代谢型谷氨酸受体 mGluR7 的基因组结构及其家族蛋白胞外区的基因组结构比较

吴清发^{①②} 董伟^{①②} 齐晓华^② 包维东^{①②} 牛宇欣^{①②} 张以琳^{①②} 张海清^{①②}
陈冲^② 刘斌^② 胡松年^② 汪建^② 杨焕明^{①②*}

(①中国科学院遗传研究所人类基因组中心, 北京 100101; ②北京华大基因研究中心, 北京 100300. * 联系人, E-mail: hmyang@genetics.ac.cn)

摘要 L-谷氨酸是哺乳动物中枢神经系统中的神经递质, 其代谢型谷氨酸受体 mGluR7 在正常中枢神经功能和多种神经退行性疾病的病理发生中起着重要作用. 为了研究 mGluR7 基因相关疾病的生物和遗传基础, 通过从 BAC 文库中筛选种子克隆和对酶切指纹图谱数据库查询等方法, 构建了覆盖 mGluR7 全长基因的基因组物理图谱, 并采用鸟枪法策略对图中 BAC 克隆进行测序与拼接组装, 最终得到准确度为万分之一的完整的基因组序列. 序列分析表明: mGluR7 基因是长达 880 kb 的超大基因; 其基因组 GC 含量为 38%, 重复序列含量为 37.5%; 由 11 个外显子和 10 个内含子组成, 其中 5 个内含子长度超过 100 kb, 内含子 1 长达 285 kb; 存在 2 个替换剪接转录本 mGluR7a 和 mGluR7b. 通过比较 mGluR 基因家族 3 组 8 个亚型受体蛋白胞外区的基因组结构, 发现它们对应的基因组结构分为 3 组, 组内各成员的基因组结构较为保守, 而组间各成员的基因组结构保守性较差. 这种基因组水平的组内保守和组间差异是和蛋白质水平上的保守和差异一致的, 提示这些受体的基本功能在进化的过程中有明显的分化. 在含 mGluR7 的组内, 尽管组内各成员的基因组结构趋于保守, 但大部分内含子长度变化幅度很大, 揭示基因组结构变化的程度高于蛋白质水平的变化, 这种变化预期是在基因表达的调控上体现其生物属性的.

关键词 代谢型谷氨酸受体 物理图谱 基因组 外显子 内含子 胞外区

L-谷氨酸是哺乳动物中枢神经系统(central nerve system, CNS)的神经递质, 参与记忆获取、学习、感觉加工、发育和突触发生等正常中枢神经系统的神经活动. L-谷氨酸的代谢失调有可能是因局部缺血、低血糖、缺氧和癫痫等引发的脑损伤的病理过程以及其他神经退行性疾病引起的^[1,2]. L-谷氨酸通过细胞膜受体起作用, 其受体可分为两大类: 配体门控阳离子通道即离子型受体(ionotropic receptor, iGluR)和代谢型谷氨酸受体(metabotropic glutamate receptor, mGluR). mGluR 是有 7 个跨膜区的 G 蛋白偶联受体, 通过 G 蛋白进行信号传导. 但其序列与其他 G 蛋白偶联受体无同源性. mGluR 有 8 个亚型, 根据它们的氨基酸序列、受体结合特性和信号传导机制的不同, 可分为 3 组, 第 1 组为 mGluR1 和 5, 它们通过 Gq 蛋白激活磷脂酶 C(PLC), 继而激活蛋白激酶 C(PKC), 引起胞内 Ca^{2+} 释放^[3]; 第 2 组为 mGluR2 和 3, 第 3 组为 mGluR4, 6, 7 和 8, 第 2 组和第 3 组的 mGluR 受体都通过 Gi/Go 抑制腺苷酸环化酶的活性, 抑制 cAMP 的

形成, 但它们的序列相似性和受体结合特性不同^[3]. 最近研究表明, 激活 mGluR7 后不仅能降低 cAMP 的水平, 而且激活 PLC^[4].

人 mGluR7 cDNA 于 1996 年被克隆^[4]. 在 8 个 mGluR 亚型中, mGluR7 表现得最为保守, 其氨基酸序列与大鼠的 mGluR7 有 99% 的相似性, 仅在 N 端有 8 个氨基酸不同. 到目前为止, 已发现 mGluR7 有 2 种选择性剪接形式: mGluR7a 和 mGluR7b^[5]. mGluR7 在脑组织不同区域都有表达, 特别在脑皮质、海马和小脑中表达水平较高^[4,6]. 对小鼠和大鼠脑组织研究表明, mGluR7 两种变体 mGluR7a 和 mGluR7b 的分布也有差异, mGluR7a 比 mGluR7b 分布广泛, mGluR7b 常与 mGluR7a 共表达, 但在小脑和侧前庭核中仅发现 mGluR7b 的表达^[7].

