[文章编号] 1007-3949(2005)13-03-0259-04

·实验研究·

胆固醇损伤内皮细胞差异表达基因的生物信息学分析

彭瑾瑜，朱 勋，唐蔚青[1]，王 抒[1]，黎 健[1]，王 仁，郭 芳，刘俊文，杨向东

（南华大学心血管病研究所，湖南省衡阳市 421001；1. 北京医院卫生部老年医学研究所，北京市 100730）

[关键词] 病理学与病理生理学；胆固醇损伤内皮细胞差异表达基因的分析；生物信息学分析；动脉粥样硬化；内皮细胞；胆固醇；细胞色素氧化酶亚基㈦

[摘 要] **目的** 利用差异表达基因克隆方法（抑制消减杂交）获得大量的动脉粥样硬化相关候选基因和表达序列标签后，探讨如何进行后续基因表达及功能的研究。**方法** 利用 Internet 网络上的数据库及生物学分析软件对胆固醇损伤内皮细胞后获得的差异表达基因进行核酸序列和蛋白质序列分析，探索差异表达基因克隆后的研究方法和思路。**结果** 通过电子延伸得到一个 684 bp 的全长 cDNA 序列；通过核酸序列分析，该序列定位在线粒体基因组的 7 587 位～8 270 位，含有一个完整的开放阅读框，编码与氧化磷酸化相关的 9 个亚基。对其中一个亚基细胞色素氧化酶㈦（COX2）分析得知，细胞色素氧化酶㈦基因编码一段 25.6 kDa 的弱酸性的信号锚蛋白，细胞色素氧化酶㈦蛋白的三维结构是一个典型的椅式结构，它含有一段疏水区域，一个跨膜结构域和一个胞质结构域。运用蛋白的进化分析，得知细胞色素氧化酶㈦蛋白胞质结构域的氨基酸序列在进化过程中高度保守。**结论** 生物信息学技术是一种高效的获取疾病相关基因信息的方法，利用生物信息学方法对细胞色素氧化酶㈦基因进行分析，获得了基因及其编码蛋白的相关信息，该蛋白参与电子传递，可能与细胞的氧化应激有关。

[中图分类号] R363 [文献标识码] A

Bioinformatic Analysis of Differentially Displayed Gene in Human Endothelial Cell Induced by Cholesterol

PENG Jin-Yu, ZHU Xun, TANG Wei-Qing[1], WANG Shu[1], LI Jian[1], WANG Ren, GUO Fang, LIU Jun-Wen, and YANG Xiang-Dong

(Institute of Cardiovascular Disease, Nanhua University, Hengyang 421001; 1. Department of Biochemistry, Beijing Institute of Geriatrics, Beijing 100730, China)

[KEY WORDS] Bioinformatics Analysis; Atherosclerosis; Endothelium Cell; Cholesterol; Cytochrome C Oxidase Subunit ㈦ Gene; Differentially Displayed Gene

[ABSTRACT] **Aim** To study the large amount of differentially displayed genes and expressed sequence tags (EST) obtained by gene clone is a difficult problem. It is important to exploit new methods for further study. **Methods** EST were acquired from human umbilical vein endothelial cells induced by cholesterol by suppression subtractive hybridization (SSH). Bioinformatics analysis of atherosclerosis related genes was done by the bioinformatical databases and bio-softwares such as BLAST, ExPasy analyse soft box, DNAMAN soft, BioEdit soft and RasMol soft. **Results** We obtained a 684 bp complete cDNA sequence by electronic clone, which was homogeneous to cytochrome C oxidase subunit ㈦ gene (COX2). The complete genome of COX2 sited in homo sapiens mitochondrion, and included a complete open reading frame (ORF) coding 227 amino acids. The analysis of its protein sequence indicated that COX2 gene encode a 25.6 kDa protein, which was a weak acid signal anchor. The three-dimensional structure of COX2 was a representative chair-like stucture, containing a hydrophobicity region, a transmembrane domain and a periplasmic domain. The periplasmic domain of the COX2 protein sequence was high conservative in the process of evolutions by the analysis of the protein evolutions. **Conclusion** We obtained the information of COX2 about nucleic acid sequence and protein sequence by bioinformatics analysis

动脉粥样硬化是一种多基因遗传病，目前对其相关基因的克隆及功能研究非常重视。用常规的生物学方法对大量的基因进行表达和功能研究，不仅耗资大、工作量大，而且存在研究的盲目性，制约了疾病相关基因在基因组、蛋白质组学方面的研究进度。生物信息学是一门综合运用生物学、数学、物理学、信息科学及计算机科学等诸多学科的理论方法的崭新交叉学科，它应用先进的数据管理技术、数学分析模型和计算机软件对各种生物信息进行获取、处理、存储、分配 、分析和解释，以达到理解数据中的生物学含义的目的。应用生物信息学技术进行基因组和蛋白质组学研究是当前的研究热点，生物信

