微生物学报 *Acta Microbiologica Sinica*
48(4)：473~479; 4 April 2008
ISSN 0001-6209; CN11-1995/Q
http://journals.im.ac.cn

里氏木霉(*Trichoderma reesei*)分泌组的预测及分析

唐雯, 严明*

(南京工业大学制药与生命科学学院, 南京 210009)

摘要:【目的】里氏木霉是一种重要的产纤维素酶工业用菌种, 研究其分泌组特性具有现实意义。【方法】应用生物信息学方法对里氏木霉基因组中 9997 个开放阅读框(ORF) 所编码的氨基酸序列进行了分析, 获得了 294 条可能的分泌蛋白序列, 并且按功能对其进行了分类, 同时用搜索模体的方法在未知功能的序列中找到具有关键模体的序列, 初步确定其潜在的功能。对获得的分泌蛋白的信号肽序列进行了分析。【结果】里氏木霉分泌组中有 188 种水解酶, 包括 114 种糖苷水解酶、42 种蛋白水解酶和 11 种脂类水解酶等; 在糖苷水解酶中包括已报道的 22 种纤维素酶和 15 种几丁质酶等, 以及 30 条具有潜在纤维素酶功能的蛋白序列。信号肽序列分析结果表明其同源性较低, 而在信号肽酶切位点附近则相对保守。【结论】通过该预测和分析开拓了里氏木霉的研究空间, 为今后的研究奠定了理论基础。

关键词：里氏木霉; 分泌蛋白; 分泌组; 预测; 信号肽; 模体
中图分类号：Q933　文献标识码：A　文章编号：0001-6209 (2008) 04-0473-07

里氏木霉(*Trichoderma reesei*)是用于生产纤维素酶最常用的菌种[1], 具有容易培养控制、产物容易分离纯化、代谢物安全无毒等特性, 近年来倍受关注。里氏木霉可以通过分泌产生大量与纤维素降解相关的酶, 因此研究其分泌组特性具有重大的意义。

传统研究分泌蛋白的方法主要是通过蛋白质双向凝胶电泳进行分析, 往往只能得到一部分分泌蛋白的信息。用生物信息学[2,3]的方法通过研究微生物的基因组、蛋白组而得到微生物的系统特性是洞悉微生物的有效途径, 即从庞大的基因组数据中通过计算获得一系列相关信息, 并经过分析和实验验证最终得到所需的结果。目前, 研究人员已经通过计算机对酿酒酵母[4]、结核分枝杆菌[5]、秀丽小杆线虫[6]、稻瘟菌[7]、根癌土壤杆菌[8]等多种微生物进行分泌蛋白的预测, 得到其分泌组的相关信息。

里氏木霉基因组的测序工作已经完成, 总大小约为33MB[9], 它的注释正在进行。本文通过对里氏木霉已公布的 9997 条 ORF 进行信号肽和跨膜区的预测, 筛选到其分泌蛋白组, 然后通过功能的预测[10]进行了初步分类, 并对其信号肽序列进行了分析, 对更全面地掌握里氏木霉的分泌特性有重大意义, 并为里氏木霉糖苷水解酶的进一步研究奠定了基础。

1 材料和方法

1.1 工具

1.1.1 真核生物基因组数据库 JGI(DOE Joint genome institute): Eukaryotic genomics(http: //genome.jgipsf.org)提供里氏木霉已注释和未注释的完整基因组序列和已注释的转录组和蛋白组信息。从中获得了的 9997 条蛋 白序列。

1.1.2 信号肽预测：SignalP 3.0 Server (http: //www.cbs.dtu.dk/services/SignalP)利用人工神经网络和HMM (Hidden Markov Models)的原理对氨基酸序列的信号肽存在情况和信号肽切割位点进行预测[11,12], 其准确度

*通讯作者。Tel: +86-25-83587695; E-mail: yanming@njut.edu.cn
作者简介：唐雯(1983–), 女, 浙江丽水人, 硕士研究生, 主要从事分子生物学的研究。E-mail: jiongjiong821@163.com
收稿日期：2007-08-27; 修回日期：2007-12-28

为 90%~91%[13]。将里氏木霉所有蛋白质序列分为 30 次提交。

1.1.3 跨膜区预测：Proteome Analyst(http: //www.cs.ualberta.ca/~bioinfo/PA)是基于 SwissProt 数据库，以 BLAST 为基本策略的预测蛋白亚细胞结构的软件[14,15]，是准确性比较高的预测工具。将从信号肽预测所得的结果提交到 Proteome Analyst 进行跨膜区的预测。