胞外区是 mGluR 受体结合谷氨酸和形成二聚体的功能域, 是 mGluR 家族 8 个不同亚型之间最为保守的结构域. 比较其家族 8 个亚型基因的胞外区的基因组结构, 可以获得基因组层次上的进化信息. 根据

实验结果或电子(in silicon)分析, 第 1 组的 *mGluR1* 位于 6q24, *mGluR5* 位于 11P, 第 2 组 *mGluR2* 位于 3P, *mGluR3* 位于 7q21.1~21.2, 第 3 组 *mGluR4* 位于 6P21.3, *mGluR6* 位于 5q35, *mGluR7* 位于 3P26.1~25.2, *mGluR8* 位于 7q31.3~32.1 (<http://www.ncbi.nlm.nih.gov/locuslink>). 我们对 *mGluR7* 的基因组序列加以测定并对完整的基因组序列进行详细的分析. 这项研究将有助于阐明 *mGluR7* 在 CNS 中时空表达的调控机制, 加速了解许多 *mGluR7* 参与的神经性疾病的发病机制, 并为开发出有针对性的治疗药物打下基础^[3,8].

1 材料与方 法

(i) 引物合成和种子克隆的筛选. 根据 *mGluR7* cDNA 3' UTR 的序列设计用于筛选种子克隆的引物对 GRM7 (引物 1: GCTTAATAAGTCACTG-ACATCAGCA; 引物 2: CCTTGCAACCTAGAAGTA-CATACG), 用于末端延伸的引物对(正向: TCAC-AGAATTGGGGAGTCATAA; 反向: CAGGGCAGAA-GGGATAAATAGT)及其他用于填补 BAC 内部缺口的引物均由中国科学院微生物研究所合成.

采用 PCR 的方法, 从 BAC 文库中筛选含特定序列的种子克隆, 该文库为 RPCI-11(<http://www.chori.org/bacpac/11framehmale.htm>), 由国际人类基因组测序协作组惠赠. 为防止 PCR 结果的假阳性, 我们一般将筛选到的克隆进行酶切分析、再次 PCR 和 FISH 定位等方法进行鉴定.

(ii) BAC DNA 的提取. 将筛选得到的阳性 BAC 的菌种在含 12.5 $\mu\text{g/mL}$ 氯霉素的固体培养基上划线培养. 次日将 BAC 单菌落接种于含相同浓度氯霉素的 500 mL LB 液体培养基, 37℃ 振荡过夜. 采用修改的碱裂解法提取 BAC DNA^[9]. 将菌液置 4℃ 的离心管中, 4℃, 4000 r/min 离心 10 min, 弃上清液. 菌体用 5 mL 溶液 I (50 mmol/L 葡萄糖, 25 mmol/L Tris-HCl, 10 mmol/L EDTA, pH 8.0) 晃动摇匀, 加入 5 mL 溶液 II (0.2 mol/L NaOH, 1% SDS) 混匀, 室温放置 5 min, 加入 5 mL 溶液 III (3 mol/L KOAc, pH 5.5) 颠倒混匀, 冰浴放置 10 min, 4℃, 12000 r/min 离心 10 min, 转移上清液至新离心管中, 加入等体积异丙醇, 室温 12000 r/min 离心 10 min, 弃上清液, 沉淀用 300 μL 的溶液 IV (10 mmol/L Tris, 50 mmol/L EDTA) 溶解后, 再加入 300 μL 的溶液 V (50 mL 7.5 mol/L KOAc, 23 mL HAc, 127 mL H₂O) 混匀, -70℃ 放置 30 min, 取出

置室温待其融化, 12000 r/min 离心 20 min, 转移上清液至新管中, 加入 2 倍体积的无水乙醇, 混匀后 12000 r/min 离心 20 min, 沉淀用 70% 乙醇洗涤, 晾干, 用含 0.1 $\mu\text{g}/\mu\text{L}$ RNase 的 40 μL TE 溶解, 37℃ 放置 30 min 备用.

