成人视网膜假定蛋白基因 ARHP 的克隆及生物信息学分析*

李 峰[1] 蒋卫红[2] 尹志华[1] 杨旭宇[1] 冯湘玲[1] 刘卫东[1] 姚开泰[1]**

（[1]中南大学湘雅医学院肿瘤研究所，长沙 410078；[2]中南大学湘雅医院耳鼻喉科，长沙 410078）

摘要 从 UniGene 库中选取编号为 BG222624 来自人鼻咽组织的表达序列标签（EST）序列，联网到 NCBI 调用 Blast 服务器分析，发现该 EST 序列是一个代表新基因的未知序列. 利用 Blast 检索 GenBank 的 nr 数据库和 EST 数据库，构建 EST 重叠群，联网到 NCBI 的 ORF finder 服务器，分析发现该 EST 重叠群具有完整的阅读框架. 分别在 cDNA 序列阅读框架的起始密码子和终止密码子的两侧设计引物，以人胎脑 cDNA 文库为模板，进行 PCR 扩增，测序确定该基因的 cDNA 全长序列. 该基因 cDNA 序列全长为 1 672 bp，阅读框架位于第 304～1 557 位之间，编码由 417 个氨基酸组成，分子质量为 46.58 ku 的蛋白质，其理论 p*I* 为 4.21. 将蛋白质序列通过 NCBI 的 Blast 服务器进行序列相似性分析，发现该基因编码的蛋白质和成年小鼠视网膜未知蛋白（BAB32214）同源. 经与国际人类基因组命名委员会协商定名为成人视网膜假定蛋白（adult retina hypothetical protein，ARHP），GenBank 登录号为 AY174896. 生物信息学分析表明，该蛋白质可能为一参与转录调控的核蛋白. ARHP 基因定位在染色体 5q35，跨越 35 163 bp，含 4 个外显子和 3 个内含子. 在基因的 5′非翻译区有 2 个 CpG 岛.

关键词 成人视网膜假定蛋白基因，基因克隆，生物信息学分析

学科分类号 Q786

UniGene 是基于实验数据而建立的一个系统. 它自动地将 GenBank 收录的序列划分成许多以基因为分类依据的非冗余序列簇. 每一 UniGene 簇包含代表某一独特基因的许多序列，以及该基因表达的组织类型和图谱定位等相关信息.

以 Homo. sapiens nasopharynx 为关键词检索 UniGene 数据库，获得了 37 个来源于人鼻咽的 UniGene 簇. 通过联网到 NCBI 调用 Blast 服务器进行序列相似性分析，发现了 16 个 UniGene 簇包含代表新基因的未知序列. 我们采用差异 RT-PCR 检测了这些 ESTs 在正常鼻咽组织和鼻咽癌组织中的表达差异，发现其中一个 UniGene 簇（Hs.163725）中的表达序列标签（EST）序列（BG222624）所代表的基因在 40 例正常鼻咽组织、鼻咽癌组织及 4 种鼻咽癌细胞株 CNE-1、CNE-2、HNE-3 和 5-8F 中，仅有 1 例鼻咽癌组织和 5-8F 细胞株表达，其余均无表达. 在淋巴结、胎心、胃、结肠、结肠癌、扁桃体、扁桃体癌、T 细胞淋巴瘤和子宫组织中也都不表达. 但在 12 例经病理诊断为神经胶质瘤的活检组织及 4 例正常大脑组织中均有表达，而且该基因在神经胶质瘤中表达上调，该基因可能与神经胶质瘤相关. 因此，对 BG222624 所代表的基因进行了生物信息学分析并克隆了该基因.

1 材料和方法

1.1 材料

50 × Advantage 2 DNA 聚合酶混合物、人胎脑 cDNA 文库均购自 Clontech 公司. PCR 扩增引物由上海博亚生物技术有限公司合成. PCR 产物测序由上海生工生物技术服务有限公司完成. Taq 酶、连接酶（ligase）、LB 培养基、pUCm-T 载体均购自上海生工生物技术服务有限公司. dNTP 为 Promega 公司产品. 胶回收试剂盒购自上海华瞬生物工程有限公司.

