生物信息学预测与实验验证相结合策略筛选鉴定新的人分泌蛋白基因*

周宇波[1,2] 刘 锋[2] 朱智东[2] 朱 弘[2] 张 新[2] 王志勤[2] 刘建华[1] 韩泽广[2**]
(1 上海交通大学生命科学技术学院 上海 200030 2 国家人类基因组南方研究中心 上海 201203)

摘要 使用生物信息学预测结合实验验证的策略筛选鉴定人新的分泌蛋白基因。用 SignalP、SOSUI、PSORT 和 BLAST 等程序对 UniProt 蛋白数据库进行生物信息学分析,筛选出用于实验验证的 14 个功能未知基因。采用 RT-PCR 方法,克隆得到 14 个基因的全长编码序列,并构建到真核表达载体 pcDNA3.1(-)/Myc-His 质粒。采用蛋白质印迹与免疫荧光分析,检测到其中 7 个基因的表达。除其中一个在细胞核表达外,其余 6 个只在细胞质中表达;其中的 4 个基因的表达产物在细胞培养液中可被检测到,鉴定为 4 个新的分泌蛋白基因。

关键词 生物信息学 分泌蛋白 逆转录-聚合酶链式反应 蛋白质印迹 免疫荧光

分泌蛋白是一类功能非常重要的蛋白,不仅参与了信号途径、形态发生、细胞凋亡、细胞分化、血液凝固、免疫防御、癌症发生等多种过程[1],而且可用作治疗药物或治疗靶点。组织型纤溶酶原激活物、红细胞生成素,肽类激素如生长激素、干扰素和白细胞介素,消化酶等分泌蛋白已成为蛋白药物治疗的主要组成[2]。

寻找与发现新分泌蛋白具有重要意义,目前常用方法包括实验筛选及计算机预测两种。信号肽序列捕捉(SST)是一种常用的大规模实验筛选方法[3,4]。方法的原理是随机克隆 cDNA 的 5′末端部分序列到包含有报告基因的表达载体中,通过相应的检测方法分析判断蛋白分泌与否。另一种实验策略是基于分泌蛋白及膜蛋白基因的 mRNA 主要与多聚核糖体及粗面内质网结合特点,构建分泌蛋白及膜蛋白基因富集的 cDNA 库,结合芯片等方法分析筛选分泌蛋白基因[5~7]。

分泌蛋白不同于细胞质蛋白的惟一共同的特征就是氨基末端信号肽序列。因此针对信号肽特点发展了一些有效的信号肽预测遗传算法[8,9]。特别是随着人类基因组测序的完成,基因组和蛋白质组的数据迅速增加更加促进了信息学预测的研究[10,11]。本文拟采用生物信息学预测和实验验证相结合策略,筛选鉴定新的分泌蛋白基因。

1 材料与方法

1.1 生物信息学预测

下载公共蛋白数据库(UniProt Knowledgebase)(ftp://ftp..ac.uk/pub/databases/uniprot/knowledgebase/)并用网上共享 web 软件 SignalP http://www.cbs.dtu.dk/services/SignalP-2.0/ 预测信号肽;SOSUI http://sosui.proteome.bio.tuat.ac.jp/sosuiframe0.html 预测穿膜区域及疏水性;PSORT http://psort.nibb.ac.jp/ 预测亚细胞定位。综合以上分析结果,筛选出用于实验验证的候选基因。

1.2 基因克隆及表达质粒构建

采用 RT-PCR 的方法从人的脾脏、心脏、脑、小肠、胸腺、骨骼肌、睾丸、肾脏、肝脏及结肠组织中获得基因的编码序列。使用 TRIZOL 试剂(Life Technologies Inc.)抽提总 RNA;逆转录采用 M-MLV 逆转录酶(Promega)系统,均按操作手册执行。根据基因编码序列,用 PrimerExpress™ 1.0 (Applied Biosystems) 软件设计包含有合适酶切位点的引物。RT-PCR 获得目的基因除终止密码子外的全长编码序列,与哺乳动物细胞表达载体:pcDNA3.1(-)/Myc-His (Invitrogen) 酶切,连接,转化大肠杆菌 DH5α。菌落 PCR 筛选阳性克隆,酶切鉴定,DNA 测

