DOI:10.16016/j.1000-5404.2002.04.003
第24卷第4期 第三军医大学学报 Vol. 24, No. 4
2002年4月 ACTA ACADEMIAE MEDICINAE MILITARIS TERTIAE Apr. 2002

文章编号:1000-5404(2002)04-0381-04

论著

铜绿假单胞菌噬菌体 PaP3 基因组的初步注释

张克斌,金晓琳,朱军民,胡晓梅,饶贤才,胡福泉 (第三军医大学基础医学部微生物学教研室,重庆市微生物工程重点实验室,重庆 400038)

提 要:目的 对已完成测序的铜绿假单胞菌噬菌体 PaP3 基因组进行初步注释。方法 利用生物信息学软件和网上工具对噬菌体 PaP3 进行开发阅读框(Open reading frame,ORFs)检索;编码序列鉴定;对公共数据库进行 BLAST 查询,分析基因的可能功能;对存在高度同源性序列的推定蛋白进行多重序列比对和系统发生树分析。结果 利用生物信息学软件在 PaP3 基因组中检索到 252 个大于 100 bp 的 ORFs,其中有 69 个 ORFs 可确认为推定基因;核苷酸对核苷酸的 BLASTn 分析未发现有意义的同源性核苷酸序列;核苷酸对蛋白质的 BLASTx 分析共发现了 4 个在公共数据库中有高度同源性序列的 ORFs,根据其同源性蛋白的功能推定 ORF13052-14761 编码产物为引物酶/解旋酶,ORF40280-41728 编码产物为末端酶大亚单位,ORF41728-42225 编码产物为溶菌酶,ORF16912-18549 编码产物为 DNA 聚合酶;绘制出了上述这四种蛋白的系统发生树。结论 噬菌体 PaP3 具有 252 个大于 100 bp 的 ORFs,其中 69 个为推定基因,有 4 个推定基因的功能被确认为分别编码引物酶/解旋酶,末端酶大亚单位,溶菌酶及 DNA 聚合酶。

关键词:噬菌体;基因组;一级注释
中图法分类号:Q755;Q939.48 文献标识码:A

Primary annotation of gnome of *Pseudomonas aeruginosa* phage PaP3

ZHANG Ke-bin, JIN Xiao-lin, ZHU Jun-min, HU Xiao-mei, RAO Xian-cai, HU Fu-quan (Department of Microbiology, Third Military Medical University, Chongqing 400038, China)

Abstract: Objective To primarily annotate the genome of *Pseudomonas areuginosa* phage PaP3 after its sequencing. Methods The primary annotation items of the genome of the bacteriophage PaP3, including identification of open reading frames (ORFs) and coding sequence, BLAST query of each ORF, multiple alignment of homology sequences and phylogenetic analysis, were carried out with bioinformatic software and analysis tool in internet. Results A total of 252 of > 100 bp ORFs and 69 coding sequences were detected in the phage PaP3 genome. BLASTn did not find any homology sequence, but BLASTx found out 4 high homology sequences, that is, ORF13052－14761 coding for primase/helicase, ORF40280-41728 for terminase large subunit, ORF41728-42225 for lysozyme and ORF16912-18549 for DNA polymerase. Multiple alignment of those 4 homology sequences was completed and their phylogenetic trees were obtained. Conclusion The 252> 100 bp ORFs and 69 coding sequences are detected, and 4 functional genes, coding for primase/helicase, terminase large subunit, lysozyme and DNA polymerase respectively, are predicated in the phage PaP3 genome.