(iii) 基因组和 BAC 末端测序. BAC DNA 测序采用鸟枪法策略. 将 10 μg BAC DNA 经超声打碎, 用 T4 DNA 聚合酶补平末端, 电泳并切下含 1.5~3 kb 片段的胶块, 用 QIAGEN 胶回收试剂盒回收相关片段. pUC18 载体用 *Sma* I 酶切并脱磷. 在 10 μL 连接体系中, 回收片段和载体按 2:1 的比例混合, 连接反应在 4℃ 进行 16 h. 用连接产物转化 *E. Coli* DH5 α 细胞. 经蓝白斑筛选后, 用牙签将阳性克隆挑入 96 孔培养板培养, 37℃, 250 r/min 振荡过夜. 用碱裂解法大规模提取质粒, 通过凝胶电泳检测模板质量, 将符合测序条件的模板转移至新的 96 孔板中, 并将各孔模板浓度调至同一浓度. 用 PCR 热循环双脱氧核苷酸方法测序, 按 ABI Prism DNA Sequencing kit 和 ET-terminator Sequencing kit 说明书进行. 反应产物经 ABI377 或 MegaBACE-1000 测序仪分离. BAC 末端测序则直接以 200 ~ 500 ng BAC DNA 为模板, 以 SP₆ 或 T₇ 序列为测序引物, 增加测序循环数到 60 个循环, 其余按 ABI Prism DNA Sequencing kits 说明书进行末端测序反应.

(iv) 拼接组装和基因组分析. 测序数据读取、拼接和监控采用不同的计算机软件. Phred 从胶图上的 ABI 或 ABD 文件读取碱基序列, 并给每个碱基一个质量评估值^[10,11]. 在拼接之前, 除去每个反应中的载体序列. Phrap 则拼接反应并生成质量文件, 装配好的每个碱基都被赋予一个质量评估值(<http://bozeman.mbt.washington.edu/Phrap.docs/Phrap.html>). 在 Consed 界面上对拼接结果进行检查修正^[12]. 同时采用 DNASTAR 软件来帮助拼接. 重复序列鉴定采用 RepeatMasker (<http://repeatmasker.genome.washington.edu/cgi-bin/RepeatMasker>), 基因外显子定位和数目分析采用 BLAST2 (<http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html>).

2 结 果

2.1 种子克隆和 BAC 物理图谱

以 GRM7 筛选 BAC 文库, 共得到 17 个 PCR 阳性 BAC 克隆. 以其中之一 RP11-124L8 作为种子克隆, 在酶切指纹数据库查询(<http://genome.wustl.edu/gsc/>)

human)得到由 129 个 BAC 组成的 BAC 重叠群, RP11-124L8 位于 BAC 重叠群中间. 因 GRM7 位于 *mGluR7* 的 3' UTR, 由此可以定位 *mGluR7* 基因 3' 端的位置, 进而得知 *mGluR7* 基因在染色体上的走向. 我们首先选择 PCR 反应都为阳性且在重叠群上分别向两边延伸的 RP11-329A2 和 RP11-25C11 克隆测序. 将 *mGluR7* 的 cDNA 与 RP11-329A2 和 RP11-25C11 的序列进行比较, 可推测出 *mGluR7* 沿着重叠群的走向. 从这一重叠群末端 RP11-489O22 开始进行第二次行走. 采用通用引物 SP₆ 和 T₇ 对 RP11-489O22 进行末端测序, 对得到的序列进行分析, 滤除重复序列和其他同源序列, 以其特异的单拷贝序列为新的引物对. 用来源于两个末端的两个引物对分别对与 RP11-489O22 重叠的克隆 RP11-24B16 进行 DNA 扩增, 来源于 RP11-489O22 SP₆ 末端的 PCR 结果为阳性, 而它的 T₇ 末端 PCR 结果为阴性, 这表明 RP11-489O22 的 T₇ 末端和 RP11-24B16 没有重叠. 因此以来源于这一克隆 T₇ 末端的引物对进行新一轮的 BAC 文库筛选,

得到阳性 BAC RP11-812I2, 再对酶切指纹数据库进行查询, 得到另一个由 121 个 BAC 组成的重叠群(图 1), 通过这些 BAC 的 *Hind*Ⅲ 指纹图谱分析, 估算出整个重叠群大小接近 1 Mb.

2.2 BAC 序列测定和序列拼接

对 BAC 克隆的测序分两个阶段进行. 首先, 每个 BAC 按 4~5 倍(4~5 ×)覆盖率进行测序来构建工作草图. 按每个 BAC 的插入长度为 150 kb, 每个测序反应的平均读长为 500 bp 计算, 即每个 BAC 约需 1200~1500 个成功反应. 然后选择最小重叠的 BAC, 补加测序反应到 7~8 倍(7~8 ×)覆盖率. 采用 Phred-Phrap-Consed 软件包进行反应拼接并生成质量文件, 对于 BAC 内部的缺口和低质量区, 采用两种方法来解决, 一是挑选合适的模板反向测序, 二是设计引物行走(walking). 最后, 所有装配好的错误率小于万分之一的 BAC 序列都经过电子酶切, 酶切结果和实际酶切的电泳结果相比较以验证是否符合组装结果. 我们将所有 BAC 的序列通过 DNASTar 整合成连续的序列(表 1).