1.2 方法

1.2.1 GenBank 数据库检索和 EST 序列拼接：将 EST 序列 BG222624（种子序列）采用 Blast 软件进行 GenBank 的 nr 数据库和 EST 数据库检索. 选择与种子序列具有较高相似性的序列（匹配序列），将种子序列和匹配序列装配产生新生序列，然后再以此新生序列作为种子序列重复上述过程，直到没有新的匹配序列入选，从而生成最后的新生序列，作为对种子序列的延伸产物（EST 重叠群）[1].

* 国家重点基础研究发展规划项目（973）（G1998051200）.

** 通讯联系人.

Tel：0731-4360094，E-mail：ktyao@fimmu.com

收稿日期：2003-04-15，接受日期：2003-05-28

1.2.2 EST 重叠群的可读框架分析：将上述 EST 重叠群与人类基因组草图进行 cDNA 序列校正（http://www.ncbi.nlm.nih.gov/genome/guide/human/）. 联网到 NCBI 的 ORF finder 服务器，分析该序列是否具有完整的阅读框架.

1.2.3 阅读框架的确定：根据得到的全长 cDNA 的阅读框架，分别在起始密码子和终止密码子段的两侧设计引物（已加入 *Eco*R Ⅰ和 *Not* Ⅰ酶切位点，以利于以后对表达的蛋白质进行研究）. 引物为 pFRW-1（5′ GCAGAATTCAAATGCCTCAGCCTAGTGTAAG 3′）和 pRVS-1（5′ ATGCGGCCGCGTAATACAAAGGTCAGTCTGG 3′）. 我们还在阅读框架内设计了 1 对引物，以验证阅读框架扩增产物的正确性. 引物为 pFRW-2（5′ CAAACAGTTCGGATCCAGATTTC 3′）和 pRVS-2（5′ CATCTTGGTAAGGACTAGAAAAG 3′）. 以人胎脑 cDNA 文库为模板，进行 PCR 扩增. 在 20 μl 的总反应体系中包含下列物质：2 μl 10 × Advantage 2 PCR 缓冲液，0.4 μl 50 × Advantage 2 DNA 聚合酶混合物，0.4 μl dNTPs（10 mmol/L），0.2 μl 20 μmol/L 的引物 pFRW，0.2 μl 20 μmol/L 引物 pRVS，0.8 μl 人胎脑 cDNA 文库的 DNA. 在 PCR 仪上扩增，条件如下：95℃ 预变性 90 s，94℃ 20 s、60℃ 退火 30 s、72℃ 3 min，35 个循环，72℃ 10 min. PCR 扩增后，取 20 μl 产物于 1.5% 琼脂糖凝胶电泳，切取目的片段，胶回收纯化 DNA，溶于 20 μl TE 缓冲液中，通过紫外分光光度计测定 DNA 溶液的 A_{260} 值，进行定量.

取 0.5 μl PCR 产物，pUCm-T 载体溶液 1 μl，T4 DNA 连接酶 14℃ 连接过夜. 采用 $CaCl_2$ 法制备 JM109 感受态菌. 再将连接产物转化进入 JM109 中. 挑选重组克隆，重组克隆在 ABI377-3 自动测序仪上完成[2].

1.2.4 基因 cDNA 序列的限制性酶切位点分析：cDNA 序列限制性酶切位点信息，为基因克隆鉴定和亚克隆提供了重要信息. 我们采用多种软件对成人视网膜假定蛋白（ARHP）基因 cDNA 序列进行了限制性酶切位点分析.

1.2.5 基因的染色体定位和基因组结构分析：将该基因 cDNA 序列与人类基因组草图进行相似性比较，进行基因的染色体定位和确定 cDNA 序列的内含子和外显子结构.

1.2.6 序列同源性及功能区的计算机分析：利用 ExPASy 服务器和 NCBI 的 Blast 软件对 ARHP 基因的 DNA 序列及蛋白质序列进行了序列同源性分析. 对蛋白质的基本性质、功能位点和蛋白质的亚细胞定位进行了预测，用 CpG Searcher 软件搜索CpG 岛.