[收稿日期] 2004-06-24 [修回日期] 2005-02-28

[基金项目] 国家自然科学基金（3002103）；湖南省自然科学基金（02JJY4011）；湖南省教育厅青年基金（02B039）

[作者简介] 彭瑾瑜，E-mail 为 pengjinyu21@yahoo.com.cn；朱勋，均为南华大学生命科学与技术学院 2000 级生物技术专业本科生。**通讯作者**杨向东，博士，副教授，硕士研究生导师，主要从事心血管疾病相关基因的克隆和细胞凋亡机制研究，联系电话 0734-8281297，E-mail 为 XDY7@263.net。

息学在基因克隆、结构分析以及功能预测中发挥重要作用[1-3]。本文利用唐蔚青等[4]采用抑制消减杂交(SSH)克隆的多个胆固醇损伤人脐静脉内皮细胞产生差异表达基因或表达序列标签(expressed sequence-tag, EST),从中挑选出一个EST(GenBank登录号为BI307819)。通过BLAST软件分析获得其全长cDNA序列,该序列包含了与氧化磷酸化密切相关的9个亚基,查询文献后挑选细胞色素氧化酶亚基㊦(COX2)进行生物信息学分析。

1 材料与方法

1.1 GenBank数据库检索和EST序列拼接

经国际互联网进入美国GenBank数据库中,利用BLAST(http://www.ncbi.nlm.nih.gov/BLAST/)检索人nr数据库和EST数据库,对获得的序列进行相似性检索,拼接后的每一个碱基至少经过2条以上的EST序列的验证。利用NCBI的基因组图谱(http://www.ncbi.nlm.nih.gov/mapview/)查询目的基因在人类基因组的定位。

1.2 cDNA序列的开放阅读框分析和人类基因组的定位

将获得的cDNA序列输入NCBI的开放阅读框分析软件ORF finder软件(http://www.ncbi.nlm.nih.gov/gorf/),查询是否具有完整的阅读框。

1.3 基因编码蛋白质的分子质量、等电点和亲疏水性分析

利用互联网上ExPaSy软件包(http://www.expasy.ch/tools/)中Compute pI/MW软件进行蛋白质的氨基酸组成、分子质量和等电点分析。利用BioEDIT软件(http://www.mbio.ncsu.edu/BioEdit/bioedit.html)进行亲疏水性分析。

1.4 基因编码蛋白质结构和功能区分析

利用Tmpred服务器(http://www.cbs.dtu.dk/services/TMHMM-2.0/)对该蛋白的跨膜区进行分析。利用nnPredict软件(http://us.expasy.org/tools/)进行蛋白序列的二级结构分析,向蛋白质立体结构数据库PDB(Protein Data Bank)提交该蛋白质序列,利用RasMol软件显示细胞色素氧化酶亚单位COX2蛋白的三维分子结构。利用PROSITE数据库分析该蛋白序列的模体符(motif),在NCBI中利用CDD(Conserved Domain Database)预测该蛋白结构功能域。

1.5 种属关系分析

利用DNAMAN软件对KOG4767家族蛋白序列多重对齐分析。

2 结 果

2.1 内皮细胞差异表达序列标签的同源性

利用BLAST软件进行核酸同源性比较,发现其与人线粒体基因(登录号为AY495330.1)同源性达98%,提示该EST是人线粒体的一段基因,基因组的定位分析显示,该基因定位在线粒体基因组的7 587~8 270位。通过BLASTn获得线粒体的全长cDNA序列,该序列编码与氧化磷酸化相关的9个亚基,通过PubMed查找相关文献,选取其中一个亚基COX2进行生物信息学分析。

2.2 细胞色素氧化酶亚基㊦基因cDNA序列的开放阅读框

利用NCBI中的ORF finder软件对COX2基因进行开放阅读框分析。该基因的最大开放阅读框是从第1位到第683位碱基,编码227个氨基酸,该开放阅读框起始密码子ATG的+4位碱基为嘌呤碱基G,符合Kozak规律:即阅读框起始密码子ATG的周围序列部分,如果-3位碱基为嘌呤碱基A或G,则是有效的翻译起始位点,否则在+4位必须出现G。表明该COX2有一个的完整阅读框(图1, Figure 1)。