1.1.4 功能预测：将获得的分泌蛋白通过 Proteome Analyst 与 SwissProt 数据库中的已知蛋白进行相似性比对，预测出分泌蛋白的大致功能。使用 MYHITS[16] (http: //myhits.isb-sib.ch/cgi-bin/motif_scan)和 MEME[17] (http: //meme.sdsc.edu/meme/meme.html) 搜索未知序列的模体(motif)。

1.1.5 信号肽分析：用 Pratt(http: //www.ebi.ac.uk/ pratt/)寻找信号肽序列中共有的模式(pattern)；用 LipoP 1.0 Server(http: //www.cbs.dtu.dk/services/ LipoP/) 将预测出来的分泌蛋白分为被 I 型信号肽酶识别的蛋白和被 II 型信号肽酶识别的蛋白；用 TatP 1.0 sever(http: //www.cbs.dtu.dk/services/TatP/) 寻找带有 RR-motif 的信号肽。

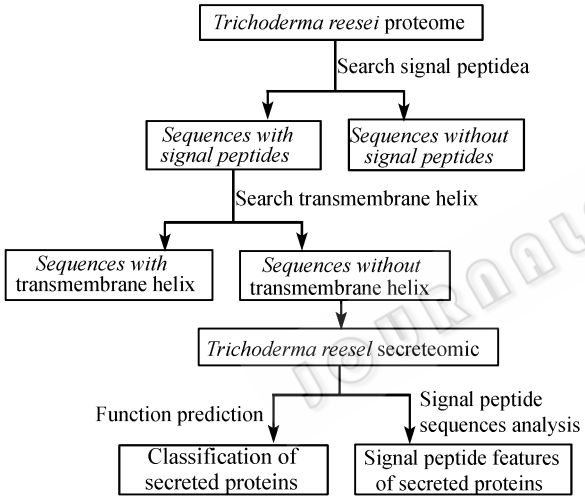

图 1 分泌蛋白预测和分析基本过程

Fig. 1 The scheme of the prediction and analysis of secreted proteins.

1.2 分析流程

使用 SignalP 对来自 JGI 的 9997 条蛋白序列进行信号肽预测，找出可能含有信号肽的序列，然后通过 Proteome Analyst 对这些序列进行分析，排除含有跨膜区的蛋白序列，找到一组分泌蛋白序列，并通过对这些序列的功能预测进行分泌蛋白的初步分类，最后进行分泌型信号肽序列的分析(图 1)。

2 结果

2.1 分泌蛋白功能预测和分类

通过预测分析，得到 294 条可能的分泌蛋白序列。由此看出，从里氏木霉蛋白组层面进行分析预测，发现它含有大量可能的分泌蛋白。使用 Proteome Analyst 通过 BLAST 对全部分泌蛋白进行功能预测，可对其进行如下的初步分类(图 2)。

通过 BLAST 分析，在预测得到的 294 条分泌蛋白的序列中，属于水解酶的蛋白序列为 188 条，主要包括蛋白水解酶、糖苷水解酶和与胆固醇代谢相关的水解酶，还有 41 条水解酶序列不知其具体功能。188 条水解酶序列中，属于蛋白水解酶的蛋白序列有 33 条，主要包括羧基肽酶(3.40%)、氨基肽酶(1.70%)、天冬氨酰蛋白酶(4.76%)等；属于糖苷水解酶的蛋白序列有 113 条，主要包括溶菌酶(1.02%)、纤维素酶(7.48%)、几丁质酶(5.10%)、木聚糖酶(0.34%)等，属于脂类代谢酶的蛋白序列有 1 条。

通过对 294 条蛋白序列的模体分析，在未知功能的糖蛋白、信号蛋白及细胞壁相关蛋白中，发现 8 条序列存在与功能相关的模体；在具体功能未知的水解酶中，发现 24 条序列存在相关的活性位点。这些分泌蛋白序列中具有潜在蛋白水解酶功能的序列有 9 条(精氨酸酶、丝氨酸酶、金属肽酶)，海藻糖酶序列 1 条，酯酶和脂肪酶序列 10 条，此外还可能存在紧密连接蛋白、免疫球蛋白、DNA 甲基化酶等。表 1 列举了其中 21 条较重要序列。

