利用烟草基因组 DNA 构建近随机多肽文库

Utilizing Tabacco Genomic DNA to Construct Nearly Random Peptide Libraries

马素参，黄海明，高友鹤*

MA Su-Can, HUANG Hai-Ming and GAO You-He*

中国医学科学院基础医学研究所中国协和医科大学基础医学院蛋白质组学研究中心，医学分子生物学国家重点实验室，北京 100005

Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences, School of Basic Medicine, Peking Union Medical College, Proteomics Research Center, Beijing 100005, China

摘 要 介绍一种新的方法构建近随机多肽文库。选取从大基因组物种的组织或细胞中提取的基因组 DNA，利用切割频率高的限制性内切酶切割，产生的短片段可以近似地认为是随机序列的片段，将它们与匹配的载体连接后转化进宿主细胞进行表达，从而获得近随机多肽文库。这样的文库可以用于蛋白质相互作用的研究。同一种基因组 DNA 可以利用不同的酶切，再分别连接到表达载体的不同读码框架，从而产生不同编码序列的多种近随机多肽文库。介绍了充分利用烟草基因组 DNA 构建两种不同酶切，三种读码框架，共六种不同编码序列的近随机多肽文库的方法。

关键词 烟草，基因组 DNA，近随机多肽文库

中图分类号 Q785 **文献标识码** A **文章编号** 1000-3061(2005)02-0332-04

Abstract We developed a novel method for constructing nearly random peptide library. Genomic DNAs extracted from tissue or cells of large genome species were digested with frequent cutter to produce short DNA fragments. These short fragments can be considered nearly random. Nearly random peptide libraries can be constructed by cloning the short fragments into appropriate expression vectors and transformation into host cells. Genomic DNA from one species can be digested with different restriction enzymes and ligated to different reading frames to produce several different libraries. In this study, we digested tobacco genomic DNA with two enzymes and cloned into three different reading frames to make totally six nearly random peptide libraries.

Key words tobacco, genomic DNA, nearly random peptide library

随机多肽文库，在研究蛋白质配体/受体相互作用、酶作用底物的分析、寻找具有生物功能的蛋白的模拟肽以及新药物的筛选等方面发挥着越来越大的作用。

目前，主要有两类构建随机多肽文库的方法。第一类方法是体外化学合成[1-3]，第二类方法是利用合成随机的寡聚核苷酸在体内表达成可溶性的融合蛋白[4-6]，但上述两类方法都有其局限性。体外人工合成多肽设计起来虽然较简单，但是要用到昂贵的仪器和复杂的方法，工作费时费力，而且

Received: October 12, 2004; Accepted: December 6, 2004.

This work was supported by Grants from the National Natural Sciences Foundation of China (No.3037030, 30270657), The National Basic Research Progrem of China (No.2004CB520804), Pilot Study for Key Basic Research Project (No.2002CCA04100) and National Natural Sciences Foundation Key Project (No.30230150).

* Corresponding author. Tel: 86-10-65296407; E-mail: gaoyouhe@pumc.edu.cn

国家自然科学基金(No.3037030, 30270657)，国家重点基础研究发展计划(No.2004CB520804)，重大基础研究前期研究专项项目计划(No.2002CCA04100)，国家自然科学基金重点项目(No.30230150)。

成本较高，合成的文库不能反复利用。用合成寡聚核苷酸在体内表达多肽的方法虽较好地解决了上述问题，但是它也有一些固有的缺点，如寡聚核苷酸序列设计起来较复杂，需要考虑简并密码子、终止密码子以及尽量避免形成回文结构等；在实验进行过程中还有退火、补平末端和酶切连接等过程，这些都会对实验的成功造成很大影响。此外，用以上方法所构建的随机多肽库的长度基本上限定在 6～8 个氨基酸，并且一种方法只能构建一种长度的文库。