收稿日期:2004-08-03 修回日期:2004-09-23
* 国家"863"计划资助项目(2002BA711A01-02)
** 通讯作者,电子信箱:hanzg@chgc.sh.cn

序确定。

1.3 细胞转染及蛋白质印迹分析

生长状态良好的 CHO 细胞, 以 1~ 1.5 × 10^5/35mm 培养皿密度接种到新的培养皿中; DMEM/10% FCS, 37℃, 5% CO_2 培养过夜。构建的表达质粒用 LipofectAMINE 试剂 (Gibco BRL) 瞬时转染 CHO 细胞。72h 后, 收集细胞培养上清, 裂解细胞, TALON 树脂 (Clontech) 浓缩样品, 15% SDS-蛋白胶分离。电转至 PVDF 膜上, 3% 脱脂牛奶/2% BSA/PBST 封闭, 室温 4h; 一抗 C-Myc (9E10) (Clontech) 反应, 室温 2h; 二抗 HRP-羊抗鼠 IgG (Gibco BRL) 反应, 室温 2h; 最后用检测试剂 ECL (Amersham Pharmacia Biotech, Ltd.) 反应, 暗室曝光。

1.4 免疫荧光分析

CHO 细胞以 1~ 1.5 × 10^5/35mm 培养皿密度接种到置有盖玻片的新培养皿中, 转染表达质粒。转染后 60h 取出培养皿, 洗涤, 加入 2% PFA (w/v) / 0.1% Tritorr-X-100/PBS 1ml, 冰上 30min 固定; 5% BSA/PBS 封闭, 室温 2h; 玻片的一半加 40~ 50μl 5% BSA/PBS 用作阴性对照, 另一半加一抗 C-Myc (9E10), 置湿盒内, 4℃过夜; 盖玻片的两半都加二抗 Cy2-驴抗鼠 IgG (H + L) (Jackson ImmunoResearch Laboratories), 4℃, 2h; 洗涤, 封片; 荧光显微镜观察。

2 结 果

2.1 信息学预测潜在分泌蛋白基因

本工作流程见图 1。用 SignalP v.2.0 对公共蛋白库进行预测, 以神经网络模型 (neural network models) 信号肽得分 (s-score) 为域值, 获得 7800 条序列。其中包含大量的已知功能基因, 包括已知的分泌蛋白。去除已知功能基因, 再用 SOSUI 预测剩余序列的疏水性及跨膜区域, 选取无跨膜区, 疏水性分值为负值或接近于 0 的序列, 共 198 条。PSORT 预测域值为至少在胞外有 30% 可能性存在, 最后保留 59 条序列。序列用 BLAST 在非冗余库 (nonredundant database) 验证, 去除不可信及功能已知序列, 最后选取 14 条序列用于实验验证。基因号 AL080121, BC013294, AL136580, BC029149, BC004336, AF231922, AF217970, AF151901, BC005069, BC032339, BC040113, AA452778, AY358591, AK074643 分别对应于 NSP059, NSP060, NSP063, NSP066, NSP070, NSP075, NSP078, NSP079, NSP081, NSP82, NSP083, NSP086, NSP088, NSP090 基因 (为方便, 实验室暂命名)。