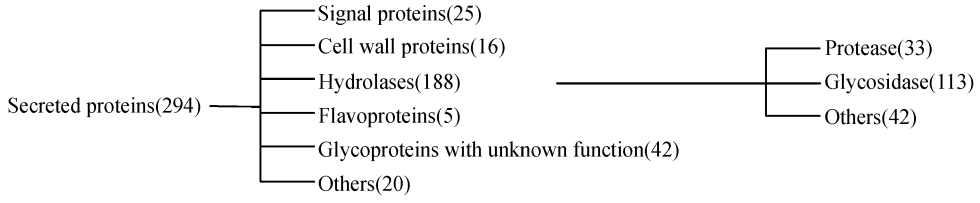

图 2 预测分泌蛋白分类

Fig. 2 Classification of predicted secreted proteins.

表 1 未知功能的分泌蛋白中具有关键模体的序列

Table 1 Sequences with key motifs in secreted proteins with unknown function

ID* of protein	Description of the motif	Possible function
40671	N-6 Adenine-specific DNA methylases signature	DNA methylases
33336	Arginase family signature 1 and 2	Arginase
41897	Serine proteases, trypsin family, histidine and serine active site	Serine protease
36557、45744、46483、42701、42985、35630	Serine proteases, subtilase family, aspartic acid and serine active site	Serine protease
40778	Lipolytic enzymes "G-D-X-G" family, histidine and serine active site	Lipolytic enzyme
41104	Neutral zinc metallopeptidases, zinc-binding region signature	Neutral zinc metallopeptidase
29811、33760、45576、41000、28334、30076、33674、35711	Carboxylesterases type-B serine active site	Carboxylesterase
45141	Lipases, serine active site	Lipase
30863	Trehalase signature 2	Trehalase

*Protein ID from JGI Eukaryotic genomics.

综合以上分析，我们得出，在 188 条水解酶序列中，蛋白水解酶共 42 条，糖苷酶共 114 条，脂类水解酶共 11 条，分别占所有分泌蛋白的 14.29%、38.78%和 3.74%。

2.2 分泌蛋白中的糖苷水解酶

从里氏木霉分泌蛋白的功能预测结果来看，糖苷水解酶所占的比例最大，这与里氏木霉降解木质纤维素的生理功能相对应。降解木质纤维素的酶有很多种，其中纤维素酶和木聚糖酶是比较重要的两种酶。在预测出来的分泌组中，存在 23 条已经确定的纤维素酶、半纤维素酶和木聚糖酶[18](表 2)。

表 2 预测的分泌组中已经确定的纤维素酶、半纤维素酶和木聚糖酶

Table 2 Identified cellulases, hemicellulases and xylanases in the predicted secreteomic

Enzymes	EC3.2.1	Amount of amino acids	GenBank ID	Reference
EGI	4	459	M15665	19
EGII	4	418	M19373	20
Cel12A	4	234	AB003694	21
EGIV	4	344	Y11113	22
EGV	4	242	Z33381	23
Cel74a	4	838	AAP57752	24
Cel61b	4	249	AAP57753	24
Cel5b	4	438	AAP57754	24
CBHI	91	513	P62694	25
CBHII	91	471	M16190	26
Xyn1	8	222	P36217	27
Xyn2	8	229	S39883	28
Xyn3	8	347	BAA89465	29
Man1	25	437	L25310	30
Bgl1	21	744	U09580	31
Bgl2	21	466	AB003110	32
Bgl4	21	833	AY281375	24
Cel1b	21	484	AY281377	24
Cel3b	21	874	AY281374	24
Cel3e	21	765	AY281379	24
Cel3d	21	700	AY281378	24
Xyl3A	37	797	Z69257	32
Xyl2	8	229	CAA49294	27

此外，在这 114 条糖苷水解酶的序列中存在 72 条与糖代谢相关而未知具体水解底物的序列，通过搜索模体的方法发现其中有 3 条序列具有多聚半乳糖醛酸酶活性位点，2 条序列具有几丁质识别结合区域,2 条序列具有半乳糖苷酶标记，1 条序列具有葡萄糖糖化酶识别标记(表 3)。使用比对的方法进一步分析这 72 条序列的模体，发现其中 6 条序列具有相同的模体，可能是具有相同水解底物的糖苷水解酶；5 条序列具有 CBM[18](carbohydrate-binding module) 区，可能是与木质纤维素降解相关的酶或蛋白；还有 25 条序列存在 1~3 个与已知纤维素酶相同的模体，可能是与纤维素降解相关的酶或蛋白。

表 3 未知功能的糖苷水解酶中具有关键模体的序列

Table 3 Sequences with key motifs in glycosidases with unknown function

ID* of protein	Description of the motif	Possible function
31101、40480、16889	Polygalacturonase active site	Polygalacturonase
20876、13945	Chitin recognition or binding domain signature	Chitinase
37303、27852	Alpha-galactosidase signature	Galactosidase
28313	Glucoamylase active site region signature	Glucoamylase

*Protein ID from JGI Eukaryotic genomics.

