

基于全基因组测序的禾谷炭疽菌中 碳水化合物酶类蛋白预测

韩长志

(西南林业大学 林学院;云南省森林灾害预警与控制重点实验室,昆明 650224)

摘 要:禾谷炭疽菌侵染玉米、小麦等农作物引起的炭疽病,给各国农业生产造成了巨大经济损失。基于前期研究结果,以630个分泌蛋白为基础序列,利用CAT预测程序,对上述蛋白进行碳水化合物酶类蛋白(CAZymes)的找寻,明确该菌中CAZymes含有267个,分为主要类别和复合类别两大类,前者包括89个GHs、53个CBMs以及41个AAs、28个CEs、11个PLs、2个GTs;后者则包括30个GHs/CBMs、10个AAs/CBMs、1个CEs/CBMs等。该研究为深入开展该病菌侵染植物的作用机制提供重要的理论基础。

关键词:全基因组测序;禾谷炭疽菌;碳水化合物酶类蛋白

中图分类号:S435.1

文献标志码:A

禾谷炭疽菌(*Colletotrichumgraminicola*(Cesati) Wilson)作为半活体营养型病菌,可以侵染玉米、小麦等禾本科植物引起炭疽病,给各国农业生产造成巨大损失^[1-5]。碳水化合物酶类(Carbohydrate-Active Enzymes, CAZymes)是病原菌侵染过程中突破寄主细胞第一道屏障——植物细胞壁的关键因素^[6]。通过对不同营养类型的植物病原真菌进行分析发现,非活体营养型病原菌基因组比活体营养型病原菌具有更多的CAZymes^[6-7]。因此,随着大量植物病原菌基因组测序的完成,通过大规模预测分析病原菌的基因组中碳水化合物酶类基因将为此类病原菌的防治提供更好的策略^[8]。

碳水化合物活性酶类数据库是专业描述结构相关的酶催化或者是碳水化合物绑定(或者是功能域)模块的数据库,这些基因功能涉及水解、修饰、转移、断裂糖苷键^[9]。碳水化合物结构(功能)活性酶类数据库(Carbohydrate-Active enZymes Database)^[9]使病原菌胞外酶细胞壁降解酶活性基因的注释与聚类分析得到极大的加强^[10]。近些年关于CAZymes的研究呈现较快的增长^[11]。根据最新研究成果,以涵盖木质纤维素降解所需要的相关酶类进行分类,主要涉及以下6大类,糖苷水解酶(Glycoside Hydrolases, GHs)^[12]、糖基转移酶(Glycosyl Transferases, GTs)^[13]、多糖裂解酶(Polysaccharide Lyases, PLs)^[14]、碳水化合物酯酶(Carbohydrate Esterases, CEs)^[14]、辅助酶类家族(Auxiliary Activities, AAs)^[15]以及碳水化合物绑定结构(Carbohydrate-Binding Modules, CBMs)^[16]。以2016年6月8日数据,CAZy数据库中共有361个家族,涉及上述类别酶的家族分别有135、99、24、16、13以及74个。本研究基于前期所获得的禾谷炭疽菌中630个分泌蛋白^[17],利用CAZymes Analysis Toolkit(CAT, <http://cricket.ornl.gov/cgi-bin/cat.cgi>)^[10]预测工具明确禾谷炭疽菌中CAZymes不同亚家族的分布情况,以期为深入开展该菌CAZymes的功能研究打下坚实的基础。

收稿日期:2016-04-13; **修回日期:**2016-06-10。

基金项目:国家自然科学基金(31560211);云南省森林灾害预警与控制重点实验室开放基金项目(ZK150004);云南省优势特色重点学科生物学一级学科建设项目(50097505);云南省高校林下生物资源保护及利用科技创新团队(2014015)。

作者简介(通信作者):韩长志(1981-),男,河北石家庄人,西南林业大学讲师,博士,研究方向为经济林木病害生物防治与真菌分子生物学, E-mail: hanchangzhi2010@163.com。

1 材料与方法

1.1 禾谷炭疽菌分泌蛋白序列来源

前期利用 Signal P, ProtComp, TMHMM, big-PI Fungal Predictor 和 TargetP 等在线预测工具对 *C. graminicola* 全基因组序列进行分泌蛋白找寻,获得的 630 个分泌蛋白^[17].