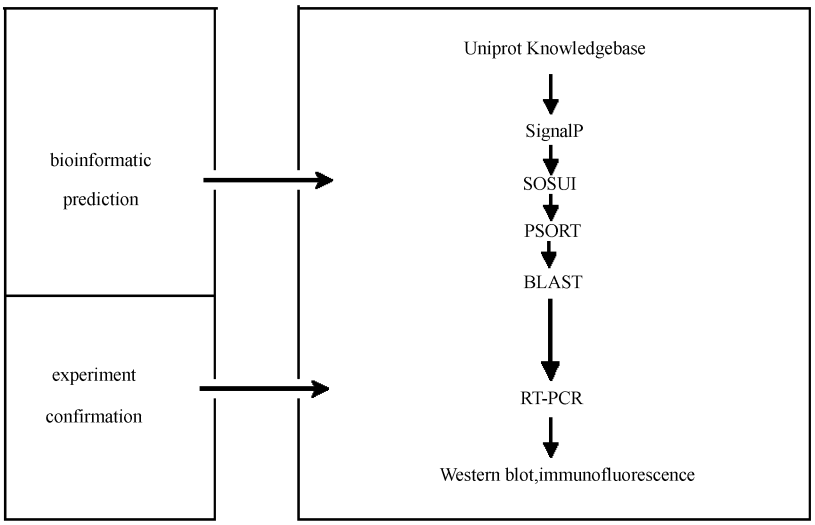

图 1 筛选鉴定人新的分泌蛋白基因策略

Fig. 1 Schematic description of combined strategy for isolating and identifying novel secreted protein genes

2.2 潜在分泌蛋白基因克隆

14 个基因的编码序列用 RT-PCR 的方法从人的脾脏, 心脏, 脑, 小肠, 胸腺, 骨骼肌, 睾丸, 肾脏, 肝脏及结肠组织中获得 (图 2)。构建的重组哺乳动物细胞表达质粒表达的融合蛋白 C 端有 Myc 和 His 标签, 分别用于蛋白的蛋白质印迹分析及免疫荧光分析和样品的浓缩 (引物序列如需要可提供)。

2.3　蛋白表达检测

免疫荧光分析用于检测融合基因的表达及蛋白的亚细胞定位。重组表达质粒瞬时转染 CHO 细胞,60h 后, 针对 Myc 标签的一抗 9E10 与细胞反应, 二抗使用偶联了荧光分子 Cy2 的驴抗鼠 IgG, 荧光显微镜观察, 共检测到 7 个基因 (*NSP060*, *NSP066*, *NSP079*, *NSP081*, *NSP082*, *NSP088*, *NSP090*) 表达(图 3)。其中除 *NSP060* 基因的蛋白主要在细胞核内表达外, 其它 6 个只在细胞质中表达。

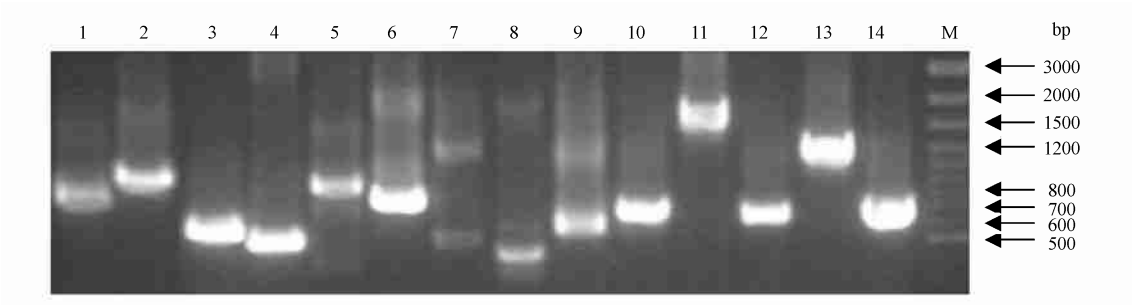

图 2　14 个候选基因克隆

Fig. 2　clone of 14 genes isolated from UniProt Knowledgebase

Lane *1*, *2*, *3*, *4*, *5*, *6*, *7*, *8*, *9*, *10*, *11*, *12*, *13*, *14* and *M* correspond to *NSP059*, *NSP060*, *NSP063*, *NSP066*, *NSP070*, *NSP075*, *NSP078*, *NSP079*, *NSP081*, *NSP082*, *NSP083*, *NSP086*, *NSP088*, *NSP090* genes and marker, respectively