通过上述分析，我们得出，在预测出来的里氏木霉分泌组中，除了已经确定的 23 条与木质纤维素降解相关的蛋白序列之外，还有 30 条蛋白序列具有潜在的纤维素酶功能，这些序列需要通过比对和实验的方法进一步分析验证。

2.3 分泌蛋白的 ORF

通过 SignalP 的预测，得到 1450 条可能含有信号肽的蛋白序列，占全部序列的 14.5%，其中 74.5%可能性在 0.8 以上，预测的准确性较高。将这 1450 条序列经过跨膜区的预测，最终得到的 294 条分泌蛋白，占全部序列的 2.94%。编码这些分泌蛋白的序列平均长度为 1200bp 左右，其中大部份的序列分布在 600bp 至 1800bp 之间(图 3)。从图中可以看出，ORF 长度的分布呈偏正态分布。除此之外，编码分泌蛋白的 ORF 长度大部份在 3000bp 以下，适合用基因工程的方法对其进行进一步的实验研究。

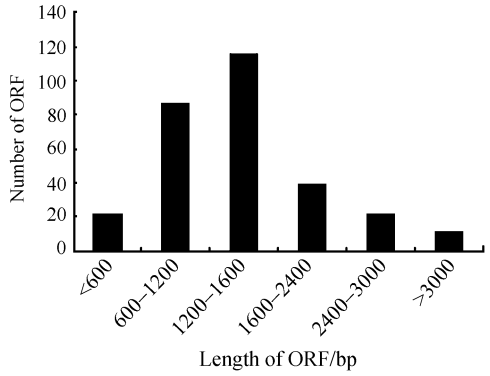

图 3 编码分泌蛋白的 **ORF** 序列长度分布

Fig. 3 Distribution of ORFs coding secreted proteins with different length.

2.4 分泌蛋白中的信号肽分析

2.4.1 分泌型信号肽的特征：引导分泌蛋白的信号肽长度(氨基酸残基的个数) 最大为 63 个氨基酸，最小为 10 个氨基酸，平均值是 20 个氨基酸。其中，17~20 个氨基酸是主要的分布区间(图 4)，占总数的 55.4%，有 19 个氨基酸的信号肽数量最多，为 60 个，占总数的 20.4%。信号肽长度的变化，说明信号肽具有高度的变异性。

分析 294 条分泌蛋白信号肽的氨基酸组成(图 5)，发现 20 种氨基酸的出现频率从高到低依次为：A→L→S→V→M→T→G→I→P→R→F→K→Q→H→N→C→Y→W→E→D。非极性氨基酸 A、L、V、M、G、I、P 的出现频率相对较高，其中丙氨酸最高，为 22.6%；而有带电侧链的氨基酸 E 和 D 的出现频率最低，分别为 0.56%和 0.35%。从中可以看出，在分泌蛋白信号肽组成中，出现频率在 5%以上的氨基酸大多为脂肪族氨基酸，并以中性和含羟基的氨基酸居多。这可能与分泌蛋白的属性相关，使信号肽更易穿过质膜，从而行使信号指导定位功能。

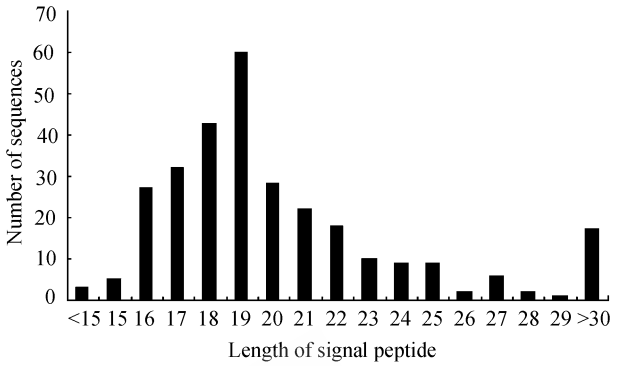

图 4 携带不同长度信号肽的蛋白序列分布

Fig. 4 Distribution of protein sequences with different length of signal peptide.