1.2 CAZymes 预测

利用 CAT 在线工具对上述分泌蛋白中的 CAZymes 开展预测工作,具体而言,首先利用上述网站中基于 BLAST 比对方法以及基于 pfam 保守结构域方法对前期所获得的 630 个分泌蛋白进行分析,并对获得的 CAZymes 类别进行后续分类统计.

2 结果与分析

2.1 禾谷炭疽菌中 CAZymes 的类型

禾谷炭疽菌中总共含有 267 个 CAZymes,其中,尤以 GHs 含量较高,所占比例为 33.33%;其次为 CBMs, AAs, CEs, PLs, GTs,所占比例分别为 19.85%、15.36%、10.49%、4.12%、0.75%(图 1).此外,在上述 CAZymes 的预测中,还有部分蛋白同属于复合酶的类型,涉及 GH/CBM, AA/CBM, CE/CBM 等(图 1).由于 CAZymes 分类系统的理论基础是氨基酸序列的相似性反映蛋白质保守的结构折叠类型,同一蛋白具有不同酶类型,推测上述蛋白在发挥功能方面具有特殊性.

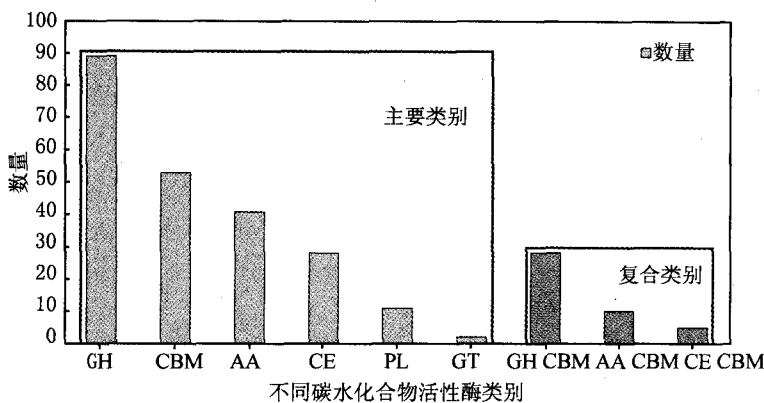


图1 禾谷炭疽菌中不同碳水化合物活性酶分布情况

2.2 禾谷炭疽菌中 CAZymes 的分布

2.2.1 禾谷炭疽菌中 GH 的分布

通过对上述糖基水解酶进行分析,发现 GH10, GH28, GH43 数量较多,为 6 个,其次为 GH3,为 5 个(表 1).值得关注的是, GH2, GH23, GH26, GH27, GH32 等类别的酶均只有 1 个蛋白,上述现象可以在一定程度上解释禾谷炭疽菌寄主广泛的原因.

2.2.2 禾谷炭疽菌中 CBM 的分布

通过对上述碳水化合物结合结构蛋白进行分析,发现 CBM50 数量较多,为 19 个,其次为 CBM13,数量均为 10 个,上述蛋白所占比例为 78.05%(表 2).此外,还涉及 CBM1, CBM2, CBM3, CBM5, CBM12, CBM32 以及 CBM57, CBM63 等类别蛋白.值得关注的是, GLRG-11041 属于 CBM12/CBM32/CBM63 类别,作为唯一一个在 CAZymes 酶类主要类别为 CBM、不同类型的蛋白,推测其在发挥 CBM 功能方面具有特殊性.