Key words: bacteriophage PaP3; genome; annotation

近期我们完成了铜绿假单胞菌溶原性噬菌体 PaP3 的全基因组测序,酶切图谱分析和测序结果表明该基因组是一 44 249 bp 长的线性平端 dsDNA 分子。但在基因组学研究中,测序工作的完成仅仅是解决了从字母到字母(From letter to letter)排列问题,得到的还只是一部没有空键、没有标点符号、没有词(Words)及词组(Phrase)、没有句子(Sentence)、没有段落(Paragraph)的“天书”。要读懂这部天书的意义,就是要解读构成这部天书的“词”、“词组”、“句子”、“段落”乃至“全文”的意义。这一过程就是对基因组的注释过程。一级注释即初步注释包括基因组的核酸种类(DNA 或 RNA)、单链或双链、环状或链状、末端分析,碱基组成分析,开发阅读框(Open reading frame,ORFs)分析,编码序列分析,通过 BLAST 检索(包括核苷酸水平和氨基酸水平)初步确定某些 ORFs 蛋白的可能功能,同源性蛋白种系发生分析等内容。在完成了铜绿假单胞菌噬菌体 PaP3 基因组的测序之后,本研究中介绍了对噬菌体 PaP3 基因组的初步注释。并将此作为对噬菌体 PaP3 功能基因研究的起点。

1 方法

1.1 基因组组成分析

用 DNAStar 软件包的 EditSeq 软件对 PaP3 基因组全长 44 249 bp 序列进行一般特性分析,包括 A、T、C、G 四种碱基的组成、G/C 含量分析等。并在完成基因的预测分析后推算推定基因的平均长度和基因组中基因的分布密度。

1.2 开放阅读框(Open reading frame,ORF)的预测

在美国国立卫生研究院网站上利用 ORF finder 软件进行 ORFs 分析。网址 http://www.ncbi.nlm.nih.gov/gorf/gorf.html。FASTA 格式输入全长基因组 44 249 bp 序列,Genetic codes 选项选择 11(Bacterial code)。ORF 的检测下限设定为 100,即检测结果只给出大于 100 bp 的 ORFs。

1.3 基因组中可能基因的预测

利用 ORF finder 分析发现的 ORFs 不一定都具有基因的功

基金项目:国家自然科学基金资助项目(30070037);重庆市攻关课题资助项目(2000)
作者简介:张克斌(1968－),男,四川省岳池县人,博士,主治医师,讲师,主要从事微生物基因组和生物信息学方面的研究,发表论文 10 篇,现在附属新桥医院检验科,重庆 400037。电话:(023)68755000-74654
通信作者:胡福泉,电话:(023)68752240,E-mail:hufuquan@yahoo.com
收稿日期:2002-02-05;修回日期:2002-03-11

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

能,为了预测哪些 ORFs 是可能的基因,在 Mark Borodovsky 实验室主页上使用 Genemark 软件进行序列预测分析[10],网址:http://genemark.biology.gatech.edu/Genemark/heruristic.cgi/。

1.4 各 ORF 的 BLAST 分析

采用美国国立卫生院生物技术信息中心的 BLAST 软件,网址是:http://www.ncbi.nlm.nih.gov/,对预测的 PaP3 基因组 DNA 中的 ORFs 分别进行 BLAST 分析。采用两种查询方式:一种是核酸序列比对核酸序列(Standard nucleotide-nucleotide BLAST,BLASTn);第二种是用核酸序列对蛋白数据库的比对(Nucleotide query-Protein db BLAST, BLASTx)。两种方法均以 FASTA 格式输入核酸序列。BLASTn 选项为:数据库选项"nr";滤器(Filter)选项"Low complexity";期望值(Expect)选项 10;Word size 选项"11"。BLASTx 选项为:翻译选"translated query-protein database";数据库选项"nr";滤器选项"Low complexity";Word size 选项"3";遗传密码选项"Standard"。