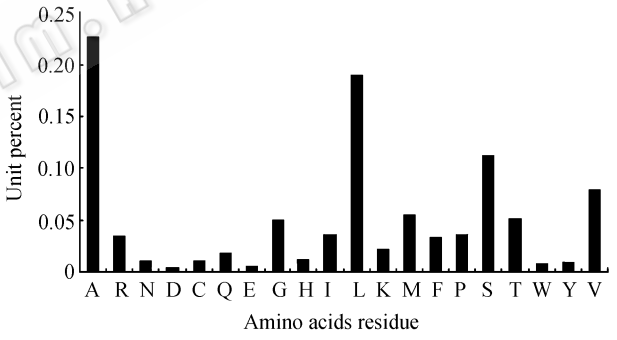

图 5 **20** 种氨基酸残基在分泌蛋白信号肽序列中的出现频率

Fig. 5 The frequency of 20 amino acids residues of signal peptides in secreted proteins.

2.4.2 信号肽序列的相似性：将 294 条蛋白序列的信号肽序列递交到 Pratt，得到 50 种模式(pattern)，其中 6 种模式比较典型(表 4)。

表 4 分泌蛋白信号肽中的典型模式

Table 4 Typical patterns of signal peptides in secreted proteins

No.	Pattern	Fitness*	Ratio in all secreted proteins/%
Pattern 1	S-x(0,3)-L-x(0,2)-L-x(0,3)-G-x(0,2)-A-x(0,4)-A	21.0203	8.50
Pattern 2	L-[AGILPV]-L-x(0,2)-G-x(0,1)-A-x(0,2)-A	20.9995	7.14
Pattern 3	S-L-x(0,2)-L-x(0,5)-G-x(0,2)-A-x(0,5)-A	20.5203	6.46
Pattern 4	L-x(0,2)-L-x(0,2)-G-x(0,2)-A-x(0,2)-A	19.3503	11.56
Pattern 5	L-x(0,1)-P-x(0,2)-L-x(0,3)-A-x(0,2)-A	17.8503	6.12
Pattern 6	L-x(0,1)-G-x(0,1)-A-x-A	15.6802	7.82

*Fitness is the score which evaluates the fitting degree of a pattern to all sequences that have the same pattern.

从表中可看出，存在 Pattern4 的蛋白序列最多，但也仅占全部分泌蛋白的 11.56%，表明信号肽的同源性比较低。将存在同一种模式的蛋白序列进行蛋白功能的分析，并没有发现明显的功能一致性。为了进一步研究信号肽序列与蛋白功能的关系，将与纤维素降解有关的蛋白序列的信号肽进行了分析，发现仅有 2~4 条蛋白序列中含有同一种模式。从以上结果可以得出，在里氏木霉中，分泌蛋白的信号肽相似程度比较低，且信号肽序列与蛋白的功能并没有直接的关系。

2.4.3 信号肽酶识别位点：根据信号肽酶识别信号肽序列的不同，可将其分为 I 型信号肽酶和 II 型信号肽酶[33]，通常 I 型信号肽酶对存在典型信号肽序列的蛋白前体进行切割，而 II 型信号肽酶对存在脂框(Lipobox)的脂蛋白前体进行切割。在预测出来的 294 条分泌蛋白的信号肽序列中，278 条序列带有 I 型信号肽酶识别位点，1 条序列带有 II 型信号肽酶识别位点，15 条序列同时含有 I 型和 II 型信号肽酶识别位点。由此表明大部分的蛋白在分泌到胞外时，是由 I 型信号肽酶将信号肽切割掉的。

2.4.4 RR-motif 型信号肽：含有 RR-motif 的信号肽，与 Tat(Twin arginine translocation) 途径的底物蛋白有关[34]，Tat 途径的底物蛋白通常是还原-氧化辅因子结合蛋白，通过该途径得到对应的辅因子，同时对蛋白进行折叠。RR-motif 通常被定义为 RRxFLK[35](x 代表任意氨基酸)，但也有其他形式的 RR-motif 存在。通过预测，在 294 条分泌蛋白的信号肽中，仅找到两条序列含有 RR-motif。结合功能分类，发现这两种蛋白均与细胞分裂有关。