我们最近开发了一种新的构建近随机多肽文库的方法，即利用切割 4 个核苷酸序列的限制性内切酶切碎基因组 DNA，与合适的载体连接，转化进宿主进行表达，从而获得随机多肽文库。该方法简单易行[7]，原理如下：用识别 4 个核苷酸的限制酶（如 *Dpn* Ⅱ）彻底切割基因组 DNA（如烟草基因组 DNA，3.7×10^9 bp[8]），理论上产生平均长度为 256bp（$4^4=256$，即 *Dpn* Ⅱ 每隔 256bp 有一个识别位点）的 DNA 片段共约 10^7 条（$3.7\times10^9/256$），这样长的 DNA 可以编码约 85.3（256/3≈85.3）个氨基酸，而平均每 64 个氨基酸密码子中有 3 个终止密码子，所以每个片段中平均应有 4 个终止密码子。因此，这样的 DNA 片段连在表达载体（质粒载体或噬菌体载体等）上翻译的时候，因为终止密码子的分布是相对随机的，多肽表达会在翻译整个片段结束前随机终止。这样就产生了平均长度在 21.33（64/3）个氨基酸残基的，长度不一、序列各异的近随机多肽文库。理论上这样的文库不是完全随机的，但是在生物学研究中可以替代随机文库使用，所以我们称它为近随机多肽文库。如果用识别 3 个核苷酸的限制性内切酶，可以更充分利用基因组 DNA，产生平均 DNA 长度为 64 个核苷酸，平均多肽长度仍为 21.33 个氨基酸残基。按此方法建立的随机多肽文库不仅具有现有随机多肽文库的特征，并具有现有随机多肽文库不具备的一些特征，如可产生不等长的随机多肽等（肽的长度从 1～21aa 都有），即不仅有序列上的随机性还有长度上的随机性，是更广泛意义上的随机多肽文库。通常可以利用此随机多肽文库中筛选出蛋白质特异的多肽结合序列等。

在本文中，我们还通过改变载体的读码框架，从而达到用同一种内切酶切割同一种基因组 DNA，连入同一类载体但却获得三种不同编码序列的随机多肽文库的目的，更充分地利用了基因组资源。同时我们用另一种不同的内切酶还可以成倍扩大这个效应。同样的烟草基因组被利用 6 次产生出 6 种不同的近随机文库。

1 材料与方法

1.1 材料

烟草基因组 DNA 由首都师范大学生物系印莉萍教授实验室惠赠，载体质粒 pGADT7 购自 Clontech Laboratories Inc，大肠杆菌菌株 DH10B 由本单位生物化学与分子生物学学系蒋澄宇教授课题组提供。限制性内切酶 *Dpn* Ⅱ、*Tsp*509 Ⅰ、T4 DNA Polymerase，Mung Bean Nuclease，小牛肠碱性磷酸酶（CIP）购自美国 New England Biolabs 公司。T4 DNA ligase 购自美国 Promega 公司。*Bam*H Ⅰ、*Eco*R Ⅰ、*Hind*Ⅲ 等限制性内切酶购自 TaKaRa Biotechnology（大连）公司。DNA Marker、质粒提取试剂盒、PCR 产物纯化试剂盒购自北京天为时代科技有限公司。

1.2 方法

1.2.1 载体质粒的修饰与改造：在本文里，为了构建可用于酵母双杂交的近随机多肽文库，我们选择了 Clontech 公司的 pGADT7 为载体。pGADT7 含有一个 GAL4 的激活结构域（Activation Domain，AD），它可以表达连在它 C-端的融合蛋白。在酵母中，融合蛋白在启动子 P_{ADH1} 的作用下高效表达，接着此融合蛋白在 SV40 和定位信号的作用下进入酵母细胞核。它还有一个 T7 启动子，HA 表位标签和一个多克隆位点。在大肠杆菌和酵母中它分别以 pUC 和 2μori 复制子进行自我复制。