1.5 各蛋白质的多重序列比对和系统发生树分析

通过上述 BLAST 查询,发现了几个 ORFs 编码的蛋白质具有高度同源性序列,根据这些同源性基因的功能,这几个基因的功能基本可以认定,于是进一步对这几组同源性蛋白序列进行多重序列比对,以期分析它们可能的功能基序和各组同源性蛋白的系统发生关系。具体是利用 DNAStar 软件包中的 MegAlign 分析软件,将目的序列在 EditSeq 软件上翻译为氨基酸序列;将此序列保存为 *.pro 格式文件;在 MegAlign 软件中打开 *.pro 格式文件,并输入各组欲比较的序列文件;比对方法选择"Clustal method",这是一种可以加空位的多重序列比对(Gapped-multiple-sequence alignment)方法。比对各组序列并绘出系统发生树。

2 结果

2.1 PaP3 基因组组成特征

利用 NCBI 网站的 ORF finder 对 PaP3 全基因组进行 ORF 检索,结果检索到 252 个大于 100 bp 的 ORFs。利用 Genemark.hmm 软件进行进一步分析,在 252 个 ORFs 中确认有 69 个 ORFs 为编码序列,推定为基因。预测编码区总长度为 42 762 bp(含终止密码),基因的平均 DNA 长度为 619.7 bp。69 个推定基因共编码 14 181 个氨基酸,基因平均氨基酸长度为 205.5 个氨基酸。整个基因组中基因分布密度为平均约 600 bp 存在一个推定基因。

2.2 各 ORF 的局域相似性比对分析结果

在核酸对核酸进行比对的 BLASTn 检索分析中,未发现有意义的同源性核苷酸序列存在,这充分表明本研究测序的 PaP3 是一个新的噬菌体。在核酸对蛋白进行比对的 BLASTx 分析中,共找到 4 个有明显同源性氨基酸序列的 ORFs,它们是:

ORF13052-14761,含 1,710 个碱基,在同源性蛋白检索中,同源性分值最高的是大肠杆菌噬菌体 SP6 的 DNA 引物酶(DNA primase),分值 156,E 值达 6e-37;同源性分值次高的是一种从海洋分离的 T7 样噬菌体 Roseophage S1O1 株的引物酶/解旋酶(Primase/helicase),分值达 148,E 值达 2e-34;随后各高分值蛋白也均为不同噬菌体的引物酶/解旋酶。因此可以推定,ORF13052-14761 编码蛋白为 PaP3 的引物酶/解旋酶。

ORF41728-42225,含 1,179 个碱基。在同源性蛋白质检索中检索到高分值同源性蛋白质 4 个,其中分值最高的是新近完成测序工作的肠道沙门氏菌噬菌体 P22 的末端酶大亚单位(Terminase large subunit),分值是 275,E 值是 1e-72;随后是两个尚未鉴定功能的蛋白,即噬菌体 LP7 的 protein GP2 和 Xylella fastidiosa 的一个假设蛋白(Hypothetical protein),分值分别为 179 和 90,E 值分别为 6e-44 和 4e-17;分值最低的是噬菌体 P22 的包装基因(Packaging gene),分值是 68,E 值是 3e-10。从这一结果可以看出,ORF13052-14761 极可能是末端酶大亚单位的编码基因。

ORF41728-42225,含 498 个碱基。获较高分值的同源性蛋白均为各噬菌体的溶菌酶,分值均在 50 分左右,E 值在 e-06 左右。本研究初步认定其为溶菌酶是基于这样一些分析:一是噬菌体中都存在溶菌酶,二是检出的同源性蛋白均是各噬菌体的溶菌酶,三是在阅读框的整个长度上都有一定程度的同源性,四是本阅读框编码的蛋白长度和检出各蛋白的长度基本相等。当然,这种推定是否正确有待于进一步实验的证实。

ORF16912-18549,含 1 638 个碱基。检索到的高分值同源性蛋白几乎均为各噬菌体或细菌的 DNA 聚合酶(DNA polymerase)。其中分值最高的是噬菌体 T3 的 gene5 编码产物,分值为 167,E 值为 3e-40,但其功能尚未鉴定;次高分值的是 T7 噬菌体的 DNA 聚合酶,分值达 161, E 值达 e-38;随后的各高分值同源性蛋白几乎均为各噬菌体或细菌的 DNA 聚合酶或 DNA 聚合酶 I。因此本研究认定该 ORF 是编码 DNA 聚合酶的基因。