2.4.5 C 区特性：信号肽有 3 个区域[36]，即 N 区、H 区和 C 区，其中 C 区包含信号肽酶切位点。通过分析 20 种氨基酸在 C 结构域的−3 位置(相对于信号肽酶切位点，“−”代表左边，“+”代表右边；“−3”即左边第 3 个位置，以下类同)、−2 位置、−1 位置、+1 位置、+2 位置和+3 位置的出现频率(表 5)，发现在−3、−1 和+1 位置上 A 的出现频率最高，分别为 46.9%、81.4%和 19.9%。在−3 位置上，除了 A 之外，出现的其它氨基酸残基为 V、T、S、G、I、C、L 等；在−1 位置上，则为 G、S、T 等。−3 和−1 位置是信号肽酶的关键识别位点，因此−3 和−1 位置上的氨基酸相对保守。

表 5 20 种氨基酸残基在信号肽酶切位点附近的出现频率(%)

Table 5 The frequency of 20 amino acid residues around the signal peptidase digesting sites(%)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
−3	46.9	0	0	0	2.1	0	0	2.8	0	2.8	1.4	0	0	0	0	5.9	7.9	0	0.3	30.0
−2	15.7	2.8	3.5	2.1	0.7	6.6	3.5	1.7	4.5	0.7	17.1	0	2.1	2.1	0	21.6	6.3	1.7	1.4	5.9
−1	81.4	0	0	0	0.7	0	0	10.0	0	0	0.3	0	0	0	0.7	5.2	1.7	0	0	0
+1	19.9	3.1	2.7	3.8	0.7	12.3	3.1	4.5	3.8	2.7	7.6	2.4	0.7	2.1	0	10.0	9.6	1.7	2.7	6.5
+2	4.1	3.4	4.8	4.5	2.1	3.8	2.7	4.5	2.1	2.1	3.1	1.7	0	3.4	26.4	10.6	7.9	1.4	3.1	8.6
+3	9.0	2.4	4.1	1.7	1.7	1.4	1.7	5.5	2.1	7.6	12.4	1.4	0.7	34	11.7	9.7	12.1	0.3	1.4	9.7

3 讨论

本文应用生物信息学的相关工具对里氏木霉已公布的 9997 条蛋白序列进行信号肽和跨膜区的分析，预测出 294 条分泌蛋白，通过蛋白功能的预测进行分类，得到以下结论：(1) 里氏木霉分泌蛋白中 64%是水解酶，包括糖苷酶(38.78%)、蛋白酶(14.29%)和脂类水解酶(3.74%)等。(2) 里氏木霉的糖苷水解酶占很大比例，包括很多与纤维素、几丁质等多糖降解相关的酶，除了已经确定的 23 条与木质纤维素降解相关蛋白序列之外，还有 30 条蛋白序列具有潜在的纤维素酶功能。

此外，通过对预测出来的里氏木霉分泌蛋白的信号肽进行分析，得到其氨基酸组成、相似性、信号肽酶类型、RR-motif、C 区等方面的特征。里氏木霉分泌型信号肽中疏水氨基酸残基含量最多；大部分是被 I 型信号肽酶识别并切割；且 RR-motif 很少，Tat 途径可能不存在或作用很小；C 区中的信号肽酶识别位点的氨基酸组成相对保守。

虽然人们已经发现里氏木霉能分泌多种纤维素酶，但就已报道的里氏木霉 23 条与木质纤维素降解相关的蛋白序列来说，只占预测出的分泌蛋白序列的 7.8%，这表明里氏木霉不仅可作为产纤维素酶的菌种，对研究和生产其他水解酶来说也有很高的利用价值。因此，通过对里氏木霉分泌蛋白的预测，掌握其分泌蛋白的相关信息，开拓了里氏木霉的研究空间，为今后的研究奠定了理论基础。而为了从分泌组层面更全面地掌握里氏木霉的特性，有必要对里氏木霉的相关序列进行进一步的比对，期望找出其具有相关功能序列的规律性，以利于更深层地利用里氏木霉。

参 考 文 献

[1] 吴石金, 罗锡平, 夏一峰. 里氏木霉产纤维素酶系各组分分泌特性. 浙江林学院学报(*Journal of Zhejiang Forestry College*), 2003, 20(2): 146−150.

[2] 陈铭. 后基因组时代的生物信息学. 生物信息学(*China Journal of Bioinformatics*), 2004, 2(2): 29−34.

[3] Ivanov AS, Veselovsky AV, Dubanov AV, *et al*. Bioinformatics platform development: from gene to lead compound. *Methods Mol Biol*, 2006, 316: 389−431.