为了得到不同读码框架的 pGADT7 载体，我们对该载体的 *Eco*R Ⅰ 位点分别进行了不同的修饰与改造，使 *Dpn* Ⅱ 切割的基因组 DNA 在载体匹配粘性末端的 *Bam*H Ⅰ 位点连接（另一对内切酶 *Tsp*509 Ⅰ 与 *Eco*R Ⅰ 具有相同的粘性末端）。（1）对 pGADT7 载体 *Bam*H Ⅰ 位点前的 *Eco*R Ⅰ 位点进行修饰，使载体 *Bam*H Ⅰ 位点插入序列的读码框架改变：（i）pGADT7 用 *Eco*R Ⅰ 切割后补平 5′突出末端：即用限制酶 *Eco*R Ⅰ 切割载体后利用 T4 DNA Polymerase 补平 5′突出的 4 个碱基，然后再用 T4 DNA ligase 将补平的载体两端连接起来。这种修饰使载体 *Bam*H Ⅰ 位点插入片段的读码框架向后移动一个碱基，我们记为 pGADT7（+B）。（ii）pGADT7 用 *Eco*R Ⅰ 切割后削平 5′突出末端：即用限制酶 *Eco*R Ⅰ 切割载体后利用 Mung Bean Nuclease 削去 5′突出的 4 个碱基，然后再用 T4 DNA ligase 将削平的载体两端连接起来。这种修饰使载体 *Bam*H Ⅰ 位点插入片段的读码框架向前移动一个碱基，我们记为 pGADT7（-B）。载体经过上面两种不同方式的修饰后，当构建烟草 *Dpn* Ⅱ 文库时，利用 pGADT7、pGADT7（+B）、pGADT7（-B）这三种质粒，就可以得到三种读码框架不同的随机多肽文库。

（2）pGADT7 载体 *Eco*R Ⅰ 位点的改造：载体 pGADT7 的 *Eco*R Ⅰ 位点前没有合适的可供修饰的酶切位点，为了改变 *Eco*R Ⅰ 位点插入片段的读码框架，我们对载体的 *Eco*R Ⅰ 位点进行了适当的改造，即人工化学合成两段寡聚核苷酸，序列分别为 5′-AATTAAGCTTGGG-3′ 和 5′-AATTAAGCTTG-3′，同时合成两段与此两条链互补的反义链，序列为 5′-AATTCCCAAGCTT-3′ 和 5′-AATTCAAGCTT-3′，经退火后形成的双链分别是：

（Ⅰ）5′-AATTAAGCTTGGG-3′
　　3′-TTCGAACCCTTAA-5′
（Ⅱ）5′-AATTAAGCTTG-3′
　　3′-TTCGAACTTAA-5′

此双链有两个与 *Eco*R Ⅰ 相匹配的粘性末端，经 5′端磷酸化后，可以克隆进 pGADT7 质粒的 *Eco*R Ⅰ 位点。插入寡聚核苷酸（Ⅰ）后可使 *Eco*R Ⅰ 位点插入片段的读码框架向后移

动一个碱基，我们记为 pGADT7(+E)。插入寡聚核苷酸(Ⅱ)后可使 *Eco*R Ⅰ位点插入片段的读码框架向前移动一个碱基，我们记为 pGADT7(−E)。插入载体的寡聚核苷酸中间设计了一个 *Hind*Ⅲ酶切位点，用于克隆后鉴定此双链是否连入载体。改造后的载体分别进行 DNA 序列测定验证。利用 pGADT7、pGADT7(+E)、pGADT7(−E)这三种质粒，当构建烟草 *Tsp*509 Ⅰ文库时，就可以得到三种不同编码序列的随机多肽文库。

1.2.2 烟草基因组 DNA 经 *Dpn* Ⅱ，*Tsp*509 Ⅰ分别酶切构建近随机多肽文库：

质粒载体的制备：将 pGADT7、pGADT7(+B)和 pGADT7(−B)三种质粒分别用限制酶 *Bam*H Ⅰ酶切；pGADT7、pGADT7(+E)和 pGADT7(−E)分别用限制酶 *Eco*R Ⅰ酶切，经小牛肠碱性磷酸酶(CIP)将酶切产物去磷酸化后利用 PCR 产物纯化试剂盒回收酶切片段。

插入片段的制备：选用与 *Bam*H Ⅰ具有匹配粘性末端的限制酶 *Dpn* Ⅱ和与 *Eco*R Ⅰ具有匹配粘性末端的限制酶 *Tsp*509 Ⅰ分别随机切割烟草基因组 DNA，利用 PCR 产物纯化试剂盒回收酶切片段。

将制备好的质粒载体和烟草基因组 DNA 片段在 16℃过夜连接(*Dpn* Ⅱ切割的基因组片段与 *Bam*H Ⅰ切割的载体相连；*Tsp*509 Ⅰ切割的基因组片段与 *Eco*R Ⅰ切割的载体相连)。连接产物脱盐后电击转化，细胞复苏后取少量菌液铺板，鉴定文库质量并测定文库容量，其余则加入甘油至终浓度 20%，于 −80℃保存，即为构建好的基因组文库。共得到六种编码序列不同的烟草近随机多肽文库。