2.3 蛋白质的多重序列比对和系统发生树分析结果

通过上述的 BLAST 分析,初步确定了 4 个基因编码的蛋白质的功能,这里进一步将它们分别与相应高分值相似性蛋白进行多重序列比对,并绘出系统发生树,结果如下:

2.3.1 引物酶/解旋酶的多重序列比对和系统发生树分析结果

从引物酶/解旋酶组同源性蛋白的多重序列比对分析结果显示,在近 N-端的 1-350aa 区域,各蛋白之间的同源性较低,而在近 C-端的 350-650aa 区域则有较高的同源性,由于各蛋白均为不同噬菌体中的引物酶/解旋酶,因此提示近 C-端区域为引物酶/解旋酶的活性区。

从图 1 的系统发生树看,各蛋白根据系统发生亲缘关系的远近可分为 5 组,第一组是大肠杆菌噬菌体 T3 和 T7 的引物酶/解旋酶,NP-041977、NP-039041、NP-039642、NP-052087 和 NP-052088 蛋白,为一组系统发生关系非常接近的蛋白;第二组是 1CR1A 和 1EOJA 两种蛋白,与第一组系统发生关系非常近。这两组蛋白均来自于肠道杆菌噬菌体。第三组为 NP-064752 蛋白,来自于从海洋中分离的噬菌体 Reseophage S1O1;第四个为 ORF13052-14761 蛋白,即本研究的噬菌体 PaP3 的引物酶/解旋酶;第五组是 AAD41902 蛋白,来自于肠道杆菌噬菌体 SP6。从图 1 横轴上可以清楚地看出各同源性蛋白的亲缘关系的相对距离分值。

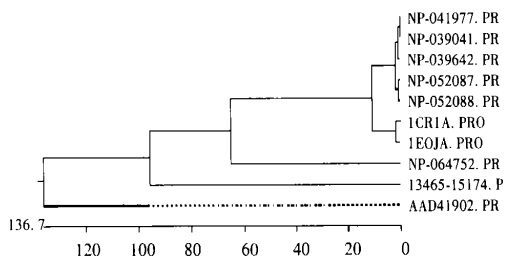


图 1 引物酶/解旋酶的系统发生树分析图

Fig 1 Phylogenetic tree for primase/helicase

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.3.2　末端酶大亚单位的多重序列比对和系统发生树分析结果　在末端酶大亚单位的多重序列比对结果中可以看到, 在全长的氨基酸序列区域内, ORF40280-41728 蛋白与其它蛋白间均有较高的相似性。它们的系统发生树关系见图 2。

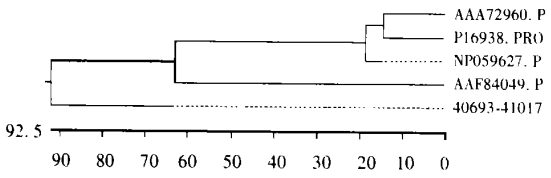


图 2　末端酶大亚基的系统发生树分析图

Fig 2　Phylogenetic tree for terminase large subunit

从系统发生树上看出, 各蛋白依系统发生关系可分为 4 组, 第一组是 AAA72960 和 P16938 蛋白, 前者是噬菌体 P22 的包装基因编码的蛋白, 后者是噬菌体 LP7 的一未知功能蛋白; 第二组是 NP-059627 蛋白, 即噬菌体 P22 的末端酶大亚单位; 第三组为 AAF84049, 是 Xylella fastidiosa 的一推定蛋白; 第四组是本研究的 ORF40280-41728 编码的蛋白。从图 2 横轴上可以清楚地看出各同源性蛋白的亲缘关系的相对距离分值。