[4] Yang J, Li Ch, Wang Y, *et al*. Computational Analysis of signal peptide-dependent secreted proteins in *Saccharomyces cerevisiae*. *Agricultural Sciences in China*, 2006, 5(3): 221−227.

[5] 王亮, 胡建平. 结核分枝杆菌 (H_{37} Rv)分泌性蛋白的生物信息学预测方法. 第四军医大学学报(*Journal of the Fourth Military Medical University*), 2006, 27(1): 86−89.

[6] 吴红芝, 李成云, 朱有勇, 等. 秀丽小杆线虫分泌蛋白组的计算机分析. 遗传(*Hereditas*), 2006, 28(4): 470−478.

[7] 苏源, 李成云, 赵之伟, 等. 稻瘟菌基因组规模分泌蛋白的预测分析. 云南农业大学学报(*Journal of Yunnan Agricultural Unversity*), 2006, 21(3): 271−275.

[8] 范成明, 李成云, 赵明富, 等. 根癌土壤杆菌 C58 Cereon 中分泌蛋白信号肽分析. 微生物学报(*Acta Microbiologic Sinics*), 2005, 45(4): 561−566.

[9] Diener SE, Chellappan MK, Mitchell TK, *et al*. Insight into *Trichoderma reesei*'s genome content, organization and evolution revealed through BAC library characterization. *Fungal Genet Biol*, 2004, 41(12): 1077−1087.

[10] 张丽娟, 成军, 罗军. 新基因功能预测的理论及方法. 医学分子生物学杂志(*Journal of Medical Molecular Biology*), 2006, 3(4): 279−282.

[11] Nielsen H, Brunak S, Heijne von G. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Engineering*, 1999, 12(1): 3−9.

[12] Menne MLK, Hermjakob H, Apweiler R. A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics*, 2000, 16(8): 741−742.

[13] Klee WE, Ellis BML. Evaluating eukaryotic secreted protein prediction. *BMC Bioinformatics*, 2005, 6(256): 1−7.

[14] Gardy LJ, Brinkman SLF. Methods for predicting bacterial protein subcellular localization. *Nat Rev Microbiol*, 2006, 4(10): 741−751.

[15] Duane Szafron, Paul Lu, Russell Greiner, *et al*. Proteome Analyst: custom predictions with explanations in a web-based tool for high-throughput proteome annotations. *Nucleic Acids Res*, 2004, 32(Web Server issue): 365−371.

[16] Falquet L, Pagni M, Bucher P, *et al*. The PROSITE database, its status in 2002. *Nucleic Acids Res*, 2002, 30: 235−238.

[17] Bailey LT, Gribskov M. Methods and statistics for combining motif match scores. *Journal of Computational Biology*, 1998, 5: 211−221.

[18] Jia Ouyang, Ming Yan, Dechong Kong, *et al*. A complete protein pattern of cellulase and hemicellulase genes in the filamentous fungus *Trichoderma reesei*. *Biotechnology Journal*, 2006, 1(3): 1266−1274.

[19] Penttila M, Lehtovaara P, Nevalainen H, *et al*. Homology between cellulase genes of *Trichoderma reesei*: complete nucleotide sequence of the endoglucanase Ⅰ gene. *Gene*, 1986, 45(3): 253−263.

[20] Saloheimo M, Lehtovaara P, Penttila M, *et al*. EGIII, a new endoglucanase from *Trichoderma reesei*: the characterization of both gene and enzyme. *Gene*, 1988, 63(1): 11−22.

[21] Okada H, Tada K, Sekiya T, *et al*. Molecular characterization and heterogonous expression of the gene encoding a low-molecular-mass endoglucanase from *Trichoderma reesei* QM9414. *Appl Environ Microbiol*, 1998, 64: 555−563.

[22] Saloheimo M, Nakari-Setala T, Tenkanen M, *et al*. cDNA cloning of a *Trichoderma reesei* cellulase and demonstration of endoglucanase activity by expression in yeast. *Eur J Biochem*, 1997, 249: 584−591.

[23] Saloheimo A, Henrissat B, Hoffren A M, *et al*. Novel, small endoglucanase gene, egl5, from *Trichoderma reesei* isolated by expression in yeast. *Mol Microbiol*, 1994, 13: 219−228.

[24] Foreman P K, Brown D, Dankmeyer L, *et al*. Transcriptional regulation of biomass-degrading enzymes in the filamentous fungus *Trichoderma reesei*. *J Biol Chem*, 2003, 278(34): 31988−31997.