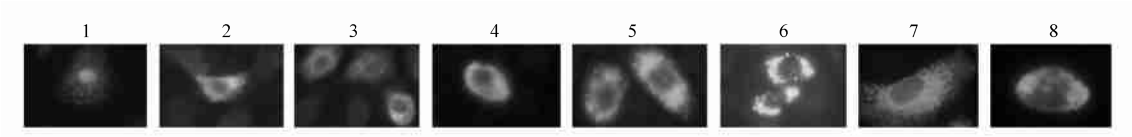

图 3　免疫荧光检测候选基因在 CHO 细胞中的表达

Fig. 3　Immunofluorescence assay of 7 genes expressed in CHO cells

Transfected CHO cells were stained with anti-G-Myc antibody and visualized with a Cy2-conjugated antibody

Lane *1*, *2*, *3*, *5*, *6*, *7* and *8* correspond to *NSP060*, *NSP082*, *NSP088*, *NSP066*, *NSP079*, *NSP090* and *NSP081* gene, respectively. Lane *4* indicates the human growth hormone gene as the positive control. Except *NSP060* protein was localized mainly in nucleolus and minor in cytoplasm, the others were all localized in cytoplasm

2.4　蛋白质印迹检测表达蛋白的分泌

为验证预测的潜在分泌蛋白基因是否分泌, 裂解转染的 CHO 细胞、收集细胞培养液、浓缩、蛋白电泳, 并用蛋白质印迹分析检测。7 个免疫荧光检测为阳性的基因中, 只有 5 个能被蛋白质印迹分析检测。除 *NSP060* 和 *NSP082*, 其它 5 个基因均能检测到表达。其中 *NSP066*、*NSP079*、*NSP081*、*NSP088* 表达的蛋白均可在细胞培养液中检测到(图 4)。

3　讨　论

分泌蛋白的筛选包括了实验与计算机两种主要预测方法。实验预测, 如信号肽捕获(SST)、分泌蛋白 mRNA 富集等方法成功获得一些新的分泌蛋白基因。尤其是 SST 法, 从 1993 年提出以来, 不断有新的改进与创新, 方法日渐成熟。但工作量大, 操作复杂, 周期长, 因为常常得到一些已知分泌蛋白, 对筛选未知分泌蛋白基因效率较低, 而且对低丰度表达及有时相性表达的分泌蛋白基因更不易捕捉。随着算法的发展及研究数据的积累, 信号肽预测, 蛋白定位等基因信息的预测也趋于更方便、快捷、准确。GO (genome Ontology) 分类 (http://www.geneontology.org./) 中包括‘细胞外’关键词的序列也达到 4651 条。SignalP 综合了神经网络模型和隐藏的马科夫模型预测准确性可达到真核生物为 78%, 原核生物为 89%, 结合 TargetP、PSORT 等预测细胞定位的分析程序, 使得基于氨基酸序列预测未知分泌蛋白的准确性更进一步提高。

我们最终目的是研究单个新分泌蛋白基因功能, 希望提高筛选效率。所以筛选对象选择易获得的数据最全公共蛋白数据库, 在信息学预测方面筛选标准更严格, 与文献报道[10,11] 不同的是组合使用

