张　悦,施　维,李　丹,等. 禾谷镰刀菌全基因组候选效应因子预测与分析[J]. 江苏农业科学,2019,47(6):81 – 84.
doi:10.15889/j.issn.1002 – 1302.2019.06.018

禾谷镰刀菌全基因组候选效应因子预测与分析

张　悦[1], 施　维[2], 李　丹[3], 徐志雄[1], 陈子牛[1]
(1. 昆明学院生命科学与技术系,云南昆明 650214; 2. 云南省个旧市农民科技教育培训中心,云南个旧 661000;
3. 云南交通技师学院,云南昆明 650300)

摘要:本研究根据已公布的禾谷镰刀菌的全基因组信息,以其全基因组蛋白序列为试验材料,通过生物信息学方法,对其候选效应分子及其功能进行预测和分析。首先利用 SignalP、TMHMM、Protcomp、big – PI Predictor、TargetP 等程序依次预测出其分泌类型的蛋白,再通过其序列大小和半胱氨酸的含量作进一步筛选,最后利用 Blastp 工具与非冗余蛋白质数据库进行比对,找出数据库中没有蛋白同源性的序列,从而获得候选效应分子。最终对禾谷镰刀菌全基因组的 14 038 个蛋白序列进行分析,预测了 126 个符合条件的候选效应分子。本研究通过 LTR – FINDER 程序对禾谷镰刀菌全基因组内的转座子进行分析,但未发现转座子存在,值得进一步分析研究。本研究采用生物信息学分析方法预测出了禾谷镰刀菌的候选效应分子并查找其基因组内转座子情况,可为进一步研究这些效应分子的功能,了解禾谷镰刀菌进化奠定基础。

关键词:禾谷镰刀菌;全基因组;候选效应因子;转座子;生物信息学;致病机制;进化历程
中图分类号: S432.1　　**文献标志码**: A　　**文章编号**:1002 – 1302(2019)06 – 0081 – 04

禾谷镰刀菌(*Fusarium graminearum*)属半知菌类丛梗孢目瘤座孢科镰刀属菌。在粮食上普遍存在,它与其他几种镰刀菌都能引起小麦、大麦、玉米等作物发生赤霉病,还可以引起水稻、高粱、豆类、茭白等发生根腐病、茎基腐病、穗腐病。由镰刀菌引起的赤霉病不仅会造成粮食产量减产,而且产生的真菌毒素也给人畜的健康造成了严重威胁[1 – 4]。

病原菌在入侵植物的过程中会分泌效应分子到寄主植物细胞中,对寄主植物细胞的生理、生化过程及细胞代谢等产生显著的影响。通过这些致病效应因子的作用病原菌可以克服寄主植物的防卫反应,从而促进和完成对寄主植物的侵染[5]。效应分子由于要分泌到细胞外起作用,因此一般具有以下特征:(1)含有 N 端信号肽;(2)无跨膜结构域;(3)无糖基磷脂酰肌醇锚定位点;(4)没有将蛋白输送至线粒体或其他胞内细胞器的预测定位信号;(5)氨基酸残基数量大约为 50 ~ 300 个氨基酸;(6)富含半胱氨酸而且特异性高于其他病原菌的效应分子[6 – 7]。因此,可以根据效应因子的一般结构特征对已完成测序的病原微生物进行分析,预测其中可能的候选效应因子,目前已有多篇报道针对多种病原微生物的侯选效应因子进行生物信息学预测分析的报道[8 – 10]。转座子是一类能够在基因组中通过转录或逆转录,在内切酶的作用下,在其他基因座上出现的 DNA 序列。通过对转子的分析,有助于了解微生物的进化历程[11 – 12]。本研究于 2017 年 8 月对已有的禾谷镰刀菌测序数据进行统计归纳,分析其基因组中的候选效应因子序列及转座子序列,以期对了解禾谷镰刀菌的致病机制及进化历程有指导意义。

1　材料与方法

1.1　禾谷镰刀菌全基因组数据

禾谷镰刀菌全基因组数据来自数据库(http://www.broad – institute),该数据库还收录了该病菌预测的 14 038 条预测基因 DNA 序列及其预测蛋白质的氨基酸序列。

1.2　信号肽预测

SignalP 4.1 Server (http://www.cbs.dtu.dk/services/SignalP/)是预测信号肽的服务器。它的功能是预测给定的氨基酸序列中是否存在潜在的信号肽剪切位点及其所在置,原核生物和真核生物都可以进行预测。以 SignalP 4.1 分析给定的氨基酸序列 C、S、Y 的最大值,以及位于 N 端和被预测