表1 禾谷炭疽菌中糖基水解酶分布情况

不同类别	ID	数量
GH2	GLRG-06433	1
GH3	GLRG-04872, GLRG-05395, GLRG-07734, GLRG-07742, GLRG-07947	5
GH5	GLRG-02165, GLRG-06263, GLRG-07690	3
GH7	GLRG-01760, GLRG-08974, GLRG-10496	3
GH10	GLRG-02054, GLRG-06931, GLRG-08031, GLRG-08139, GLRG-08165, GLRG-08914	6
GH11	GLRG-01738, GLRG-01976, GLRG-05348, GLRG-05524, GLRG-09962	5
GH12	GLRG-02687, GLRG-09921, GLRG-11951	3
GH13	GLRG-07266, GLRG-09452	2
GH16	GLRG-00338, GLRG-02215, GLRG-03555, GLRG-11730	4
GH17	GLRG-05648, GLRG-09366	2
GH18	GLRG-01873, GLRG-07132, GLRG-07749, GLRG-11050	4
GH20	GLRG-01100, GLRG-06280	2
GH23	GLRG-08514	1
GH26	GLRG-09226	1
GH27	GLRG-11396	1
GH28	GLRG-00800, GLRG-01092, GLRG-06645, GLRG-09911, GLRG-10528, GLRG-11389	6
GH30	GLRG-07683, GLRG-11082	2
GH31	GLRG-03534, GLRG-05958	2
GH32	GLRG-10490	1
GH35	GLRG-02255, GLRG-06813	2
GH43	GLRG-01733, GLRG-06934, GLRG-07768, GLRG-09259, GLRG-11211, GLRG-11363	6
GH45	GLRG-00636, GLRG-02944	2
GH53	GLRG-04044	1
GH55	GLRG-04185, GLRG-07606, GLRG-07939	3
GH62	GLRG-07682	1
GH67	GLRG-04392	1
GH75	GLRG-06284, GLRG-07202	2
GH76	GLRG-01135, GLRG-05002, GLRG-06063, GLRG-09468	4
GH78	GLRG-07254	1
GH79	GLRG-00365, GLRG-04782	2
GH92	GLRG-08346, GLRG-08815	2
GH93	GLRG-01144, GLRG-05999	2
GH105	GLRG-06036, GLRG-10310	2
GH128	GLRG-08167	1
GH131	GLRG-00140, GLRG-08878	2
GH132	GLRG-04405	1
合计	—	89

表2 禾谷炭疽菌中碳水化合物绑定结构分布情况

不同类别	ID	数量
CBM1	GLRG-03559, GLRG-07469, GLRG-11267	3
CBM2	GLRG-05319, GLRG-07984, GLRG-09466	3
CBM3	GLRG-05463	1
CBM5	GLRG-04240, GLRG-09702	2
CBM12	GLRG-00881, GLRG-01786, GLRG-06648, GLRG-07172, GLRG-07698, GLRG-11936	6
CBM13	GLRG-00289, GLRG-00821, GLRG-04809, GLRG-05320, GLRG-05523, GLRG-06040, GLRG-07001, GLRG-10303, GLRG-10476, GLRG-11933	10
CBM32	GLRG-06151, GLRG-06643, GLRG-09307, GLRG-10305	4
CBM50	GLRG-00313, GLRG-00789, GLRG-00923, GLRG-01347, GLRG-02116, GLRG-02289, GLRG-02947, GLRG-03701, GLRG-04797, GLRG-06219, GLRG-06670, GLRG-06967, GLRG-07575, GLRG-07767, GLRG-09161, GLRG-09932, GLRG-11022, GLRG-11140, GLRG-11587	19
CBM57	GLRG-06995, GLRG-11861	2
CBM63	GLRG-01658, GLRG-11975	2
CBM12/CBM32/CBM63	GLRG-11041	1
合计	—	53

2.2.3 禾谷炭疽菌中 AA 的分布

通过对上述辅助酶家族进行分析,发现 AA9 数量较多,为 16 个,其次为 AA1,AA7,数量均为 8 个,上述蛋白所占比例为 78.05%(表 3)。此外,还涉及 AA2,AA3 以及 AA5,AA10 等类别蛋白。

表 3 禾谷炭疽菌中辅助酶类家族分布情况

不同类别	ID	数量
AA1	GLRG-03376, GLRG-04158, GLRG-04766, GLRG-06593, GLRG-08623, GLRG-08638, GLRG-08949, GLRG-11502	8
AA2	GLRG-00293, GLRG-09626	2
AA3	GLRG-06195, GLRG-08186, GLRG-09759, GLRG-10297	4
AA5	GLRG-03175, GLRG-11288	2
AA7	GLRG-01068, GLRG-05347, GLRG-06312, GLRG-06381, GLRG-06552, GLRG-09236, GLRG-10142, GLRG-10946, GLRG-00823, GLRG-01827, GLRG-04195, GLRG-05009, GLRG-06274, GLRG-08089, GLRG-08605, GLRG-09070, GLRG-09165, GLRG-09214, GLRG-09215, GLRG-10314, GLRG-10700, GLRG-11110, GLRG-11180, GLRG-11551	8
AA9	GLRG-09214, GLRG-09215, GLRG-10314, GLRG-10700, GLRG-11110, GLRG-11180, GLRG-11551	16
AA10	GLRG-06483	1
合计	—	41