图 1 覆盖整个 *mGluR7* 基因的 BAC 重叠群物理图谱及有关 STS 的定位
虚线所示的 BAC 不包含 *mGluR7* 基因组的序列

表 1 覆盖 *mGluR7* 基因的 BAC 克隆及大小

BAC 名称	登录号	插入片段长度/bp	成功反应数	碱基错误率(1/10000)
RP11-25C11	AC027121	162771	1260	0.01
RP11-124L8	AC026165	166330	4829	0.08
RP11-329A2	AC026157	199831	1456	0.05
RP11-194I5	AC023333	152246	2206	0.22
RP11-157E16	AC068313	183690	1055	0.19
RP11-53E12	AC018504	178133	4696	0.96
RP11-291P21	AC026184	179563	2973	0.07
RP11-590A20	AC026204	171055	1440	0.49
RP11-669E3	AC034194	179249	1896	0.89
RP11-24B16	AC066585	134846	1836	0.56
RP11-489O22	AC069247	180234	1214	0.04
RP11-812I2	AC026217	193373	1231	0.82
RP11-534J21	AC026200	188379	1439	0.21
RP11-137G22	AC023514	172077	2021	0.07
RP11-13F6	AC018840	134841	3456	0.34

2.3 mGluR7 基因组结构分析

在 *mGluR7* 基因组区域, 平均 GC 含量为 38.0%, 略低于人整个基因组的 GC 含量平均值 41.0%. 分别计算各外显子和内含子的 GC 含量, 可发现这段基因组序列中内含子的平均 GC 含量为 37.3%, 相当恒定, 而 *mGluR7* 基因的 5' 端外显子的 GC 含量为 67.5%, 和周围的基因组序列形成强烈的对比; 而 3' 端的外显子的 GC 含量则降为 38.0%, 和周围基因组基本一致(图 2). 前面 8 个外显子, 主要编码受体的胞外区域. 这段基因组的重复序列含量为 37.5%(SINE 8.92%, LINE 15.85%, LTR 8.47%, DNA 2.39%, 其他为 1.96%). 计算各内含子中的各重复序列的含量, 可见 SINE 和 DNA 转座子含量相对恒定, 而 LINE 和 LTR 则有相应的波动(图 3).

图 2 *mGluR7* 外显子和内含子 GC 含量分布

■ 示外显子, ● 示内含子

图 3 *mGluR7* 内含子各类重复序列含量分布

■ 示 SINE, ● 示 LINE, ▲ 示 LTR, ▼ 示 DNA

将 *mGluR7* 的两个 cDNA 序列(A43649, X94552) 和基因组序列比较, 可知 *mGluR7* 共有 11 个外显子. A43649 编码 *mGluR7b*, 有外显子 10, 但没有外显子 11; X94552 编码 *mGluR7a*, 有外显子 11, 没有外显子

10. 这一预测与实验结果完全一致^[4,5].

所有剪接受体和供体位点均符合 GT-AG 规则. *mGluR7* 的启动区无 TATA 盒, 但转录起始区 GC 含量高且含许多 Sp1 结合位点, 这种启动子模式和很多持家基因的启动子模式相同, 如二氢叶酸还原酶(DHFR) 基因和 *Dp71* 基因^[13]. 翻译起始密码子 ATG 在外显子 1 中, 终止密码子分别位于外显子 10 和 11 中; 外显子 10 是最小的外显子(71 bp), 外显子 8 是最大的外显子(936 bp). *mGluR7* 蛋白结构可分为跨膜区、胞外区和膜内区 (<http://www.cbs.dtu.dk/services/TMHMM-1.0/>), 外显子 1 ~ 8 编码胞外区的 589 个氨基酸, 外显子 8 的 3' 端和外显子 9 的 5' 端编码跨膜区的 261 个氨基酸, *mGluR7a* 的胞内区 64 个氨基酸是由外显子 9 的 3' 端和外显子 11 编码, 而 *mGluR7b* 的胞内区 71 个氨基酸是由外显子 9 的 3' 端和外显子 10 编码. 内含子大小幅度变化很大, 内含子 9 是最小的内含子, 长为 6 kb, 而内含子 1, 2, 4, 7 和 8 都大于 100 kb, 内含子 1 最大, 长达 285 kb (表 2).

mGluR7 的两种变体 *mGluR7a* 和 *mGluR7b* 的氨基酸端 899 个氨基酸序列完全一致, 但它们 3' 末端分别由外显子 11 和外显子 10 编码, 导致 *mGluR7a* 末端的 16 个氨基酸和 *mGluR7b* 末端 23 个氨基酸的序列差异. *mGluR7a* 的 3' 端氨基酸不仅与 *mGluR7a* 定位于轴突有关, 而且通过与含 PDZ 结构域的 PICK1 结合, 使 *mGluR7a* 聚集成簇^[14,15]. *mGluR7b* 的 3' 端氨基酸也有指导蛋白定位于轴突的信号序列和与 PICK1 结合的能力^[16,17], 但与 *mGluR7a* 相比, *mGluR7b* 与 PICK1 结合能力仅为 50%^[17].