收稿日期:2017 – 11 – 01
基金项目:云南省高校特色生物资源开发与利用重点实验室开放基金项目(编号:GXKJ201606);云南省昆明市滇池水生态恢复重点实验室开放项目(编号:SST201607)。
作者简介:张　悦(1981—),男,云南昆明人,博士,副教授,主要从事植物病理学研究。E – mail:www151519190@163.com。

remedies[J]. Land Use Policy,2016,57(3):694 – 701.
[12]Mesas – Carrascosay F J,Notario – Garcia M D,de Larriva J E M,et al. Validation of measurements of land plot area using UAV imagery [J]. International Journal of Applied Earth Observation and Geoinformation,2014,33(5):270 – 279.
[13]Mao W H,Wang Y M,Wang Y Q. Real time detection of between – row weeds using machine vision[C]. Las Vegas,Nv July,2003.
[14]龙满生,何东健. 玉米苗期杂草的计算机识别技术研究[J]. 农业工程学报,2007,23(7):139 – 144.
[15]杨　柳,陈延辉,岳德鹏,等. 无人机遥感影像的城市绿地信息提取[J]. 测绘科学,2017,42(2):59 – 64.
[16]江　洪,汪小钦,吴　波,等. 地形调节植被指数构建及在植被覆盖度遥感监测中的应用[J]. 福州大学学报(自然科学版),2010,38(4):527 – 532.

的剪切位点间 S 曲线的中间值,以此区分信号肽和非信号肽。而信号肽剪切位点则位于预测的含有信号肽蛋白的 Y 曲线的最大值处,本试验中使用默认设置[13]。

1.3　蛋白的跨膜结构域预测

TMHMM Server (http://www. cbs. dtu. dk/services/TMHMM/)主要用于预测蛋白的跨膜结构域[14]。

1.4　亚细胞定位预测

ProtComp (http://www. softberry. com/berry. phtml?topic = protcompan&group = programs& subgroup = proloc)主要是对动物或真菌中蛋白的亚细胞定位进行预测。它可将蛋白按以下归属进行划分:细胞核、质膜、胞外分泌、细胞质、线粒体、内质网、过氧化物酶体、溶酶体、高尔基体等[15]。

1.5　是否有脂质锚定修饰预测

本研究通过 Big - PI Predictor(http://mendel. imp. ac. affgpi/fungi - server. html) 对在真菌糖基磷脂酰肌醇(glycosylphosphatidylinosi - tol,简称 GPI)修饰位点进行预测,判断预测蛋白是否有脂质锚定修饰。如果存在糖基磷脂酰肌醇脂质锚定修饰,真核生物中蛋白质须要在内质网中[16]。

1.6　确定预测蛋白的亚细胞定位

用 TargetP 1.1 Server(http://www. cbs. dtu. dk/services/TargetP/)进一步确定预测蛋白的亚细胞定位。可将蛋白的定位归属于线粒体、叶绿体、胞外分泌以及其他亚细胞定位。位置分配是基于相应的 N 末端的前序列预测存在,如叶绿体转运肽、线粒体靶向肽、分泌途径的信号肽[17]。

1.7　计算半胱氨酸含量

通过 CalMolWtCalMolWt (http://www. cnhupo. cn/CalMW/MYMW. asp)对蛋白质分子量、氨基酸组成进行计算,用于分析筛选出序列的半胱氨酸含量[9]。

1.8　序列比对

Blastp (http://blas. ncbi. nlm. nih. gov/Blasucgi?PROGRAM = blastp&PAGETYPE = BlastSearch&LINKLOC = blasthome)是使用蛋白序列在其蛋白数据库中进行查询的一种工具。每条所查序列能与数据库中已存在的每条已知序列进行序列比对,可以通过所得结果结合参数得到与此相关的信息。

1.9　转座子的筛选分析

本研究是通过 LTR - FINDER 程序(http://tlife. fudan. edu. cn/ltr_finder/)对禾谷镰刀菌全基因组序列进行转座子的筛选分析[18]。