2.2.4 禾谷炭疽菌中 CE 的分布

通过对上述碳水化合物酯酶蛋白进行分析,发现 CE5 数量较多,为 8 个,其次为 CE10,数量为 7 个,上述蛋白所占比例为 53.57%(表 4)。此外,还涉及 CE1,CE4,CE12,CE16 等类别蛋白。

表 4 禾谷炭疽菌中碳水化合物酯酶的分布情况

不同类别	ID	数量
CE1	GLRG-01823, GLRG-08032, GLRG-08913, GLRG-10654	4
CE4	GLRG-00386, GLRG-04776, GLRG-05587, GLRG-07915	4
CE5	GLRG-02120, GLRG-06811, GLRG-08340, GLRG-09296, GLRG-09578, GLRG-09929, GLRG-10468, GLRG-11172	8
CE8	GLRG-02404, GLRG-10770	2
CE10	GLRG-00351, GLRG-01925, GLRG-02614, GLRG-05333, GLRG-06562, GLRG-08221, GLRG-09350	7
CE12	GLRG-07704, GLRG-09386	2
CE16	GLRG-01795	1
合计	—	28

2.2.5 禾谷炭疽菌中 PL 的分布

通过对上述多糖裂解酶蛋白进行分析,发现 PL1 数量较多,为 5 个,其次为 PL3,数量为 4 个,上述蛋白所占比例为 81.82%(表 5)。此外,还涉及 PL4,PL9 等类别蛋白,其数量较少,均为 1 个。

2.2.6 禾谷炭疽菌中 GT 的分布

通过对上述糖基转移酶蛋白进行分析,发现仅包含 GT2,GT4 两个蛋白,其 ID 分别为 GLRG-06582, GLRG-01825(表 6)。与其他酶类进行比较,该酶类较少,推测其在禾谷炭疽菌危害植物过程中发挥着特有功能。

表5 禾谷炭疽菌中多糖裂解酶的分布情况

不同类别	ID	数量
PL1	GLRG-00973、GLRG-06347、GLRG-06555、GLRG-09012、GLRG-11475	5
PL3	GLRG-00244、GLRG-00836、GLRG-01732、GLRG-11725	4
PL4	GLRG-09637	1
PL9	GLRG-09775	1
合计	—	11

表6 禾谷炭疽菌中糖基转移酶的分布情况

不同类别	ID	数量
GT2	GLRG-06582	1
GT4	GLRG-01825	1
合计	—	2

2.2.7 禾谷炭疽菌中复合类别蛋白的分布

通过对上述 CAZymes 进行分析,发现除含有典型的 GH,CBM,AA,CE,GT 以外,还含有一些复合类别的蛋白,涉及 GH/CBM,AA/CBM,CE/CBM 三大类,尤以 GH/CBM 类别中的蛋白数量居多,为 30,所占比例高达 69.77%(表 7).上述复合类别的蛋白所具有的具体类别呈现出多样性特点,推测其在禾谷炭疽菌危害植物过程中发挥着多种功能.