[25] Shoemaker S, Schweickart V, Ladner M, *et al*. Molecular cloning of exo-cellobiohydrolase I derived from *Trichoderma reesei* strain L27. *Biotechnology*, 1983, 1: 691−696.

[26] Teeri TT, Lehtovaara P, Kauppinen S, *et al*. Homologous domains in *Trichoderma reesei* cellulolytic enzymes: gene sequence and expression of cellobiohydrolase Ⅱ. *Gene*, 1987, 51(1): 43−52.

[27] Torronen A, Mach RL, Messne R, *et al*. The two major xylanases from *Trichoderma reesei*: characterization of both enzymes and genes. *Biotechnology*, 1992, 10(11): 1461−1465.

[28] Saarelainen R, Paloheimo M, Fagerstrom R, *et al*. Cloning, sequencing and enhanced expression of the *Trichoderma reesei* endoxylanase Ⅱ (p Ⅰ 9)gene xln2. *Mol Gen Genet*, 1993, 241(5-6): 497−503.

[29] 陆长梅, 袁生, 赵庆新. 用 Overlap-PCR 法从 *Trichoderma reesei* QM9414 基因组 DNA 中克隆并表达木聚糖酶Ⅲ. 生物工程学报(*Chinese Journal of Biotechnology*), 2004, 20(5): 764−769.

[30] Stalbrand H, Saloheimo A, Vehmaanpera J, *et al*. Cloning and expression in *Saccharomyces cerevisiae* of a *Trichoderma reesei*

beta-mannanase gene containing a cellulose-binding domain. *Appl Environ Microbiol*, 1995, 61: 1090–1097.

[31] Takashima S, Nakamura A, Hidaka M, *et al*. Molecular cloning and expression of the novel fungal beta-glucosidase genes from *Humicola grisea* and *Trichoderma reesei*. *J Biochem*, 1999, 125(4): 728–736.

[32] Margolles-Clark E, Tenkanen M, Nakari-Setala T, *et al*. Cloning of genes encoding alpha-L-arabinofuranosidase and beta-xylosidase from *Trichoderma reesei* by expression in *Saccharomyces cerevisiae*. *Appl Environ Microbiol*, 1996, 62(10): 3840–3846.

[33] Harold T, Albert B, Jan DH, *et al*. Signal peptide-dependent protein transport in *Bacillus subtilis*: a genome-based survey of the Secretome. *Microbiology and Molecular Biology Reviews*, 2000, 9: 515–547.

[34] Bendtsen JD, Nielsen H, Widdick D. Prediction of twin-arginine signal peptides. *BMC Bioinformatics*, 2005, 6: 167.

[35] Berks BC. A common export pathway for proteins binding complex redox cofactors. *Mol Microbiol*, 1996, 22(3): 393–404.

[36] Paetzel M, Dalbey R E, Strynadka NC. Crystal structure of a bacterial signal peptidase in complex with a beta-lactam inhibitor. *Nature*, 1998, 396: 186–190.

Prediction and Analysis of the Secreteomic in *Trichoderma reesei*

Wen Tang, Ming Yan*

(*College of Life Science and Pharmaceutical Engineering*, *Nanjing University of Technology*, *Nanjing* 210009, *China*)

Abstract: [Objective] We studied the secretomics' properties of *Trichoderma reesei*, an important industrial microorganism used for cellulase production. **[Methods]** We analyzed the amino acid sequences coded by 9997 ORFs in *Trichoderma reesei* genome with bioinformatics approaches, identified 294 possible secreted protein sequences, and classified them by functions. We also applied motif search methods to search key motifs in the function-unknown sequences and preliminary predicted their functions. Moreover, we analyzed the signal peptide sequences of the secreted proteins. **[Results]** There were 188 hydrolytic enzymes in *Trichoderma reesei*'s secretomics, including 114 glycosidases, 42 proteases, and 11 lipoidases. The glycosidases included 22 reported cellulases, and 15 chitinases, as well as 30 other protein sequences probably related to cellulose degradation. The homology of signal peptides of secreted proteins was low whereas sequences near the digesting site of signalase were conservative. **[Conclusion]** This method gave insights into the whole secreted proteome of *Trichoderma reesei* and provided basis for further studies on secretomic features at a genome level.

Keywords: *Trichoderma reesei*; secreted protein; secreteomic; prediction; signal peptide; motif

*Corresponding author. Tel: +86-25-83587695; E-mail: yanming@njut.edu.cn m
Received: 27 August 2007/ Revised: 28 December 2007