2　结果与分析

由表 1 可知,数据库公布的禾谷镰刀菌全基因组共有 36.45 Mbp,分为 433 个重叠群,预测基因数量共计 14 038 个。

表 1　禾谷镰刀菌全基因组信息

项目	大小或数量
基因组装配大小(Mbp)	36.45
基因组序列的重叠群数量(个)	433
预测基因数量(个)	14 038

通过 SignalP 4.1 Server 对 14 038 个蛋白序列进行分析,预测得到 1 271 个编码含 N 端信号肽的蛋白,占全基因组蛋白序列的 9.05%。对所得的结果进行分析发现,序列长度主要集中在 100 ~ 600 aa 之间,占全部的 83.08%,其中长度在 200 ~ 300 aa 之间的序列最多,占全部的 19.27%。序列长度在 900 ~ 1 000 aa 的最少,仅仅只有总数的 0.78%(图 1)。

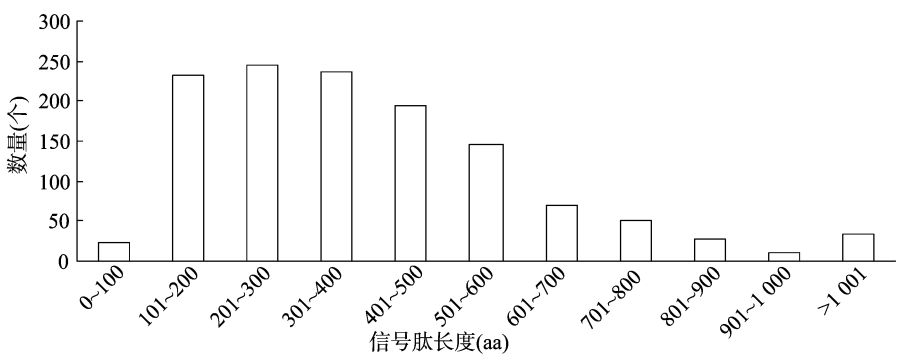

图1　禾谷镰刀菌含信号肽蛋白的序列长度分布

在含有信号肽的分泌型蛋白质序列中,如果含有跨膜区则表明该蛋白可能为膜受体,也可能是膜上的锚定蛋白或离子通道蛋白。本研究使用 TMHMM Server v2.0 来预测蛋白序列的跨膜螺旋结构,排除具有跨膜结构域的蛋白序列。从蛋白跨膜结构域分析结果可以看到,在含信号肽的 1 271 个蛋白序列中,有 92 个蛋白序列含有 2 个或多个跨膜域,157 个蛋白序列只含有 1 个跨膜域,而有 1 022 个蛋白序列则不含跨膜域。只含 1 个跨膜域的蛋白序列,其所具有的跨膜结构域位置均位于 N 端,该区域可能为前期所预测的信号肽序列。由于服务器并不能完全对信号肽序列和所属跨膜域序列进行区分,因此,本研究选择不含跨膜域和只含有 1 个跨膜域的 1 179 个蛋白序列进行下一步研究。

将上述初步筛选出的 1 179 个蛋白序列进一步用 ProtComp v9.0 进行分析,如表 2 所示,共预测到 733 个信号肽分泌至胞外,4 个转运至液泡膜,4 个转运至液泡,5 个转运至溶酶体,16 个转运至溶酶体膜,7 个转运至细胞核,10 个转运至内质网,31 个转运至内质网膜,11 个转运至高尔基体,18 个转运至高尔基体膜,11 个传输至过氧化物酶体,29 个转运至线粒体,123 个转运至线粒体膜,49 个转运至细胞质,128 个转运至细胞质膜。继而进行 GPI 锚定蛋白的预测,以判断这些被初步推断为分泌蛋白的是否为胞外蛋白。将 733 个分泌蛋白用 Big - PI Predictor 程序进行分析,发现有 63 个为 GPI 锚定蛋白,而 670 个为非 GPI 锚定蛋白。

由于效应分子的氨基酸残基数量一般在 50 ~ 300 aa 之间,因此将这 733 个序列按照其长度进行排列,明确了其中有 349 个蛋白的氨基酸残基数量在50 ~ 300 aa之间。此外,根

表 2　具有信号肽的蛋白质亚细胞定位预测

预测定位部位	蛋白数量（个）	占比（%）
液泡膜	4	0.34
液泡	4	0.34
溶酶体	5	0.42
细胞核	7	0.59
内质网	10	0.85
高尔基体	11	0.93
过氧化物酶体	11	0.93
溶酶体膜	16	1.36
高尔基体膜	18	1.53
线粒体	29	2.46
内质网膜	31	2.63
细胞质	49	4.16
线粒体膜	123	10.43
细胞质膜	128	10.86
细胞外	347	29.43
细胞外膜	386	32.74