表7 禾谷炭疽菌中复合类型的分布情况

不同类别	具体类别	ID	数量
GH/CBM	GH43/CBM13	GLRG-08138	1
	GH74/CBM1	GLRG-07406	1
	GH7/CBM1	GLRG-05335、GLRG-11038	2
	GH2/CBM35/GH27	GLRG-07024	1
	GH11/CBM1	GLRG-01857	1
	GH62/CBM13	GLRG-04764	1
	GH6/CBM1	GLRG-00951、GLRG-08518	2
	GH5/CBM2	GLRG-10073	1
	GH5/CBM1	GLRG-08534	1
	GH43/CBM35	GLRG-03535、GLRG-06015、GLRG-07933	3
	GH10/CBM1	GLRG-08966、GLRG-10506	2
	GH18/CBM18	GLRG-03690、GLRG-05986、GLRG-06994、GLRG-11080	4
	GH10/CBM2	GLRG-08206	1
	GH15/CBM20	GLRG-06159	1
	GH71/CBM24	GLRG-01017、GLRG-09772	2
	GH16/CBM6	GLRG-04807、GLRG-09790	2
	GH43/CBM6	GLRG-00820	1
	GH18/CBM50/CBM18	GLRG-11994	1
	AA/CBM	AA3/AA8/CBM1	GLRG-07538
AA9/CBM1		GLRG-01889、GLRG-09681、GLRG-09732、GLRG-10079、GLRG-11469	5
AA9/CBM18		GLRG-059061	1
CE/CBM	AA5/CBM32	GLRG-02805、GLRG-05590、GLRG-11847	3
	CE4/CBM18	GLRG-02208、GLRG-11238、GLRG-11241	3
合计	—	43	

3 讨论

上述研究结果与前人报道不仅在 CAZymes 总数量方面有所差别,而且在具体的酶的类别数量方面也有一定的差别^[18],究其原因,就总数量而言,由于本研究在数量统计方面,以 CAZymes 主要类别和复合类别进行的分别统计,使得归属类别方面存在一定的差异,若将复合类别中的 CAZymes 分别归属到主要类别中,那么总的数量为 321 个,当然该统计存在 54 个重复蛋白;此外,就不同类别中 CAZymes 数量而言,按照上述统计方法,本研究所获得 GHs、GTs、CBMs、CEs、PLs 以及 AAs 则分别含有 130、2、98、29、11、51 个。可以发现,本研究所获得 GTs 数量明显少于前人所报道的 53 个,原因尚不明确,有待于进一步研究。同时, O'Connell 等报道该菌中含有 349 个 CAZymes^[18],鉴于前人开展 CAZymes 的具体类别有 5 个,而根据最新的分类类别,有 6 个,造成具体不同类别分项数量方面的差异。就预测方法而言,本研究是基于前期通过全基因组序列预测所获得分泌蛋白,再依据现有网站中两种不同的预测方法所获得,是否为禾谷炭疽菌中的全部,均有待于进一步开展试验进行验证。

4 结论

目前,学术界关于禾谷炭疽菌中 Septin 蛋白^[19]、RGS 蛋白^[20]、14-3-3 蛋白^[21]等生物信息学分析的报道较多,对于 CAZymes 的报道尚未见报道。前期明确禾谷炭疽菌中含有 630 个分泌蛋白^[17],本研究利用 CAT 在线预测程序,明确该菌中 CAZymes 含有 267 个,分为主要类别和复合类别两大类,前者包括 89 个 GHs、53 个 CBMs 以及 41 个 AAs、28 个 CEs、11 个 PLs、2 个 GTs;后者则包括 30 个 GH/CBM、10 个 AA/CBM、1 个 CE/CBM 等。该研究为深入开展禾谷炭疽菌中的 CAZymes 的功能研究以及进一步找寻新的药剂作用靶标打下了坚实的理论基础。