2.4 mGluR 家族的胞外区基因组结构比较

mGluR 位于细胞膜上, 其胞外区是结合谷氨酸受体和形成二聚体的区域, 是 *mGluR* 受体分子中最为保守的区域. 对所有 *mGluR* 来说, 已发现的所有替换剪接形式都是针对胞内参与信号传导的部分. 这些蛋白都可分为跨膜区、胞外区和膜内区三部分 (<http://www.cbs.dtu.dk/services/TMHMM-1.0/>). 各个蛋白胞外区大小稍有不同, *mGluR1* 最大, 为 594 个氨基酸, *mGluR2* 最小, 为 566 个氨基酸. 第 1 组的 *mGluR1* 和 5 的胞外区由 7 个外显子参与编码; 第 2 组的 *mGluR2* 和 3 的胞外区由 3 个外显子参与编码; 第 3 组的 *mGluR4*, 6, 7 和 8 的胞外区则有 8 个外显子参与编码. 编码所有 *mGluR* 受体胞外区的第 1 个外显

子都含有起始密码子 ATG，编码胞外区的最后的外显子都是占用编码跨膜区的外显子的 5'端. 在各组内，外显子大小和结构都很保守，但在不同组之间，外显子大小和结构则完全不同，含 *mGluR7* 的第 3 组和另两组相比，仅有 2 个内含子的插入位置保守(图 4). 在各组内部，虽然外显子大小和结构都很保守，

但其内含子大小变化幅度很大，以第 3 组最为典型(表 3). 进化分析的结果表明 *mGluR6* 出现的时间早于 *mGluR4* 和更晚的 *mGluR7* 和 *mGluR8*(数据未显示). 从表 3 中可以看出，内含子增大趋势极为明显，*mGluR7* 和 *mGluR8* 中各有 4 个内含子长于 100 kb，*mGluR7* 的内含子 1 更是长达 285 kb.

表 2 *mGluR7* 基因外显子-内含子结构

外显子	外显子起始位置	内含子-外显子交界处序列	外显子长度/bp	外显子-内含子交界处序列	内含子长度/bp	内含子相
1	1	GTGAGC	757	TTCCAGgtagggtgc	284713	0
2	285321	gtctttgcagATCCCC	217	AGGCAGgtaggatgag	152011	1
3	437550	tcttaacagGTGGAC	142	TATAAAgtaagaataa	7673	2
4	445364	ctttccacagGCAGAT	152	TGGAAGgtatgggttt	108386	1
5	553903	ctctgtttagGTTTG	141	GCACAGgtaatttaat	37452	1
6	591495	taatgtgcagGACAGG	201	TCAATGgtgagtctcc	8810	1
7	600505	tcttccttagGTAGTG	140	CTCAATgtgagtctcg	116700	0
8	717344	tatgttacagATAGAA	936	GAAAAgtaagtgaag	100695	0
9	818974	tgctectagCTCTAC	247	CAAACAgtagtaact	6062	1
10	825282	tttctgtagACTGTA	71	3'UTR after stop	53930	
11	879281	tttcttcagGCCCTG	130	3'UTR after stop

图 4 *mGluR* 家族 8 个不同亚型受体蛋白的胞外区的比对及其内含子插入位置比较示意图

各个片段代表每个外显子编码的蛋白部分；连线分别表示保守的内含子插入位置、翻译起始位点 ATG 和跨膜区起始位置 TM；箭头所示连线为组间保守的内含子插入位置

表 3 *mGluR4*, 6, 7 和 8 受体基因胞外区的内含子大小

基因名称	内含子名称及长度/bp						
	1	2	3	4	5	6	7
<i>mGluR6</i>	2366	307	672	1184	121	1952	84
<i>mGluR4</i>	40890	13124	2765	2291	~6930	234	3573
<i>mGluR8</i>	135981	201813	422	1293	132477	160291	75475
<i>mGluR7</i>	284713	152011	7673	108386	37452	8810	116700

<https://engine.scichina.com/doi/10.1360/csb2002-47-7-508>

3 讨论

3.1 序列的测序与组装

在用鸟枪法策略对 BAC 测序时, 我们一般按 1 : 1 比例安排测序反应, 即一半模板用通用引物同时进行正反向测序, 另一半只进行正向测序. 这样一方面保证模板分布的随机性, 另一方面可通过同一模板的正反向序列在拼接时提供“连锁信息”, 并用这些序列在 BAC 序列中的相对距离来监控拼接质量. 这样就在随机性和“连锁信息”中找到了一个平衡点. 单向测序能够增加模板的随机性而且有助于填补缺口. 如果模板的另一端恰好覆盖一个缺口, 可通过测定模板另一端序列来填补缺口. 一般情况下, 4 倍覆盖率(4 ×)能够包含 95% 的 BAC 区域, 则有助于和它重叠的 BAC 的拼接, 而且能够提供 SNP 信息.