据效应分子富含半胱氨酸的特点，利用 CalMolWt 计算所有候选序列的半胱氨酸含量。将不含半胱氨酸的序列排除之后，得到 336 个含有半胱氨酸残基数量在 1 ~ 24 个之间的蛋白序列。最后，将上述符合条件的氨基酸序列在 NCBI 数据库中利用 Blastp 工具与非冗余蛋白质数据库进行比对，找出那些与数据库中没有同源性的序列，要求其 E - value 值小于 1×10^{-5}，最终得到 126 个符合上述所有 6 个条件的候选效应分子。其中有 16 个蛋白序列在数据库中没有任何与之同源的序列，剩余的 110 个蛋白序列除了与禾谷镰孢属的序列有同源性外，与其他物种都没有同源性。

现已明确大多数物种的信号肽主要是通过 4 种类型的信号肽酶识别位点被信号肽酶所识别并被切割，从而使成熟蛋白穿过膜转运到细胞不同的部位。本研究通过对 126 个候选效应分子所含的信号肽氨基酸长度进行分析，如图 2 所示，含有信号肽长度为 16 ~ 20 aa 的蛋白质序列数量最多，所占比例为 79.36%，其中尤以所含信号肽长度为 18 aa 的蛋白序列居多，所占比例为 24.60%。

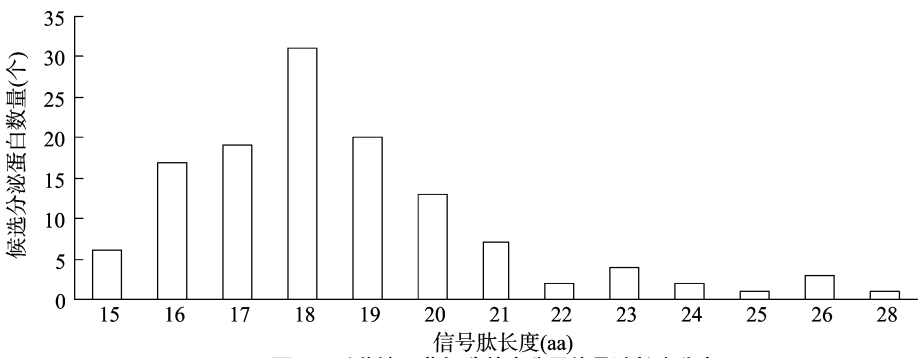

图2　禾谷镰刀菌候选效应分子信号肽长度分布

利用 LipoP 1.0 Server 对上述分泌蛋白进行信号肽酶识别位点的预测分析，结果显示 114 个蛋白序列含有 SpI 型信号肽识别位点，7 个含有 CYT 型信号肽识别位点，5 个含有 SpII 型信号肽识别位点，所占比例分别为 90.48%、5.56%、3.97%，说明禾谷镰刀菌中的候选效应分子大部分是由 SpI 型信号肽酶进行识别。

此外还对 20 种氨基酸在 126 个候选效应分子信号肽中出现的频率进行了分析。如图 3 所示，在组成信号肽的氨基酸中，丙氨酸的数量最多，为 3 422 个，占 10.15%；其次为甘氨酸，有 3 275 个，占 9.73%；其后依次为丝氨酸、苏氨酸、亮氨酸、脯氨酸、缬氨酸、异亮氨酸、天冬酰胺、天冬氨酸、谷氨酸、精酰胺、赖氨酸、苯丙氨酸、酪氨酸、谷氨酰胺、蛋氨酸、组氨酸、半胱氨酸、色氨酸，分别占 8.41%、8.15%、6.02%、5.46%、5.29%、4.98%、4.98%、4.81%、4.38%、4.19%、4.16%、4.07%、3.11%、2.92%、2.67%、2.52%、2.49%、1.43%。统计分析发现，非极性、疏水氨基酸（丙氨酸、缬氨酸、亮氨酸、甘氨酸、异亮氨酸、脯氨酸）的出现频率最高，占 41.63%；其次为极性、不带电荷的氨基酸（丝氨酸、苏氨酸、半胱氨酸、蛋氨酸、天冬酰胺、谷氨酰胺），占 29.62%；带正电荷的碱性氨基酸（赖氨酸、精酰胺、组氨酸）占 10.87%；带负电荷的酸性氨基酸（天冬氨酸、谷氨酸）占 9.19%；芳香族氨基酸（色氨酸、苯丙氨酸、酪氨酸）占 8.61%。