1.2.3 文库容量的测定：取 1μL 电击转化后复苏的文库菌液，置于 99μL 新鲜 LB 液体培养基中，混匀后再取出 10μL 稀释后的文库菌液置于 90μL 新鲜 LB 液体培养基中，将其全部均匀涂布于 LB 固体培养基(Amp^r)中，37℃倒置培养过夜。对培养皿中的菌落数 N 计数，按下列公式计算文库容量：

文库容量(cfu) = (N/100μL) × 10^3(稀释倍数) × 10^3(μL/mL)

1.2.4 文库质量的鉴定：

烟草 *Dpn* Ⅱ文库的鉴定：因为在 pGADT7 多克隆位点两端分别有一个 *Hind*Ⅲ识别位点，它们之间的长度恰好为 800bp，如果在多克隆位点中有外源片段的插入，插入片段中无 *Hind*Ⅲ位点，用 *Hind*Ⅲ单酶切后将产生一条大于 800bp 的酶切片段。插入片段中如果存在 *Hind*Ⅲ位点，用 *Hind*Ⅲ单酶切后将会产生两条以上的酶切片段，酶切片段长度之和应大于 800bp。因此，我们在用 *Dpn* Ⅱ所建的文库中随机挑取数个单克隆，提取质粒后用 *Hind*Ⅲ单酶切，结果如图 1。

烟草 *Tsp*509 Ⅰ文库的鉴定：在 pGADT7 多克隆位点两端分别有一个 *Hind*Ⅲ识别位点，它们之间的长度为 800bp。因为我们在对载体改造时在克隆进载体的那段寡聚核苷酸中加入了一个 *Hind*Ⅲ酶切位点，这时用 *Hind*Ⅲ切割，就会产生 500bp 和 300bp 两个酶切片段。因为外源片段是在中间及下游的 *Hind*Ⅲ位点之间插入，因此如果在多克隆位点中有外源片段的插入，插入片段中无 *Hind*Ⅲ位点，*Hind*Ⅲ单酶切后将产生一条 500bp 的酶切片段和一条大于 300bp 的酶切片段。外源片段中如果存在 *Hind*Ⅲ位点，用 *Hind*Ⅲ单酶切后将会产生 3 条以上的酶切片段，除 500bp 的片段外，其余酶切片段长度之和应大于 300bp。该文库鉴定结果如图 2。

2 结果

2.1 6 种烟草近随机多肽库的构建

我们利用这一新的近随机多肽文库构建方法，共构建了六种编码序列完全不同的烟草随机多肽文库(表 1)。

表 1 六种烟草随机多肽文库

Table 1 The six tobacco random peptide libraries were constructed

Library	Digest of genomic DNA	Vector	Digest of vector	Modify of vector	Transformed clones
Tobacco(1)	*Dpn* Ⅱ	pGADT7	*Bam*H Ⅰ	No	1.02×10^7
Tobacco(2)	*Dpn* Ⅱ	pGADT7(+B)	*Bam*H Ⅰ	Yes	1.5×10^7
Tobacco(3)	*Dpn* Ⅱ	pGADT7(−B)	*Bam*H Ⅰ	Yes	1.07×10^7
Tobacco(4)	*TSP*509 Ⅰ	pGADT7	*Eco*R Ⅰ	No	5×10^7
Tobacco(5)	*TSP*509 Ⅰ	pGADT7(+E)	*Eco*R Ⅰ	Yes	1.02×10^7
Tobacco(6)	*TSP*509 Ⅰ	pGADT7(−E)	*Eco*R Ⅰ	Yes	1.1×10^7

2.2 烟草 *Dpn* Ⅱ文库的鉴定结果

由图 1 可看出，八个质粒中均有外源片段的插入，说明用 *Dpn* Ⅱ所建文库中的重组率比较高，外源插入片段的长度约在数百 bp 之间。

2.3 烟草 *Tsp*509 Ⅰ文库的鉴定结果

由图 2 可看出，挑取的八个质粒中均有外源片段的插入，3 号样品出现一条带是因为插入片段加上 300bp 后正好是 500bp，而 6、7、9 出现三条带有可能是在插入片段中存在 *Hind*Ⅲ识别位点。我们对此类文库测序产生的 24 个序列分析表明，最短的多肽为 6，最长的为 73 个，平均为 19.25 个氨基酸。