对 BAC DNA 拼接时, 通常先去除反应碱基质量值低于 Q20 的那些反应. 这些低质量的反应不仅无助于 BAC 的拼接, 而且常常引起错误拼接. 我们用 Phrap 软件拼接时, Phrap 软件给每个碱基作一个质量评估值, 这为我们提供了一个客观的评价标准: 完成 BAC 拼接的错误率小于 1/10 kb. 通过 e-PCR (<http://www.nvbi.nim.nih.gov/genome/sts/epcr.cgi>)检测, 可在完整的基因组序列发现 GRM7 和其他 25 个 STS, 这些 STS 先前定位在 3 号染色体上或位置未定; 将 mGluR7 的 cDNA 和完整的基因组序列比较, 各外显子转录本和基因组的排序保持一致, 进一步确证我们得到的基因组序列的正确性. 整个基因组片段定位于 Genethon 遗传图谱上的 16.5 ~ 19.0 cM 的 2.5 cM 区间内. 这个区域 500 kb 的遗传距离大约相当 1 cM, 其遗传重组率是人染色体平均遗传重组率的 2 倍.

3.2 基因组编码区的鉴定

对基因组序列中编码序列的鉴定是基因组学的一个重要且艰巨的任务. 目前对基因组中基因识别鉴定有两种方法, 一是基于公用数据库中已知的信息来识别和鉴定基因的组织结构, 这不仅包括将基因组序列和数据库中已知基因 cDNA 和 EST 在核酸序列水平上作相似性比较, 而且包括将基因组序列和来源于其他模式生物的序列都以不同的可读框翻译后在蛋白质序列水平上比较. 因为在进化过程中, 不同物种中来源于同一祖先的基因在蛋白质水平比核酸水平更为保守, 从蛋白水平比较, 能发现很多在核酸水平上无法揭示的信息. 另一种方法为完全从

新起始基因预测, 常用 Genscan 等软件对基因组中可能的蛋白编码区进行识别.

Genscan 用于基因预测的概率模型包括基因的基因组特征, 这些基因组特征与物种紧密相关. 这些特征包括: 典型基因密度, 每个基因的平均外显子数, 不同类型的外显子大小分布, 编码区及非编码区(内含子和基因间序列)的 6 核苷酸组成的差异, 转录起始和终止信号, TATA 盒, Cap 位点, PolyA 位点, 供体和受体等信息. GC 含量不同区域的基因密度和基因结构相差很大, 在预测时需采用不同的模型参数. 在不同的概率模型中, 模型参数相差较大, 如类别 1(G + C < 43%)中的内含子长度为 2069 bp, 而在类别 4 (G + C > 57%)中, 内含子长度为 518 bp^[18]. 由于重复序列在基因组中所占的百分比通常为 40% ~ 50%, 而且它们具有或多或少的基因信号特征, 如 LINE 的转录起始信号和 PolyA 位点等, 因此它们存在与否, 不仅与局部的 GC 含量密切相关, 而且影响局部的基因组特征. 我们用 Genscan 对 mGluR7 屏蔽重复序列前后的基因组序列进行预测, 两者得到的编码区数目/基因数目分别为 105/29 和 46/8; 将预测的编码区序列和 mGluR7 的 cDNA 序列比较, 未屏蔽的基因组序列给出 mGluR7 基因的 8 个外显子, 缺少外显子 5, 10 和 11; 而屏蔽后的基因组序列给出 mGluR7 基因的 9 个外显子, 缺少外显子 10 和 11. 各个比对上的编码区都分别位于每个独立的预测基因中. 将预测得到的编码区序列和 EST 数据库比较, 未找到在核酸和蛋白水平上相似的其他序列.

将基因组序列和公用数据库做相似性搜索, 发现 mGluR7 基因的内含子 1 序列和 EST(BG214418)序列完全匹配. 该 EST 由 2 个外显子组成, 转录方向和 mGluR7 相反, 该 EST 和数据库中其他序列无任何同源性. 由于此 EST 是采用对基因组随机激活的方法得到, 该基因在何组织中表达及与 mGluR7 基因表达的关系有待进一步研究^[19].