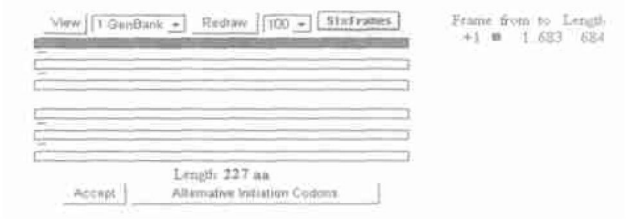

图1. 细胞色素氧化酶㊦基因的开放阅读框分析

Figure 1. ORF analysis of COX2 gene

2.3 细胞色素氧化酶㊦基因编码蛋白的分子质量和等电点

利用ExPaSy软件包中Compute pI/MW软件进行氨基酸组成、分子质量和等电点分析。结果显示,该蛋白为分子质量25.56502品种效益年kDa,等电点为4.67,为一略酸性的蛋白质;该蛋白序列含有较多的中性非极性氨基酸,特别是亮氨酸(Leu)含量达到14.54%,异亮氨酸(Ile)为9.69%,为其二级结构的α螺旋和β折叠的形成打下基础。

2.4 细胞色素氧化酶㊦基因编码蛋白的亲疏水性

利用BioEdit软件对COX2蛋白进行亲疏水性分析,基于BioEdit软件的亲疏水性分析的计算方式,基线上方+1.5以上代表亲水性,基线下方-1.5以下代表疏水性。得知该蛋白在30~43位间以及68~75位间各有一疏水性区域(图2, Figure 2)。

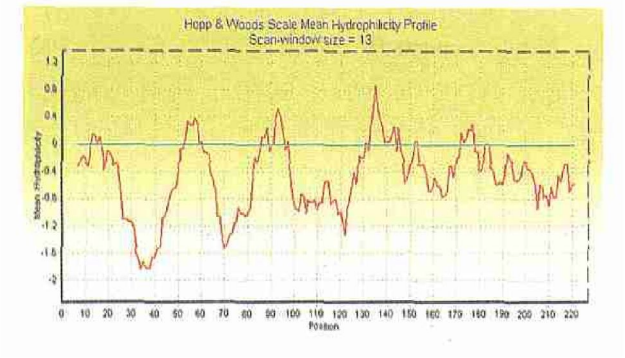

图 2. 细胞色素氧化酶Ⅱ蛋白序列的亲疏水性分析

Figure 2. Hydrophile and hydrophobe analysis of COX2 sequence

2.5 细胞色素氧化酶Ⅱ基因编码蛋白的跨膜区段

利用 Tmpred 服务器对该蛋白的跨膜区进行分析显示: 有两个跨膜区段, 分别在第 26 位~ 第 48 位和第 63 位~ 第 85 位(图 3, Figure 3) 。

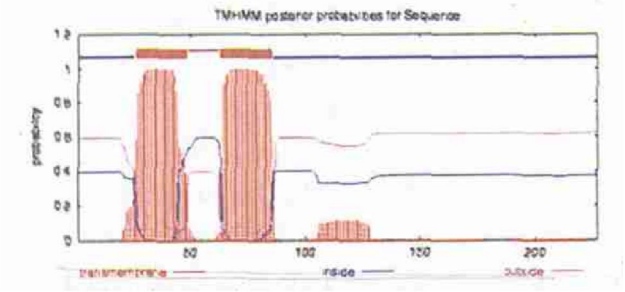

图 3. 细胞色素氧化酶Ⅱ蛋白序列的跨膜区分析

Figure 3. Trans-membrane analysis of COX2 sequence

2.6 细胞色素氧化酶Ⅱ基因编码蛋白的二级结构

用 nnPredict 软件进行蛋白序列二级结构分析显示, 该蛋白序列氨基端的 90 个氨基酸以 α 螺旋为主, 后 130 个氨基酸以 β 折叠为主(图 4, Figure 4) 。