2 结 果

2.1 GenBank 数据库检索和 EST 序列拼接

以 BG222624 序列为种子序列，采用 Blast 软件进行 GenBank 的 nr 数据库和 EST 数据库检索. 我们得到了 BG222624、BQ072564、BI001001 和 BU675873 共 4 个具有高度相似性的 EST，并将其组装为 1 699 bp 的 EST 重叠群.

2.2 全长 cDNA 序列的获得和阅读框架的确定

利用 NCBI 的 ORF finder 服务器，分析发现该序列具有完整的阅读框架. 根据 EST 序列拼接的 EST 重叠群，在 cDNA 阅读框架的起始密码子和终止密码子的两侧设计引物，以人胎脑 cDNA 文库为模板，进行 PCR 扩增. 其他引物对的扩增结果验证了阅读框架扩增产物的正确性（图 1）. 将 PCR 产物（1 287 bp）克隆到 pUCm-T 载体，转化 JM109，进行 PCR 扩增鉴定（图 2），测序结果与利用 NCBI 的 ORF finder 服务器分析结果一致，从而确定了基因的全长 cDNA 序列（图 3）.

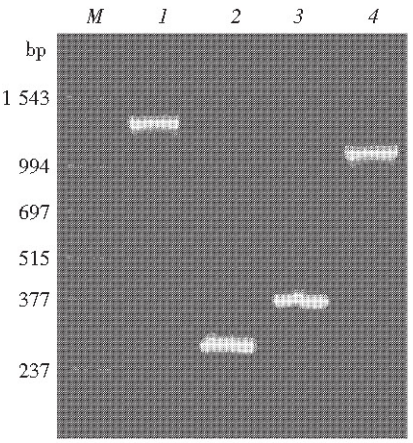

Fig. 1 Identification of PCR products and their correctness

1: pFRW-1 and pRVS-1; *2*: pFRW-2 and pRVS-2; *3*: pFRW-1 and pRVS-2; *4*: pFRW-2 and pRVS-1. *M*: PCR marker.

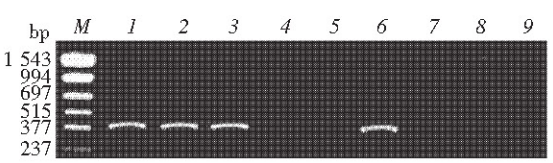

Fig. 2 Identification of recombination clone by PCR amplification using pFRW-1 and pRVS-2

1～*9*: clone number. *M*: PCR marker.