2.3.3　溶菌酶的多重序列比对和系统发生树分析结果　在溶菌酶的多重序列比对结果中, 由于蛋白 163 L、240 L、1L22 和 1C69A 均为 T4 噬菌体溶菌酶的不同区域的突变体, 它们有极高的同源性, 而 ORF41728-42225 编码产物在全长氨基酸序列上与上述四种蛋白有一定同源性。从图 4 中可看到, ORF41728-42225 编码的溶菌酶与第一组大肠杆菌 T4 噬菌体溶菌酶的 4 种突变蛋白的亲缘关系的相对距离分值为 77。

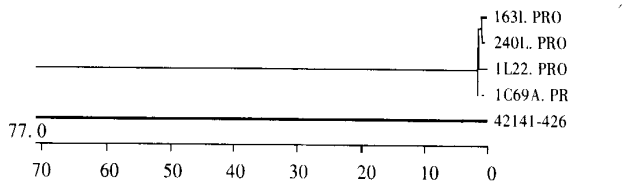
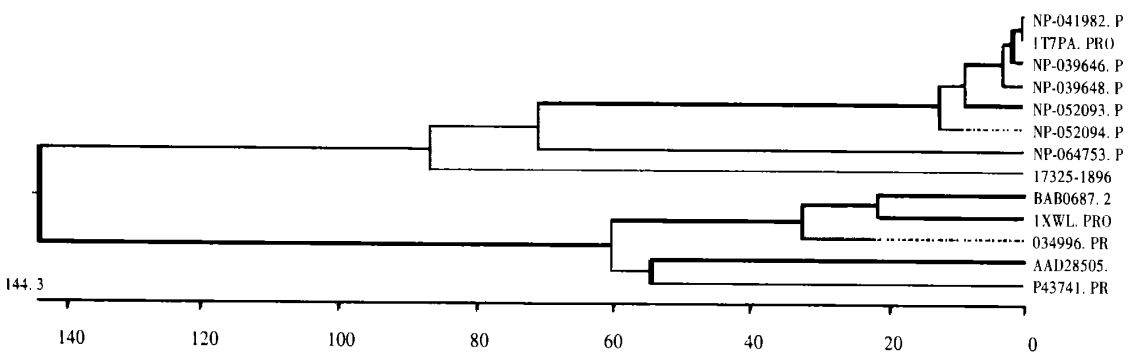


图 3　溶菌酶的系统发生树分析图

Fig 3　Phylogenetic tree for lysozyme

2.3.4　DNA 聚合酶组的多重序列比对和系统发生树分析结果　DNA 聚合酶组的多重序列比对结果显示, 在各段氨基酸序列中 ORF16912-18549 编码蛋白与其它各蛋白均有一定相似性, 但以靠近 C-端的 700aa-960aa 为高, 推测该段区域可能为 DNA 多聚酶的活性区域。

在该系统发生树中, 几乎所有蛋白均鉴定为 DNA 聚合酶或 DNA 聚合酶 I, 从系统发生关系上看, 可分为两组, 第一组是噬菌体 DNA 聚合酶, 第二组是细菌 DNA 聚合酶或 DNA 聚合酶 I。在第一组中大肠杆菌噬菌体 T3 和 T7 DNA 聚合酶关系最近, 而噬菌体 phiYeO3-12DNA 聚合酶与大肠杆菌噬菌体 T3 和 T7 的 DNA 聚合酶系统发生关系有一定距离, 而 ORF16912-18549 编码蛋白归属于噬菌体组, 说明其系统发生关系离噬菌体蛋白近, 离细菌蛋白远。