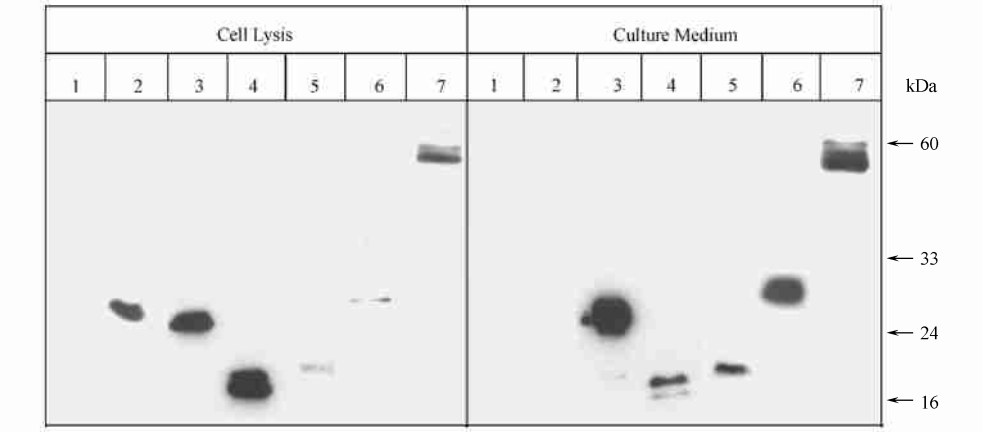

图 4 蛋白质印迹检测基因表达产物在 CHO 细胞和培养液中的分布

Fig. 4 Western blot of 5 genes expressed in CHO cells and culture medium

Expressed proteins were detected with anti-C-Myc antibody (9E10). Lane *2*, *4*, *5*, *6*, *7* correspond to *NSP090*, *NSP079*, *NSP066*, *NSP081* and *NSP088* gene, respectively. Lane *1* indicates the pcDNA3.1(-)/Myc-His as the negative control. Lane *3* indicates the human growth hormone gene as the positive control

了不同的预测软件如 SignalP, SOSUI, PSORT 等。无论用何种策略,候选基因都需要通过各种实验验证,即使 SST 方法获得的阳性结果仍需要克隆全长基因证明预测的正确性。所以本实验室确立了信息学预测,实验验证的技术路线(图 1)。初步的实验结果也证实了方法的简便、快速、准确特点。

14 个候选基因中仅有 *NSP082*, *NSP083*, *NSP088* 和 *NSP090* 被 SPDI[10] 预测为分泌蛋白基因,说明我们生物信息学预测结果的独特和可信。目前只有 *NSP088* 被实验验证,虽然实验验证不分泌并不能肯定该基因一定不是分泌蛋白基因,但已验证的基本可以确定为分泌蛋白。通常分泌蛋白不会定位到细胞核,所以 NSP060 蛋白的亚细胞定位暗示 *NSP060* 基因不是分泌蛋白基因。亚细胞定位与蛋白质印迹分析检测相结合的检测方法会进一步提高结果的可靠性。

本方法的限速步骤是候选基因的克隆。原因有以下几点:引物设计不合理;PCR 条件不合适;RT 的组织来源不合适和基因自身的表达特点等都会影响到基因克隆进展。将人的心、肝、脾、肺、肾、脑、小肠、结肠、胸腺、骨骼肌、睾丸等组织的 cDNA 等量混合,采用梯度 PCR 方法在一定程度上简化了 PCR 的条件,避免了组织来源不适当等问题,相对提高了克隆效率。

构建的 14 个表达基因载体中有 7 个不能检测到表达,经测序分析表明序列完全正确,没有移码,编码框没有提前中止。可能原因有:CHO 细胞不适合目的基因的表达或表达量低;商品化的表达载体并不适合所有基因;检测表达的方法灵敏度有限等。4 个分泌蛋白的表达特点各不一样(图 4),有的以分泌型为主,有的主要以细胞内形式存在,有的表达蛋白条带不止一条。但通常表观分子量较理论计算的分子量大,这是因为蛋白翻译后的各种修饰,特别是糖基化修饰造成。