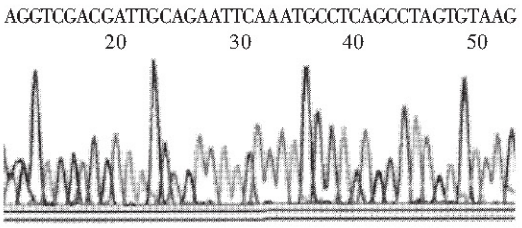

Fig. 3　5′ partial sequencing of pUCm-T/ARHP plasmid

2.3　DNA 序列和基因组结构分析

该基因 cDNA 序列全长为 1 672 bp，命名为 ARHP 基因，GenBank 登录号为 AY174896. 阅读框架位于第 304 ~ 1 557 位之间，编码 417 个氨基酸，预测分子质量为 46.58 ku（图 4）. ARHP 基因组跨越 35 163 bp，含 4 个外显子和 3 个内含子，内含子和外显子交界区符合 ag-gt 规则（表 1）.

```
gggaacccgtcaggaaggacataaacaaaacaaacccgaggcagcatggagaggggccgt

ggcccctgcagcggaaccggacccagtccctgagccgcccctacacccacagacagcatc
                                        *
gcacagaattatttttaaaaaaaagcagtgatccaagcaattgaattggaagcactctggg
                  *           *
gaaacctgctgtttattgtggaaatcatcttcgatcttggaattgaaagtaaagctggaa

aggaatttacaaacaagaaaaaaaagaagtttggaatcggattcacaggatctgggcttg

gaaatgcctcagcctagtgtaagcggaatggatccgcctttcggggatgcctttcgaagc
    M  P  Q  P  S  V  S  G  M  D  P  P  F  G  D  A  F  R  S
cacacctttcggaacaaactctgatgagcacagatctcttagcaaacagttcggatcca
 H  T  F  S  E  Q  T  L  M  S  T  D  L  L  A  N  S  S  D  P
gatttcatgtatgaactggatagagagatgaactaccaacagaatcctagagacaacttt
 D  F  M  Y  E  L  D  R  E  M  N  Y  Q  Q  N  P  R  D  N  F
ctttctttggaggactgcaaagacattgaaaatctggagtctttcacagatgtcctggat
 L  S  L  E  D  C  K  D  I  E  N  L  E  S  F  T  D  V  L  D
aatgagggtgctttaacctcaaactgggaacagtgggatacatactgtgaagacctaacg
 N  E  G  A  L  T  S  N  W  E  Q  W  D  T  Y  C  E  D  L  T
aaatataccaaactaaccagctgtgacatctggggaacaaaagaagtggattacttgggt
 K  Y  T  K  L  T  S  C  D  I  W  G  T  K  E  V  D  Y  L  G
cttgatgacttttctagtccttaccaagatgaagaggttataagtaaaactccaacttta
 L  D  D  F  S  S  P  Y  Q  D  E  E  V  I  S  K  T  P  T  L
gctcaacttaatagtgaggactcacagtctgtttctgattcccttattaccccgattca
 A  Q  L  N  S  E  D  S  Q  S  V  S  D  S  L  Y  Y  P  D  S
cttttcagtgtcaaacaaaatcccttaccctcttcattccctggtaaaaagatcacaagc
 L  F  S  V  K  Q  N  P  L  P  S  S  F  P  G  K  K  I  T  S
agagcagctgctcctgtgtgttcttctaagactctgcaggctgaggtccctttgtcagac
 R  A  A  A  P  V  C  S  S  K  T  L  Q  A  E  V  P  L  S  D
tgtgtccaaaaagcaagtaaacccacttcaagcacacaaatcatggtgaagaccaacatg
 C  V  Q  K  A  S  K  P  T  S  S  T  Q  I  M  V  K  T  N  M
tatcataatgaaaaggtgaacttttcatgttgaatgtaaagactatgtaaaaaaggcaaag
 Y  H  N  E  K  V  N  F  H  V  E  C  K  D  Y  V  K  K  A  K
gtaaagatcaacccagtgcaacagagccggccccttgttgagccagattcacacagatgca
 V  K  I  N  P  V  Q  Q  S  R  P  L  L  S  Q  I  H  T  D  A
gcaaaggagaacacctgctactgtggtgcagtggcaaagagacaagagaaaaaaggggatg
 A  K  E  N  T  C  Y  C  G  A  V  A  K  R  Q  E  K  K  G  M
gagcctcttcaaggtcatgccactcccgctttgccttttaaagaaacccaggaactatta
 E  P  L  Q  G  H  A  T  P  A  L  P  F  K  E  T  Q  E  L  L
ctaagtccctgcccaggaaggtcctgggtcacttgcagcaggagagagcagcagtctt
 L  S  P  L  P  Q  E  G  P  G  S  L  A  A  G  E  S  S  S  L
tctgccagtacatcagtctcagattcatcccagaaaaaagaagagcacaattattctctt
 S  A  S  T  S  V  S  D  S  S  Q  K  K  E  E  H  N  Y  S  L
tttgtctccgacaacttgggtgaacagccaactaaatgcagtcctgaagaagatgaggag
 F  V  S  D  N  L  G  E  Q  P  T  K  C  S  P  E  E  D  E  E
gacgaggaggatgttgatgatgaggaccatgatgaaggattcggcagtgagcatgaactg
 D  E  E  D  V  D  D  E  D  H  D  E  G  F  G  S  E  H  E  L
tctgaaaatgaggaggaggaagaagaggaagaggattatgaagatgacaaggatgatgat
 S  E  N  E  E  E  E  E  E  D  Y  E  D  D  K  D  D  D
attagtgatactttctctgaaccaggtattataatgcttgcaagcttaccagactgacct
 I  S  D  T  F  S  E  P  G  I  I  M  L  A  S  L  P  D  *
ttgtattactattttgaaatagaaaggtttttgtttctgttttgtttggataatttcttt
 attttagtttgggaattaaatgacttaaaccttg(a)18
```