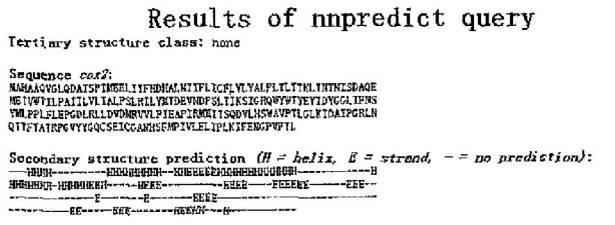

图 4. 细胞色素氧化酶Ⅱ蛋白的二级结构 H 表示 α 螺旋结构, E 表示 β 折叠片

Figure 4. The secondary structure prediction of COX2 sequence

2.7 细胞色素氧化酶Ⅱ基因编码蛋白的结构域

在线(http://smart.embl-heidelberg.de/) 分析蛋白质的结构功能域位点显示, 该蛋白序列跨膜结构域位于 1~ 83 位, 胞浆结构域位于 95~ 226 位。结合二级结构与功能结构域分析, 可知 COX2 蛋白的跨膜结构域主要以 α 螺旋为主, 胞浆结构域以 β 折叠为主。在 NCBI 中利用 CDD (Conserved Domain Database) 预测该蛋白结构功能域显示, COX2 蛋白序列含有一胞质结构域和跨膜结构域, 该蛋白与 KOG4767 家族(此家族含有真核生物细胞色素氧化酶亚单位Ⅱ以及一些与能量产生和传递相关的蛋白) 以及 CyoA 家族(COG 编号为 COG1622, 单细胞生物中血红素/铜型细胞色素氧化酶亚基Ⅱ的同源蛋白蛋白家族) 有较高同源性(图 5, Figure 5) 。

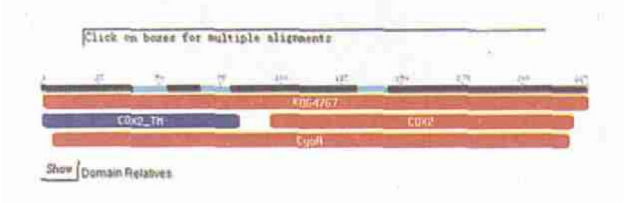

图 5. 细胞色素氧化酶Ⅱ蛋白结构域分析

Figure 5. Structural domains analysis of COX2 sequence

2.8 细胞色素氧化酶Ⅱ基因编码蛋白的模体符

利用 PROSITE 数据库分析该蛋白序列的模体符(motif) 显示, COX2 蛋白序列含有多种不同的模体符, 其中包括双核铜中心和蛋白激酶 C(PKC) 模体符。从 PROSITE 数据库中我们获知 COX2 蛋白胞质结构域中有一个双核铜中心(在 159~ 207 位) , 其特征序列为: V-x-H-x(33, 40) -C-x(3) -C-x(3) -H-x(2) -M。查询蛋白质结构数据库 CATH 数据库, 得知此双核铜中心结构域是一反平行 β 桶结构(希腊钥匙型) , 见图 6(Figure 6) 。

图 6. 细胞色素氧化酶Ⅱ蛋白双核铜中心的三级结构

Figure 6. The thirdly structural domains prediction of COX2 protein

2.9 细胞色素氧化酶Ⅱ基因编码蛋白的三维结构

向蛋白质立体结构数据库 PDB (Protein Data

Bank) 提交该蛋白质序列, 利用 RasMol 软件显示细胞色素氧化酶亚单位 COX2 蛋白的三维分子结构。如图 7(Figure 7) 所示, COX2 蛋白的三维分子构象图呈现一个典型的椅式结构。

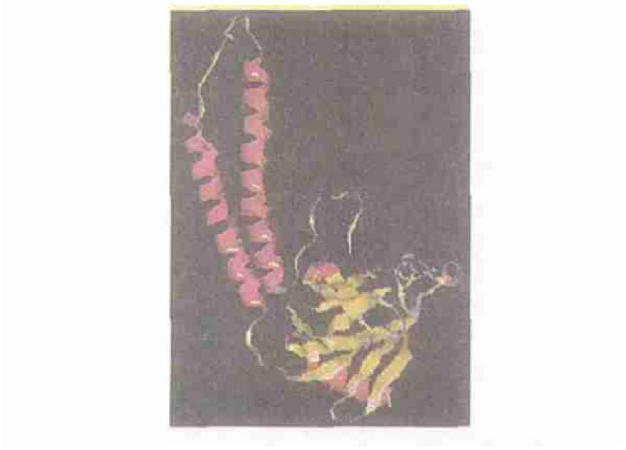

图 7. 细胞色素氧化酶㊣蛋白三维结构示意图 图中深红色弹簧状表示 α-螺旋; 黄色板状表示 β-折叠; 蓝色表示 β-转角; 其他残基颜色为白色

Figure 7. The three-dimensional structural prediction of COX2 protein

2.10 细胞色素氧化酶㊣基因编码蛋白及 KOG4767 家族蛋白序列多重对齐分析

利用 DNAMAN 软件对 KOG4767 家族蛋白序列多重对齐分析发现, 所有参与对比的序列与人的 COX2 的胞质结构域区段的氨基酸显示出高度的序列同源性, 而该区域正是双核铜中心功能域的位置所在, 并且此结构的主链氨基酸 V-x-H-x(33, 40)-C-x(3)-C-x(3)-H-x(2)-M 的序列同源性几乎达 100%, 从而强烈提示 COX2 的双核铜结构域为一重要的功能域[5]。