另外，对禾谷镰刀菌的 433 个重叠群的基因组序列利用 LTR - FINDER 程序进行了转座子序列筛选，结果并没有发现转座子的存在。

3　结论与讨论

禾谷镰刀菌全基因组测序的完成和公布，为研究禾谷镰刀菌的分泌蛋白、效应分子、致病因子及与植物之间互作提供了重要的数据基础。

本研究通过对禾谷镰刀菌的基因组序列的分析与筛选，对 14 038 个蛋白序列进行分析，预测得到 1 271 个编码含 N 端信号肽的蛋白。其中不含跨膜域和只含有 1 个跨膜域的有 1 179 个蛋白，进一步分析预测，其中 733 个蛋白分泌至胞外，这些蛋白有可能是与禾谷镰刀菌致病相关的候选效应分子。继而进行 GPI 锚定蛋白的预测，发现有 63 个为 GPI 锚定蛋白，而 670 个为非 GPI 锚定蛋白。真菌细胞壁上的 GPI 锚定蛋白对真菌的黏附、形态转换和细胞壁合成有着重要的影响。微生物的黏附是其致病性最重要的决定因素之一[19-20]，真菌病原体被确定的黏附素很少，因此预测 670 个非 GPI 锚定蛋白为候选效应因子。

利用 CalMolWt 计算所有候选序列的半胱氨酸含量，将不含半胱氨酸的序列排除之后，得到 336 个含有半胱氨酸残基数量 在1 ~24个之间的蛋白序列。最后，将上述符合条件的

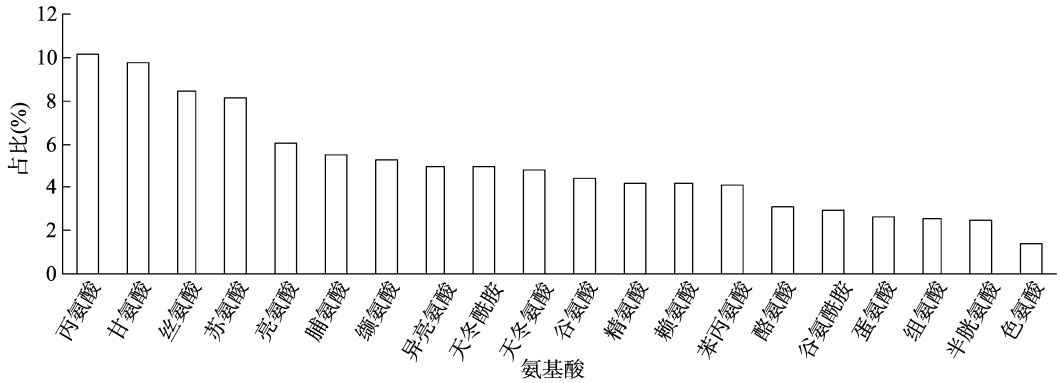

图3　禾谷镰刀菌候选效应分子信号肽氨基酸分布频率

氨基酸序列在 NCBI 数据库中利用 Blastp 工具与非冗余蛋白质数据库进行比对,找出那些与数据库中没有同源性的序列,最终得到 126 个符合条件的候选效应分子。其中有 16 个蛋白序列在数据库中没有任何与之同源的序列,剩余的 110 个除了与禾谷镰孢属的序列有同源性外,与其他物种都没有同源性。将分泌蛋白进行信号肽酶识别位点的预测分析,结果显示 114 个蛋白序列含有 SpI 型信号肽识别位点,7 个含有 CYT 型信号肽识别位点,5 个含有 SpII 型信号肽识别位点。说明禾谷镰刀菌中的效应分子大部分是由 SpI 型信号肽酶进行识别的。

此外,还对 20 种氨基酸在 126 个候选效应分子信号肽中出现的频率进行了分析,得出丙氨酸的数量最多,含量最低的是色氨酸。最后测得 126 个符合要求的禾谷镰刀菌候选效应分子大多属于小型蛋白,其信号肽集中在 16 ~ 20 个氨基酸且含有大部分的 SpI 型信号肽识别位点。这些效应分子可能是禾谷镰刀菌的致病因子。

本研究还对禾谷镰刀菌的 433 个公布的重叠群序列利用 LTR - FINDER 程序进行转座子查找分析,结果并没有发现转座子的存在,该结果值得进一步研究分析。

通过一系列的筛选与分析,可以更好地从分子水平系统地了解禾谷镰刀菌基因与蛋白质的结构与组成。并对进一步研究禾谷镰刀菌与寄主植物之间的关系有一个更好的基础,为今后研究其致病性以及其病源危害奠定基础。

参考文献:

[1]张大军,邱德文,蒋伶活. 禾谷镰刀菌基因组学研究进展[J]. 安徽农业科学,2009,37(17):7892 - 7894.

