测序最近刚刚完成, 将会呈现出迄今为止最为详尽的大熊猫全基因组遗传学图谱. 本研究完成的基于第 1 代 Sanger 测序技术的大熊猫 9 个 BAC 的测序和组装, 将为大熊猫全基因组组装准确性的评估提供一个良好的资源.

本研究所有的序列和注释数据, 已经上传到 GenBank 数据库中(Accession number: GQ181172-GQ181180).

致谢 感谢深圳市政府和盐田区政府对本项目的支持, 深圳华大基因研究院生物信息中心的樊伟和张国捷提供的指导和修改建议, 深圳华大基因研究院基因组技术平台的田埂在大熊猫 BAC 文库构建时提供的技术平台支持, 深圳华大基因研究院系统支持平台的叶辰为数据存储和传输提供的支持.

参考文献

1 Zhan X, Li M, Zhang Z, et al. Molecular censusing doubles giant panda population estimate in a key nature reserve. Curr Biol, 2006, 16: R451—452[DOI]

2 Liu W, Zhao Y, Liu Z, et al. Construction of a 7-fold BAC library and cytogenetic mapping of 10 genes in the giant panda (*Ailuropoda melanoleuca*). BMC Genomics, 2006, 7: 294[DOI]

3 Zeng C J, Pan H J, Gong S B, et al. Giant panda BAC library construction and assembly of a 650-kb contig spanning major histocompatibility II complex class II region. BMC Genomics, 2007, 8: 315[DOI]

4 Peng R, Zeng B, Meng X, et al. The complete mitochondrial genome and phylogenetic analysis of the giant panda (*Ailuropoda melanoleuca*). Gene, 2007, 397: 76—83[DOI]

5 Krause J, Unger T, Nocon A, et al. Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. BMC Evol Biol, 2008, 8: 220[DOI]

6 Zhang B, Li M, Zhang Z, et al. Genetic viability and population history of the giant panda, putting an end to the "evolutionary dead end"? Mol Biol Evol, 2007, 24: 1801—1810[DOI]

7 Wan Q H, Zeng C J, Ni X W, et al. Giant panda genomic data provide insight into the birth-and-death process of mammalian major histocompatibility complex class II genes. PLoS ONE, 2009, 4: e4147[DOI]

8 Osoegawa K, Woon P Y, Zhao B, et al. An improved approach for construction of bacterial artificial chromosome libraries. Genomics, 1998, 52: 1—8[DOI]

9 http://www.repeatmasker.org

10 Ewing B, Hillier L, Wendl M C, et al. Base-calling of automated sequencer traces using phred. Ⅰ. Accuracy assessment. Genome Res, 1998, 8: 175—185

11 Ewing B, Green P. Base-calling of automated sequencer traces using phred. Ⅱ. Error probabilities. Genome Res, 1998, 8: 186—194

12 http://www.phrap.org/

13 Jurka J, Kapitonov V V, Pavlicek A, et al. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res, 2005, 110: 462—467[DOI]

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

15 Lowe T M, Eddy S R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res, 1997, 25: 955—964[DOI]

16 Griffiths-Jones S, Moxon S, Marshall M, et al. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res, 2005, 33: D121—124[DOI]

17 Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res, 2004, 14: 988—995[DOI]

18 Salamov A A, Solovyev V V. Abinitio gene finding in *Drosophila* genomic DNA. Genome Res, 2000, 10: 516—522[DOI]

19 Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics, 2003, 19: ii215—225

20 Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res, 2000, 28: 45—48[DOI]

21 Lindblad-Toh K, Wade C M, Mikkelsen T S, et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature, 2005, 438: 803—819[DOI]

22 Gorn A H, Lin H Y, Yamin M, et al. Cloning, characterization, and expression of a human calcitonin receptor from an ovarian carcinoma cell line. J Clin Invest, 1992, 90: 1726—1735[DOI]

23 Masi L, Becherini L, Colli E, et al. Polymorphisms of the calcitonin receptor gene are associated with bone mineral density in postmenopausal Italian women. Biochem Biophys Res Commun, 1998, 248: 190—195[DOI]

24 Ballow D J, Xin Y, Choi Y, et al. Sohlh2 is a germ cell-specific bHLH transcription factor. Gene Expr Patterns, 2006, 6: 1014—1018[DOI]

25 Kamei Y, Tsutsumi O, Taketani Y, et al. cDNA cloning and chromosomal localization of neural adhesion molecule NB-3 in human. J Neurosci Res, 1998, 51: 275—283[DOI]