3 讨 论

本文利用生物信息学方法对 COX2 进行基因序列注释和蛋白质序列的功能注释, 从分析结果中得知, 该蛋白序列是一个分子质量为 25.6 kDa、等电点为 4.67 的略酸性蛋白。COX2 亚基的 N 端, 含有由两个 α 螺旋组成的疏水性跨膜结构域, 该结构可能有助于细胞色素氧化酶与线粒体膜紧密结合。COX2 亚基的胞质结构域中含有一个由反平行 β 桶构成的双核铜中心 CuA, 并且通过序列多重对齐分析, COX2 亚基及 KOG4767 家族蛋白的双核铜中心主链氨基酸有高度同源性, 从而强烈提示此结构为一重要功能域。我们推测反平行 β 桶结构可能有利于形成非特异性的电子通道, 使 COX2 亚基作为细胞色素 C 氧化酶电子入口, 接受来自细胞色素 C 的电子, 经双核铜 CuA 中心传递电子, 从而参与电子传递。同时, 我们发现 COX2 蛋白具有一个 PKC 的 motif, 提示 COX2 胞质域中的离子通道可能受 PKC 的调控, 而 PKC 的激活在细胞增殖和氧化应激的信号传递中具有十分重要的意义[6]。

利用生物信息学一些基本分析方法, 我们获得了大量 COX2 基因及其编码蛋白的相关信息, 为进一步研究 COX2 蛋白与动脉粥样硬化的关系提供了实验依据。在以前的研究中[7], 我们曾用电子克隆(电子延伸) 获得泡沫化细胞差异表达基因的全长 cDNA 序列; 利用生物信息学软件分析了新的人突触相关蛋白(FRG4) 抗原表位, 用于合成抗原多肽制备抗体。尽管近几年有很多生物信息学相关专著出版, 但是该领域知识更新很快, 因此我们认为仍有必要介绍用最新软件和方法进行的分析。在实际操作过程中, 我们也总结了一些经验: 进行生物信息学分析需要较好的计算机基础, 熟悉有关基因、蛋白质分析的基础知识; ㊃由于数据库中数据的生物学功能注释远远落后于自动测序仪产生的大量序列数据. 所以当进行序列同源性分析得到与这类缺乏注释的数据相关的信息时, 其信息的可用性受到一定影响; ㊃生物信息学分析的基础是通过人类基因组计划获得的大量核酸和蛋白质数据以及专门的数据库, 由于数据库及数据处于不断更新的状态, 因此我们用生物信息学分析获得的数据在某些情况下需要进行校正; 极少数的序列可能由于数据库中缺乏相关数据而无法进行分析; 生物信息学分析需要可相互操作的生物信息系统, 各种类型的数据转换工具, 以及不断改进的统计分析方法和优化算法; 生物信息学分析可以成为经典实验的研究指南, 但并不能完全取代经典生物学研究方法, 很多时候生物信息学分析结果要经过实验验证。

[参考文献]

[1] 赵国屏, 等. 生物信息学. 北京: 科学出版社, 2002

[2] 张成岗, 贺福初. 生物信息学方法与实践. 北京: 科学出版社, 2002

[3] 李越中, 闫章才, 高培基. 基因组研究与生物信息学. 济南: 山东大学出版社, 2001

[4] 唐蔚青, 王 抒, 杨向东, 陈保生, 李凌松. 胆固醇诱导血管内皮细胞基因的差异表达. 中国动脉硬化杂志, 2003, **11** (1): 1-4

[5] 李连之, 宋爱新, 黄仲贤. 光谱法研究细胞色素 C 氧化酶 CuA 结构域蛋白的稳定性. 光谱实验室, 2004, **21** (1): 135-137

[6] Ren S, Shatadal S, Shen GX. Protein kinase C-beta mediates lipoprotein-induced generation of PAI-1 from vascular endothelial cells. *Am J Physiol Endocrinol Metab*, 2000, **278** (4): E656-662

[7] 闾宏伟, 杨向东, 何淑雅, 杨永宗. cDNA 文库基础上运用热启动聚合酶链反应末端延伸快速分离全长 cDNA 序列. 中国动脉硬化杂志, 2002, **10** (5): 396-399

(此文编辑 胡必利, 文玉珊)