[2]王路遥,王　超,申成美,等. 引发小麦赤霉病和茎基腐病禾谷镰孢菌的生物防治初探[J]. 麦类作物学报,2014,34(5):703 - 708.

[3]纪武鹏,于　琳,王　平. 玉米茎腐病主要致病菌——禾谷镰孢菌的生物学特性研究[J]. 现代化农业,2014(9):67 - 69.

[4]张志博,高增贵,张小飞,等. 分离自小麦赤霉病和玉米茎基腐病的禾谷镰孢菌的致病性研究[J]. 辽宁农业科学,2010(6):1 - 4.

[5]于钦亮,马　莉,刘　林,等. 禾谷镰刀菌基因组中含寄主靶向模体分泌蛋白功能的初步分析[J]. 生物技术通报,2008(1):160 - 165,180.

[6]刘玉岭,柳云帆,谢建平. 粟酒裂殖酵母全基因组中含信号肽蛋白质的研究[J]. 遗传,2007,29(2):250 - 256.

[7]吕伟强,刘　聪,黄丽丽,等. 内生菌 KM - 1 - 2 全基因组 ORFs 信号肽和分泌蛋白预测及功能分析[J]. 微生物学报,2017,57(3):411 - 421.

[8]闫丽斌,肖淑芹,薛春生. 玉米大斑病菌全基因组候选效应分子的预测和分析[J]. 沈阳农业大学学报,2017,48(1):15 - 20.

[9]陈琦光,王陈骄子,杨　媚,等. 希金斯刺盘孢全基因组候选效应分子的预测[J]. 热带作物学报,2015,36(6):1105 - 1111.

[10]陈琦光,舒灿伟,杨　媚,等. 植物病原真菌效应分子的研究进展[J]. 基因组学与应用生物学,2016,35(11):3105 - 3114.

[11]马　欣,高学文. 转座子随机突变芽孢杆菌的研究进展[J]. 中国生物防治学报,2015,31(3):394 - 403.

[12]何虎翼,谭冠宁,唐洲萍,等. 植物转座子与基因表达调控[J]. 生物技术通报,2017,33(4):38 - 43.

[13] Petersen T N, Brunak S, von Heijne G, et al. SignalP 4. 0: discriminating signal peptides from transmembrane regions [J]. Nature Methods,2011,8(10):785 - 786.

[14] Krogh A, Larsson B, Heijne G V, et al. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes[J]. Journal of Molecular Biology, 2001, 305(3): 567 - 580.

[15] Zhang M Q. Computational prediction of eukaryotic protein - coding genes[J]. Nature Reviews Genetics, 2002, 3(9): 698 - 709.

[16] Eisenhaber B, Bork P, Eisenhaber F. Sequence properties of GPI - anchored proteins near the omega - site: constraints for the polypeptide binding site of the putative, transamidase[J]. Protein Engineering Design & Selection, 1998, 11(12): 1155 - 1161.

[17] Emanuelsson O, Brunak S, von Heijne G, et al. Locating proteins in the cell using TargetP, SignalP and related tools [J]. Nature Protocols, 2007, 2(4): 953 - 971.

[18] Xu L, Zhang Y E, Su Y, et al. Structure and evolution of full - length LTR retrotransposons in rice genome[J]. Plant Systematics and Evolution, 2010, 287(1/2): 19 - 28.

[19]刘　昭,于小番,张　宇,等. 丝氨酸蛋白酶编码基因 *prtP* 对副干酪乳杆菌黏附特性的影响[J]. 江苏农业学报,2017,33(3):683 - 689.

[20]李鹏成,杨　倩,侯继波. 乳酸杆菌 S - 层蛋白对产肠毒素大肠杆菌黏附 Caco - 2 细胞的协同作用[J]. 江苏农业学报,2017,33(2):384 - 388.