Fig. 4　cDNA and predicted protein sequence of ARHP

Stop codon is indicated by asterisk（*）

Table 1　Exon-intron junction of ARHP gene

Exon	Exon size/bp	5′splice donor	Intron size/bp	3′splice acceptor	Intron
1	112	CCACAG**gt**gagc	24 105		1
2	200	CCTCAG**gt**aaat	5 715	tgac**ag**ACAGCA	2
3	126	GAACTG**gt**aagc	3 689	ttgt**ag**CCTAGT	3
4	1 216			tctc**ag**GATAGA	

Uppercase and lowercase letters indicate exon and intron sequences respectively. Conserved splice donor and acceptor dinucleotide sequences are indicated in bold.

2.4 ARHP 基因的染色体定位

将 ARHP 基因序列与人类基因组草图进行相似性比较（http://www.ncbi.nlm.nih.gov/genome/guide/human/），发现该基因与人类 5 号染色体上的 NT_023132 序列高度相似，将该基因定位于 5q35.

2.5 ARHP 基因 cDNA 的主要限制性酶切位点

图 5 表明 ARHD 基因 cDNA 序列的几个主要限制性酶切位点.

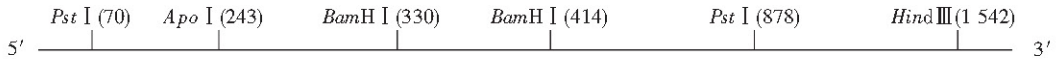

Fig. 5 Main restriction enzyme sites of ARHP cDNA

The number in bracket represents the position of RE sites in cDNA.

2.6 序列同源性及功能区的计算机分析

ExPASy 服务器和 NCBI 的 Blast 软件分析结果表明，在基因的 5′非翻译区有 2 个 CpG 岛. ARHP 基因的蛋白质与一种成年小鼠视网膜未知蛋白（BAB32214）同源（相似性为 70%）. 与其他已知蛋白无明显同源性. 蛋白质的理论 p*I* 为 4.21，为一种酸性蛋白质. 蛋白质在其近羧基端有一疏水区. 生物信息学 Neural Networks 分析表明，该蛋白质可能为一参与转录调控的核蛋白. 蛋白质功能位点分析结果见表 2.

Table 2 Functional sites analysis of ARHP protein

Functional sites	Pattern
N-glycosylation site	N[^P][ST][^P] 35 NSSD 336 NYSL
cAMP- and cGMP-dependent protein kinase phosphorylation site	[RK]{2}.[ST] 175 KKIT
Protein kinase C phosphorylation site	[ST].[RK] 162 SVK 178 TSR 187 SSK 329 SQK
Casein kinase Ⅱ phosphorylation site	[ST].{2}[DE] 7 SGMD 21 TFSE 37 SDPD 61 SLED 73 SFTD 86 SNWE 93 TYCE 105 TSCD 149 SVSD 324 SVSD 353 SPEE 375 SEHE 380 SENE 403 TFSE 414 SLPD
Tyrosine kinase phosphorylation site	[RK].{2,3}[DE].{2,3}Y 331 KKEEHNY
N-myristoylation site	G[^EDRKHPFYW].{2}[STAGCN][^P] 82 GALTSN 309 GSLAAG
Amidation site	.G[RK][RK] 173 PGKK

3 讨 论

基因组信息学的首要任务之一就是发现新基因. 传统试验方法发现新基因极其繁杂和耗时，EST 战略的提出改变了这种状况，大大缩小了研究范围，减轻了工作量，加速了发现新基因的步伐[3,4].

我们通过联网到 NCBI 调用 Blast 服务器对 BG222624 进行序列相似性分析，发现该 EST 序列是一个代表新基因的未知序列. 以 BG222624 序列为种子序列，我们得到了 1 699 bp 的 EST 重叠群. 利用 NCBI 的 ORF finder 服务器，分析发现该序列具有完整的阅读框架. 根据 EST 序列拼接的 EST 重叠群，在 cDNA 阅读框架的起始密码子和终止密码子的两侧设计引物，以人胎脑 cDNA 文库为模板，进行 PCR 扩增，将 PCR 产物克隆到 pUCm-T 载体，测序结果与利用 NCBI 的 ORF finder 服务器分析结果一致，从而通过实验确定了 ARHP 基因的 cDNA 全长序列.