用本实验室建立的生物信息学预测与实验验证结合的筛选策略,从公共蛋白库中克隆了 14 个潜在分泌蛋白,采用蛋白质印迹与免疫荧光分析,检测到其中 7 个基因的表达,目前初步确定其中的 4 个为新的分泌蛋白(GenBank accession no: BC029149, AF151901, BC005069, AY358591),证明了方法的可行性。这些新的分泌蛋白基因可能在生物过程中扮演重要角色,也可能是潜在的治疗靶点或药物。

参考文献

[1] Klein R D, Gu Q, Goddard A, et al. Selection for genes encoding secreted proteins and receptors. Proc Natl Acad Sci U S A, 1996, 93: 7108~7113

[2] Grabley S, Thiericke R. Bioactive agents from natural sources: trends in discovery and application. Adv Biochem Eng Biotechnol, 1999, 64: 101~154

[3] Tashiro K, Tada H, Heilker R, et al. Signal sequence trap: a cloning strategy for secreted proteins and type I membrane proteins. Science, 1993, 261: 600~603

[4] Kojima T, Kitamura T. A signal sequence trap based on a

constitutively active cytokine receptor. Nat Biotechnol, 1999, 17: 487~ 490

[5] Egland K A, Vincent J J, Strausberg R, et al. Discovery of the breast cancer gene BASE using a molecular approach to enrich for genes encoding membrane and secreted proteins. Proc Natl Acad Sci U S A, 2003, 100: 1099~ 1104

[6] Kopczynski C C, Noordermeer J N, Serano T L, et al. A high throughput screen to identify secreted and transmembrane proteins involved in Drosophila embryogenesis. Proc Natl Acad Sci U S A, 1998, 95: 9973~ 9978

[7] Diehn M, Eisen M B, Botstein D, et al. Large-scale identification of secreted and membrane-associated gene products using DNA microarrays. Nat Genet, 2000, 25: 58~ 62

[8] Nielsen H, Brunak S, von Heijne G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein Eng, 1999, 12: 3~ 9

[9] Ladunga I. PHYSEAN: PHYsical SEquence ANalysis for the identification of protein domains on the basis of physical and chemical properties of amino acids. Bioinformatics, 1999, 15: 1028 ~ 1038

[10] Clark H F, Gurney A L, Abaya E, et al. The secreted protein discovery initiative (SPDI), a large-scale effort to identify novel human secreted and transmembrane proteins: a bioinformatics assessment. Genome Res, 2003, 13: 2265~ 2270

[11] Grimmond S M, Miranda K C, Yuan Z, et al. The mouse secretome: functional classification of the proteins secreted into the extracellular environment. Genome Res, 2003, 13: 1350~ 1359

Isolate and Identify Human Novel Secreted Protein Genes with Combined Strategy of Bioinformatics Prediction and Experimental Confirmation

ZHOU Yu-bo[1,2]　LIU Feng[2]　ZHU Zhi-dong[2]　ZHU Hong[2]　ZHANG Xin[2]
WANG Zhi-qin[2]　LIU Jian-hua[1]　HAN Ze-guang[2]
(1 College of Life Science and Biotechnology, Shanghai Jiaotong University, Shanghai 200030, China)
(2 Chinese National Human Genome Center at Shanghai, Shanghai 201203, China)

Abstract The strategy of bioinformatics prediction combined with experimental confirmation was used to isolate and identify novel potential human secreted protein genes. After bioinformatics prediction with SignalP, SOSUI, PSORT and BLAST programs on the UniProt Knowledgebase, 14 genes were isolated for experimental identification. With RT-PCR amplification, all full coding sequences of these genes were cloned into mammalian cell expression plasmid, pcDNA3.1 (-)/Myc-His vector. 7 genes were successfully expressed, judged by western blot and immunofluorescence analysis, in which one gene was expressed in nucleolus, the other six genes were expressed only in cytoplasm, and 4 expressed proteins of which were secreted into the culture medium and identified as novel secreted proteins.

Key words Bioinformatics　Secreted protein　RT-PCR　Western blot　Immunofluorescence