NP041982: gene5, DNA polymerase (bacteriophage T7); 1T7PR: Chain A, T7 DNA polymerase complexed to DNA primertemplate... (bacteriophage T7); NP039646: 5(bacteriophage T3); NP039648: 5B(bacteriophage T3); NP052093: DNA polymerase(bacteriophage phiYeO3-12); NP052094: 5B protein(bacteriophage phiYeO3-12); NP064753: DNA polymerase(Roseophage SI01); BAB06872: DNA polymerase I(*Bacillus halodurans*); 1XWL: DNA polymerase I(*Bacillus Stearothermophilus*); 034996: DNA polymerase(*Bacillus subtilis*); AAD28505: DNA polymerase I(*Rhodothermus marinus*); P43741: DNA polymerase(*Haemophilus influenzae*)

图 4　DNA 聚合酶的系统发生树图

Fig 4　Phylogenetic tree for DNA polymerase

3　讨论

在完成铜绿假单胞菌溶原性噬菌体 PaP3 的全基因组测序和序列组装的同时, 一些初步注释的内容, 如核酸种类(DNA 或 RNA)、单链或双链、环状或链状、末端分析, 碱基组成分析等工作已经完成, 内容已在文献[1]中介绍。在本研究中, 首先采用 ORF finder 检索 PaP3 全序列中的 ORFs, 共发现了 252 个大于 100 bp 的 ORFs. 然后是将各个 ORF 分别输入与公共数据库中进行序列局域相似性比对分析, 根据同源性基因的功能来推测未知基因的功能。在 NCBI 数据库中, 可提供多种比对检索方式, 我们根据需要, 选择了两种查询方式: BLASTn 即核苷酸对核苷酸的比对检索和 BLASTx 即核苷酸对蛋白的比对检索方式。将基因组

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

预测分析获得的 252 个ORFs 逐一对核苷酸和蛋白数据库进行检索比对。在 BLASTn 分析中，未发现任何一个 ORF 在现有数据库中存在有意义的同源性核苷酸序列，结合 BLASTx 分析结果，说明在 PaP3 与其它生物包括噬菌体之间存在遗传密码使用上的差异，即遗传密码偏嗜性的问题。新测序的 PaP3 基因组是全新的，也意味着目前我们对这个基因组的注释还只是部分注释，完全注释还有大量艰巨的工作要做。

而在 BLASTx 分析中，发现有四个 ORFs 存在有意义同源性蛋白，这四种同源性蛋白均是噬菌体生长繁殖过程所必需的酶，都具有极为重要的生物学作用。

引物酶/解旋酶的生物学作用是在 DNA 复制过程中，在复制叉的部位解开 DNA 双链，与此同时，在后随链(Lagging strand)的复制中需合成一些较短的冈崎片段(Okazaki fragments)。在大肠杆菌 T7 噬菌体中，这一过程需要两种模板依赖性酶的协同作用，首先由引物酶合成 RNA 引物，再由 DNA 多聚酶延伸而成冈崎片段[2,3]。T7 噬菌体 gene4 编码一个具有双重功能的蛋白，称为引物酶/解旋酶，组成一个环形六聚体结构，具有解旋和催化 RNA 引物合成的双重功能[3]。本研究中 ORF13052-14761 编码蛋白是否具有相同的双重功能，尚需进一步实验证实和研究。

末端酶(Terminase)在将噬菌体核酸正确装入头部衣壳的过程中具有重要作用。其全酶具有 ATP 酶(ATPase)、DNA 链分离和位点特异性的核酸酶活性，以使病毒基因组完整、正确地装入病毒颗粒中。此外，末端酶各亚单位是一系列参与核酸包装的核蛋白的组成成份，这些核蛋白能在高速运转的核酸褒装过程中完成基因组 DNA 的迅速定位，而使包装迅速有序。由核蛋白参与的这一有序过程被认为是在病毒 DNA 包装、其它生物染色体复制的启动及转录的启动过程中共同的生物学机制[4~7]。Gual A 等曾对噬菌体 SPP1 的末端酶大亚单位的功能有过较深入的研究，发现一个 6 组氨酸-大亚单位复合物在体内具有内切酶、ATP 酶和双链 DNA 结合活性，能介导一定特异性的位点剪切[5]。