根据公用数据库(如 GenBank)中的序列, mGluR7 基因位于长度为 3 Mb, 平均 GC 含量为 38% 的基因组区域中. 在 mGluR7 基因组区域的端粒侧, 有长为 1.4 Mb 的基因空白区, 在 mGluR7 另一侧, 有 600 kb 的基因空白区. 此区段是否真正“空白”或者存在不同形式的基因分布, 如何从这些大段的基因空白区找出基因, 是一个亟待解决的问题^[20]. 根据 Genscan 对 mGluR7 分析的结果, Genscan 预测的方法无法准

确地预测大基因组跨度的 mRNA 完整序列, 但能预测出大部分的编码区^[21]. 因此对那些大的基因空白区, 可先用 Genscan 对屏蔽重复序列后的基因组序列预测编码区, 以预测的各编码区序列作为探针, 和不同组织来源和发育阶段的 mRNA 做反向 Northern 杂交, 来获得可能的全长 cDNA.

在 GenBank 中, 还有另一条编码 *mGluR7* 的 cDNA (U92458), 它的编码区和 X94552 完全一样, 但其 5' 及 3' UTR 和 X94552 差异较大. U92458 的 3' UTR 长度为 1200 bp, 与 *mGluR7a* 的全长 3' UTR 匹配 (<http://compbio.ornl.gov/Grail-1.3>). U92458 的 5' UTR 包含额外的 303 bp 序列, 未发现与 3 号染色体序列的相似性, 但和染色体 19 的序列完全匹配. 推测此现象可能是两个独立的克隆错误拼接产生的^[22].

3.3 基因组进化

对 8 个 *mGluR* 受体亚型的胞外区基因组结构比较时, 我们先用 DNASTAR 软件中 Clustal 方法依据 PAM250 矩阵对它们的氨基酸序列进行比对, 可以看到它们长度虽然变化不同, 但其保守的氨基酸位置都可以对齐, 特别是胞外区域的 17 个 Cys 都很保守. 在考虑外显子保守性时, 是以其编码的氨基酸是否在比对中落在同一位置为判断标准. 如 *mGluR2* 和 7 中有一个保守的内含子插入位点, 在 *mGluR2* 为第 429 个氨基酸处, 而在 *mGluR7* 为第 459 个氨基酸处. 组内的内含子插入位点都很保守, 但组间的大部分内含子插入位点并不保守, 这说明组内各成员的蛋白质结构域与基因组结构都趋向高度保守.

基因的进化首先是基因功能的进化. 基因功能的进化表现在不同的水平. 蛋白质水平的进化与基因组水平的进化, 或者说基因结构的进化有本质上的不同. 前者导致新的结构域的出现, 而后者则改变对基因表达的调控, 两者又都通过基因的重组和自然选择而遗传下来. 对于比较基因组学研究来说, 植物基因组的进化规律与动物基因组的进化规律会有本质上的不同^[20]. 在 *mGluR7* 所属的第 3 组内, 各个成员在蛋白水平上显得较为保守, 但其基因组大部分内含子增长幅度很大. *mGluR7* 和 8 是 *mGluR* 家族中最迟出现的, *mGluR7* 基因中含有 5 个超过 100 kb 的内含子, 特别是内含子 1 长达 285 kb, 如此长的内含子, 可能在调控基因的表达上有重要作用.

mGluR7 是 *mGluR* 家族中最为保守的亚型, 主要分布在谷氨酸类神经突触的前突触部位, 通过与谷氨酸结合, 抑制谷氨酸的进一步释放^[3]. 适当的 *mGluR7* 表达水平对维护正常神经系统的功能是必不可少的. 细胞中 mRNA 的表达水平主要由启动子类型、转录完整基因所需的时间和其半衰期等因数决定. 对于高度保守的 *mGluR7* 来说, 它的启动子类型为“持家基因”的启动子模式, 因此它在细胞中的 mRNA 水平主要通过转录时间长短来调节. 增大内含子, 可以有效地增加转录完整基因所需的时间, 使细胞中 mRNA 的拷贝数保持在其功能相对应的水平. 如基因组长度为 2.4 Mb 的营养障碍基因 (dystrophin), 完全转录需要 16 h^[23], 其半衰期为 16 h, 以保持在每个细胞含有 8 个营养障碍基因拷贝^[24]. *mGluR7* 基因中 5 个超过 100 kb 的内含子可能发挥这种调节基因表达水平的作用.

致谢 本工作为国家高技术研究发展计划资助项目(批准号: 863-J19)和中国科学院创新工程资助项目(批准号: KSCX1-D4).