3 讨论

本文中我们介绍了一种全新的构建近随机多肽文库的方法。由于很多生物的基因组 DNA 较大，当它们被酶切成很小的片段时，每个片段的序列可以认为是接近随机的，这

样的随机片段可以用来构建随机多肽文库。该方法消除了现有建库技术中的弊端并具备以下特点:

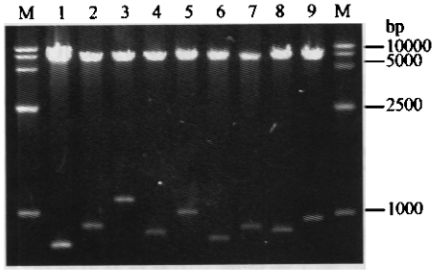

图 1 烟草 DpnⅡ文库鉴定结果

Fig.1 Library (made by *Dpn* Ⅱ) identification by *Hind*Ⅲ

M: marker; 1: pGADT7 plasmid; 2~9: recombinant pGADT7 plasmids.

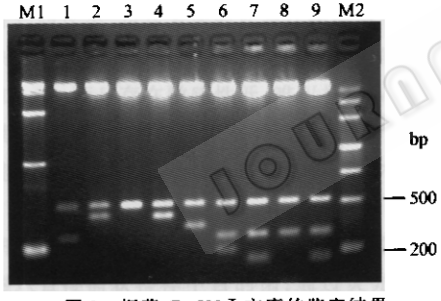

图 2 烟草 *Tsp*509Ⅰ文库的鉴定结果

Fig.2 Library (Made by *Tsp*509Ⅰ) identification by *Hind*Ⅲ

M1: marker1; M2: marker2; 1: pGADT7 plasmid;

2~9: recombinant pGADT7 plasmids.

首先,它利用自然界存在的天然产物——各物种的基因组 DNA,不需要人工合成;

第二,由于基因组 DNA 本身的复杂性保证了所建文库的复杂性;

第三,由于基因组 DNA 是双链产物,所以建库时不需要退火、补平末端等过程,而是直接酶切、连接,大大简化了试验程序;

第四,将质粒载体稍加改变还可以更充分利用基因组资源,如本文中将载体在插入位点之前修饰或改造,改变它的读码框架,则用同一酶切的基因组 DNA 连入载体后,可以产生 3 种不同序列的随机多肽文库。由于读码框架不同,在这 3 个随机肽库中产生的多肽也不一样,将它们混合以后可以产生容量更大的随机多肽文库;

第五,我们构建的随机肽库不仅有序列的随机性,还有序列长度的多样性,可以满足不同试验的需要。

REFERENCES(参考文献)

[1] Fodor SP, Read JL, Pirrung MC *et al*. Light-directed, spatially addressable parallel chemical synthesis. *Science*, 1991, **251**: 767 - 773

[2] Houghten RA, Pinilla C, Blondelle SE *et al*. Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. *Nature*, 1991, **354**: 84 - 86

[3] Lam KS, Salmon SE, Hersh EM *et al*. A new type of synthetic peptide library for identifying ligand-binding activity. *Nature*, 1991, **354**: 82 - 84

[4] Parmley SF, Smith GP. Filamentous fusion phage cloning vectors for the study of epitopes and design of vaccines. *Adv Exp Med Biol*, 1989, **251**: 215 - 218

[5] Cwirla SE, Peters EA, Barrett RW *et al*. Peptides on phage: a vast library of peptides for identifying ligands. *Proc Natl Acad Sci USA*, 1990, **87**: 6378 - 6382

[6] Scott JK, Smith GP. Searching for peptide ligands with an epitope library. *Science*, 1990, **249**: 386 - 390

[7] Huang H, Gao Y. A method for the generation of arbitrary peptide libraries using genomic DNA. *Molecular Biotechnology*, accepted

[8] Croy RRD. Plant Molecular Biology LabFax (Labfax Series). Academic Press, 1994