噬菌体溶菌酶是噬菌体在吸附宿主菌后，破坏菌壁，注入核酸之所必需的酶，是噬菌体感染和复制后释放过程中不可缺少的酶。其作用机制是通过发挥其 N-乙酰胞壁酸-L-丙氨酸酰胺酶(N-acetylmuramyl- L-alanine amidase)活性来消化宿主肽聚糖而裂解宿主菌[8]。本研究在对噬菌体溶菌酶进行多重序列比对分析和进化树绘制时，BLASTx 检索、比对分析所获得的同源性匹配物都是 T4 噬菌体溶菌酶的不同位点突变体。基本可以确定 ORF41728-42225 是编码溶菌酶的基因，该基即是噬菌体重要的功能基因，也是看家(House-keeping gene)，在医疗实践中可能有重要的应用前景。

DNA 聚合酶的生物学功能较为复杂，运用也非常广泛，主要包括：① 模板依赖性 5′-3′DNA 聚合酶活性；②5′-3′及 3′-5′DNA 外切酶活性。该酶广泛用于 cDNA 第二链合成、DNA 测序、PCR 及 DNA3′凹端的补平和 3′端末端标记等。在大肠杆菌 T4 噬菌体中，发现结合 DNA 的 DNA 聚合酶 gp43 也能和 mRNA 的非翻译区 5′先导序列(Untranslated 5′-leader of the mRNA)结合并抑制自身 mRNA 的转录[9]。总之，DNA 聚合酶在 DNA 的合成和修复以及基因表达的调控方面都具有重要作用。

从蛋白质多重序列比对和系统发生树来看，这几种蛋白在进化程度上非常古老，均与其它噬菌体蛋白保持着较近的系统发生关系，而与细菌等其它生物的蛋白的系统发生关系较远。这说明各物种蛋白在进化上始终保持了其物种的基本特征。本研究通过 BLAST 分析所获得的这几种蛋白及其功能的预测结果是一种基于数据库检索和计算机理论分析的结果，尚需要进一步用实验来加以证实。

参考文献：

[1] 张克斌，金晓琳，朱军民，等.铜绿假单胞菌噬菌体 PaP3 基因组测序[J].第三军医大学学报，2002，24(4)：378－380.

[2] Sawaya M R，Guo S，Tabor S，*et al*. Crystal structure of the helicase domain from the replicative helicase－primase of bacteriophage T7[J]. Cell，1999，99(2)：167－177.

[3] Kato M，Frick D N，Lee J，*et al*. A complex of the bacteriophage T7 primase-helicase and DNA polymerase directs primer utilization[J]. J Biol Chem，2001，276(24)：21 809－21 820.

[4] Catalano C E. The terminase enzyme from bacteriophage lambda：a DNA-packing machine[J]. Cell Mol Life Sci，2000，57(1)：128－148.

[5] Gual A，Camacho A G，Alonso J C. Functional analysis of the terminase large subunit，G2P，of Bucillus subtilis bacteriophage SPP1[J]. J Biol Chem，2000，275(45)：35 311－35 319.

[6] Lin H，Rao V B，Black L W. Analysis of capsid portal protein and terminase functional domains：interaction sites required for DNA packaging in bacteriophage T4[J]. J Mol Biol，1999，289(2)：249－260.

[7] Schmieger H，Taleghani K M，Meierl A，*et al*. A molecular analysis of terminase cuts in headful packing of Salmonella phage P22[J]. Mol Gen Gene，1990，221(2)：199－202.

[8] DeMartini M，Halegoua S，Inouye M. Lysozymes from bacteriophages T3 and T5[J]. J Virol，1975，16(2)：459－461.

[9] Pavlov A R，Karam J D. Nucleotide-sequence-specific and non－specific interactions of T4 DNA polymerase with its own mRNA[J]. Nucleic Acids Res，2000，28(23)：4 657－4 664.

（编辑 冷怀明）

?1994-2016 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net