参 考 文 献

- [1] Bergstrom G C, Nicholson R L. The biology of corn anthracnose: knowledge to exploit for improved management[J]. *Plant Disease*, 1999, 83(7): 596-608.
- [2] Callaway M, Smith M, Coffman W. Effect of anthracnose stalk rot on grain yield and related traits of maize adapted to the northeastern United States[J]. *Canadian Journal of Plant Science*, 1992, 72(4): 1031-1036.
- [3] Keller N, Bergstrom G. Developmental predisposition of maize to anthracnose stalk rot[J]. *Plant Disease*, 1988, (72): 972-980.
- [4] Leonard K, Thompson D. Effects of temperature and host maturity on lesion development of *Colletotrichum graminicola* on corn[J]. *Phytopathology*, 1976, (66): 635-639.
- [5] Warren H, Nicholson R, Ullstrup A, et al. Observations of *Colletotrichum graminicola* on sweet corn in Indiana[J]. *Plant Disease Reporter*, 1973(57): 143-144.
- [6] van den Brink J, de Vries R P. Fungal enzyme sets for plant polysaccharide degradation[J]. *Applied Microbiology and Biotechnology*, 2011, 91(6): 1477-1492.
- [7] 陈相永, 陈捷胤, 肖红利, 等. 植物病原真菌寄生性与分泌蛋白组 CAZymes 的比较分析[J]. *植物病理学报*, 2014, 44(2): 163-172.
- [8] Schmidt S M, Panstruga R. Pathogenomics of fungal plant parasites: what have we learnt about pathogenesis[J]. *Curr Opin Plant Biol*, 2011, 14(4): 392-399.
- [9] Cantarel B L, Coutinho P M, Rancurel C, et al. The Carbohydrate-Active EnZymes database (CAZY): an expert resource for Glycogenomics[J]. *Nucleic Acids Research*, 2009, 37(Database issue): D233-238.
- [10] Park B H, Karpinet T V, Syed M H, et al. CAZymes Analysis Toolkit (CAT): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZY database[J]. *Glycobiology*, 2010, 20(12): 1574-1584.
- [11] Lombard V, Golaconda Ramulu H, Drula E, et al. The carbohydrate-active enzymes database (CAZY) in 2013[J]. *Nucleic Acids Research*, 2014, 42(Database issue): D490-495.
- [12] Bourne Y, Henriessat B. Glycoside hydrolases and glycosyltransferases: families and functional modules[J]. *Current opinion in structural biology*, 2001, 11(5): 593-600.
- [13] Campbell J A, Davies G J, Bulone V, et al. classification of nucleotide-diphospho-sugar glycosyltransferases based on amino acid sequence similarities[J]. *Biochemistry Journal*, 1997, 326: 929-939.
- [14] Lombard V, Bernard T, Rancurel C, et al. A hierarchical classification of polysaccharide lyases for glycogenomics[J]. *Biochemistry*

- Journal, 2010, 432(3): 437-444.
- [15] Levasseur A, Drula E, Lombard V, et al. Expansion of the enzymatic repertoire of the CAZy database to integrate auxiliary redox enzymes[J]. *Biotechnology for Biofuels*, 2013, 6(1): 41.
- [16] Boraston A B, Bolam D N, Gilbert H J, et al. Carbohydrate-binding modules: fine-tuning polysaccharide recognition[J]. *Biochemistry Journal*, 2004, 382(Pt3): 769-781.
- [17] 韩长志. 全基因组预测禾谷炭疽菌的分泌蛋白[J]. *生物技术*, 2014, 24(2): 36-41.
- [18] O'Connell R J, Thon M R, Hacquard S, et al, et al. Lifestyle transitions in plant pathogenic *Colletotrichum fungi* deciphered by genome and transcriptome analyses[J]. *Nature Genetics*, 2012, 44(9): 1060-1065.
- [19] 韩长志. 禾谷炭疽菌中 3 个典型 Septin 的生物信息学分析[J]. *河南师范大学学报(自然科学版)*, 2014, 42(6): 108-114.
- [20] 韩长志. 禾谷炭疽菌 RGS 蛋白生物信息学分析[J]. *微生物学通报*, 2014, 41(8): 1582-1594.
- [21] 韩长志. 禾谷炭疽菌 14-3-3 蛋白生物信息学分析[J]. *河南师范大学学报(自然科学版)*, 2014, 42(3): 109-114.

Prediction for CAZymes Protein from *etotrichum graminicola* Genome

HAN Changzhi

(College of Forestry, Southwest Forestry University, The Key Laboratory of Forest Disaster Warning and Control of Yunnan Province, KunMing 650224, China)

Abstract: *Colletotrichum graminicola* can infect corn, wheat and other food crops, which can cause tremendous economic losses in agricultural production in many countries Based on preliminary 630 secreted proteins sequence in *C graminicola*, 267 Carbohydrate-Active Enzymes (CAZymes) proteins were obtained with CAT prediction algorithm, which were divided into two categories including main category and composite category. The former category includes 89 GHs, 53 CBMs and 41 AAs, 28 CEs, 11 PLs, 2 GTs; the latter includes 30 GH/CBMs, 10 AA/CBMs, 1 CE/CBMs and so on. The study provides an important theoretical basis for the further conduct of the action mechanism of plant pathogen infection.

Keywords: *Colletotrichum graminicola*; CAZymes; prediction algorithm