ARHP 基因的进一步生物信息学分析结果表明，ARHP 基因定位在染色体 5q35，基因组跨越 35 163 bp，含 4 个外显子和 3 个内含子. 在基因的 5′非翻译区有 2 个 CpG 岛，是可能的甲基化位点，参与基因的表达调控. 该蛋白质有多个磷酸化、糖基化、肉豆蔻酰基化位点和 1 个酰胺化位点，蛋白质的磷酸化、糖基化和酰基化等具有十分重要的生物学意义. 特别是蛋白质的磷酸化与去磷酸化过程是生物体内普遍存在的信号转导调节方式，几乎涉

及所有的生理及病理过程，如细胞的生长发育、基因表达甚至癌变等等. 研究表明，转录因子在细胞周期调控、细胞增殖及肿瘤发生、转移等生理病理过程中发挥重要作用. ARHP 蛋白可能为一参与转录调控的核蛋白，在神经胶质瘤中表达上调的原因是什么？ARHP 基因是否与神经胶质瘤及视网膜母细胞瘤相关？其具体生物学功能值得深入研究.

参 考 文 献

1 张成岗，贺福初. 生物信息学. 北京：科学出版社，2002. 72～74
Zhang C G, He F C. Bioinformatics. Beijing: Science Press, 2002. 72～74

2 刘上峰，李麓芸，傅俊江，等. 利用巢式 PCR 技术和人类基因组草图搜索法快速获得人类睾丸生精细胞凋亡相关基因 TSARG2. 生物化学与生物物理学报，2002，**34**（3）：378～382
Liu S F, Li L Y, Fu J J, *et al*. Acta Biochem Biophys Sin, 2002, **34**（3）: 378～382

3 王华春，陈清轩. 充分利用 EST 数据库资源. 生物化学与生物物理进展，2000，**27**（4）：442～444
Wang H C, Chen Q X. Prog Biochem Biophys, 2000, **27**（4）: 442～444

4 Adams M D, Kelley J M, Gocayne J D, *et al*. Complementary DNA sequencing: expressed sequence tags and Human Genome Project. Science, 1991, **252**（5013）: 1651～1656

Cloning and Bioinformatics Analysis of a Novel *H. sapiens* Adult Retina Hypothetical Protein Gene ARHP*

LI Feng[1)], JIANG Wei-Hong[2)], YIN Zhi-Hua[1)],
YANG Xu-Yu[1)], FENG Xiang-Ling[1)], LIU Wei-Dong[1)], YAO Kai-Tai[1)]**

([1)] *Cancer Research Institute, Xiangya School of Medicine, Central South University, Changsha* 410078, China;
[2)] *Department of Otolaryngology, Xiangya Hospital, Central South University, Changsha* 410078, China)

Abstract BLAST analysis suggested that a cDNA fragment (GenBank accession number BG222624) derived from human nasopharynx might represent a novel human gene. Applying the bioinformatics and experimental technique, a novel human gene have been cloned from the fetal brain cDNA library. Since this fragment contained a complete open reading frame (ORF) of 1 254 bp with a stop codon in its upstream and poly (A) signal in its downstream, it could be concluded that it is a full-length gene (GenBank accession number AY174896), which was named as adult retina hypothetical protein (ARHP). The full-length cDNA of ARHP gene is 1 672 bp, coding for a 417 amino acids polypeptide with a predicted molecular mass of 46.58 ku and isoelectric point of 4.20. The deduced amino acid showed 70% homology with a *M. musculus* protein BAB32214. Bioinformatics analysis suggested that the protein may be a nucleus protein regulating gene transcription. The new gene is comprised of four exons, with three intervening introns and it is localized to chromosome 5q35. It have been found that there are two CpG islands in 5′ UTR of this gene.

Key words ARHP, cloning, bioinformatics analysis

* This work was supported by a grant from The Special Funds for Major State Basic Research of China (G1998051200).

** Corresponding author. Tel: 86-731-4360094, E-mail: ktyao@fimmu.com

Received: April 15, 2003 Accepted: May 28, 2003