参 考 文 献

- Meldrum B, Garthwaite J. Excitatory amino acid neurotoxicity and neurodegenerative disease. *Trends Pharmacol Sci*, 1990, 11: 379~387
- Monaghan D T, Bridges R J, Cotman C W. The excitatory amino acid receptors: their classes, pharmacology and distinct properties in the functions of the central nervous system. *Annu Rev Pharmacol Toxicol*, 1989, 29: 365~402
- Conn P J, Pin J P. Pharmacology and functions of metabotropic glutamate receptor. *Annu Rev Pharmacol Toxicol*, 1997, 37: 205~237
- Makoff A, Pilling C, Harrington K, et al. Human metabotropic glutamate receptor type 7: molecular cloning and mRNA distribution in the CNS. *Mol Brain Res*, 1996, 40: 165~170
- Flor D J, Putten V, Ruegg D, et al. A novel splice variant of a metabotropic glutamate receptor human *mGluR7b*. *Neuropharmacology*, 1997, 36: 153~159
- Phillips T, Makoff A, Murrison E, et al. Immunohistochemical localisation of *mGluR7* protein in rodent and human cerebellar cortex using subtype specific antibodies. *Mol Brain Res*, 1998, 57: 132~141
- Kinoshita A, Shigemoto R, Ohishi H, et al. Immunohistochemical localization of metabotropic glutamate receptor, *mGluR7a* and *mGluR7b*, in the central nervous system of the adult rat and mouse: a light and electron microscopic study. *J Comp Neurol*, 1998, 393: 332~352
- Moroni F, Cozzi A, Pellegrini-Giampietro P, et al. Group 1 and group 2 metabotropic glutamate receptor agents: pharmacological studies *in vitro* and *in vivo*. *European Neuropsychopharmacology*, 1997, 7(Suppl): s94-7-508

- 9 Sambrook J, Fritsch E F, Maniatis T. Molecular Cloning: A Laboratory Manual. 2nd ed. New York: Cold Spring Harbor Laboratory Press, 1989
- 10 Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*, 1998, 8: 186~194
- 11 Ewing B, Hillier L, Wendl M, et al. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res*, 1998, 8: 175~185
- 12 Gordon D, Abajian C, Green P. Consed: a graphical tool for sequence finishing. *Genome Res*, 1998, 8: 195~202
- 13 Lederfein D, Yaffe D, Nudel U. A housekeeping type promoter, located in the 3' region of the duchenne muscular dystrophy gene, controls the expression of Dp71, a major product of the gene. *Hum Mol Genet*, 1993, 2: 1883~1888
- 14 Stowell J N, Craig A M. Axon/dendrite targeting of metabotropic glutamate receptors by their cytoplasmic carboxy-terminal domains. *Neuron*, 1999, 22: 525~536
- 15 Boudin H, Doan A, Jun X, et al. Presynaptic clustering of mGluR7a requires the PICK1 PDZ domain binding site. *Neuron*, 2000, 28: 485~497
- 16 Shigemoto R, Kinoshita A, Wada E, et al. Differential presynaptic localization of metabotropic glutamate receptor subtypes in the rat hippocampus. *J Neurosci*, 1997, 17(19): 7503~7522
- 17 Far O E, Airas J, Wischmeyer E, et al. Interaction of the C-terminal tail region of the metabotropic glutamate receptor 7 with the protein kinase C substrate PICK1. *Eur J Neurosci*, 2000, 12: 4215~4221
- 18 Burge C, Karlin S. Prediction of complete genes structures in human genomic DNA. *J Mol Biol*, 1997, 268: 78~94
- 19 John J H, Bruce S, Stephen R, et al. Creation of genome-wide protein expression libraries using random activation of gene expression. *Nature Biotechnol*, 2001, 19: 440~445
- 20 Wong G K, Passey D A, Huang Y, et al. Is "junk" DNA mostly intron DNA? *Genome Res*, 2000, 10: 1672~1678
- 21 Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*, 1997, 268: 78~94
- 22 Wu S, Wright R A, Rockey P K, et al. Group III human metabotropic glutamate receptor 4, 7 and 8: molecular cloning, functional expression, and comparison of pharmacological properties in RGT cells. *Mol Brain Res*, 1998, 53: 88~97
- 23 Tennyson C N, Klamut H J, Worton R G. The human dystrophin gene require 16 hours to be transcribed and is co-transcriptionally spliced. *Nature Genet*, 1995, 9: 184~190
- 24 Tennyson C N, Shi Q, Worton R G. Stability of the human dystrophin transcript in muscle. *Nucleic Acids Res*, 1996, 24: 3059~3064

(2001-09-17 收稿, 2002-02-06 收修改稿)