doi:10.3969/j.issn.1672-5565.2013.11.

纤维二糖水解酶 I 基因的系统进化分析

陈笑笑,田兴军*
(南京大学 生命科学学院,南京 210093)

摘 要:纤维二糖水解酶 I(CBH I)是生物降解纤维素的一种重要的外切酶,它作用于纤维素分子末端,水解 β-1,4-糖苷键。纤维二糖水解酶由 3 个部分组成:具有催化活性的催化结构域,作用为锚定纤维素的纤维素结合域以及连接这两个结构域的一段短肽。已知催化结构域属于糖基水解酶家族 7(GH7),纤维素结合域属于糖类结合模块家族 1(CBM1)。为进一步探索 CBHI 编码基因之间的进化关系,本研究依据 CBHI 的结构域在 GenBank 数据库中搜索并鉴定 CBHI 编码基因并据此构建系统发育树。序列的平均长度为 1 776 bp,平均 GC 含量为 57.64%,平均转换颠换比为 0.71,平均遗传距离为 0.424。得出结论 CBHI 编码基因只存在于真菌中,是一个相对活跃的基因,它的进化与物种的进化有着密切的关系。

关键词:纤维二糖水解酶 I(CBHI);系统进化;序列比对
中图分类号:Q343.1+5 **文献标识码**:A **文章编号**:1672-5565(2013)-04-300-09

Phylogeny tree analysis of cellobiohydrolase I

CHEN Xiao-xiao, TIAN Xing-jun*
(*School of life Science ,NanJing University, NanJing* 210093,*China*)

Abstract: Cellobiohydrolase I (CBHI) is an important exo-acting enzyme, playing a role in the conversion of cellulose to glucose by cutting the disaccharide cellobiose from the non-reducing end of the cellulose polymer chain. CBHI is composed of three parts: catalytic domain (CD), cellulose-binding domain (CBD) and a glycosylated linker peptide connecting CD and CBD. The CD of CBHI belongs to glycoside hydrolase family 7 (GH7) while the CBD belongs to carbohydrate-binding module family 1 (CBM1). For the discovery of phylogenetic relationship between CBHI genes, this article searched and identified CBHI genes from GenBank according to the domain of CBHI and then did phylogeny tree analysis. The average length of the sequences is 1776bp and G+C content of them is 57.64%. The average transition/transversion rate ratio is 0.71. The average genetic distance is 0.424.The result stated that CBHI genes exist only in fungus. CBHI genes are relatively active and the evolution of CBHI genes is significantly related to the evolution of species.

Keywords: Cellobiohydrolase I (CBHI); Phyletic Evolution; Sequence Alignment

前 言

纤维素是广泛存在于自然界的多糖化合物,纤维素水解能产生大量的小分子单糖,是重要的生物资源。[1]与纤维素的工业降解相比,生物降解纤维素具有清洁、快速、反应条件低、转化率高等优点,是近年来生物研究的一个重要课题。纤维素酶主要由三类具有不同催化反应功能的酶组成,即内切葡聚糖酶(endo-1,4-β-D-glucanase, EC3.2.1.4)、外切葡聚糖酶(exo-1,4-β-D-glucanase,EC3.2.1.91)和 β-葡萄糖苷酶((β-1, 4-glucosidase, EC3.2.1.21)[2]。外切葡聚糖酶作用于纤维素链末端,水解 β-1,4-糖苷键,每次切下一个纤维二糖分子,因此又被称为纤维二糖水解酶(CBH)。纤维二糖水解酶分为 CBHI 和 CBHII,其中 CBHI 作用于纤维素的还原性末端,

收稿日期:2012-06-06;**修回日期**:2013-08-25.
作者简介:陈笑笑,女,硕士研究生,研究方向:生物信息学;E-mail:cxx880421@163.com.
*通讯作者:田兴军,男,博士生导师,研究方向:生态学;E-mail:tianxj@nju.edu.cn.

而 CBHII 作用于非还原性末端[3],相比 CBHII,CBHI 在纤维素的降解中具有更重要的作用[4-5]。

CBHI 一般有两个结构域,催化结构域(,CD)和纤维素结合域(,CBD)。两个结构域之间由一段长度为 33~300 氨基酸不等的富含脯氨酸或羟基氨基酸的连接短肽相连[6]。CBHI 的 CD 属于糖基水解酶家族 7(Glycoside hydrolase family7,GH7),大约由 433 个氨基酸组成,其三围结构是由 5 个 α 螺旋和 7 条 β 链组成的具有很大凹面的 β 三明治。β 三明治的一面有两个从 GH7 蛋白内部延伸至表面的环形成的一条长约 50A 的隧道,GH7 蛋白的活性位点位于隧道的一端,作用于纤维素链的还原端,而纤维素链的非还原端的葡萄糖残基锚定在 GH7 蛋白上[7]。CBHI 的 CBD 位于 CBHI 的 C 端,它属于糖类结合模块家族 1(1,CBM1),能同时识别晶体和非晶体纤维素,插入纤维素的结晶区,加速纤维素的去结晶化。CBD 具有复杂的楔形结构,表面具有一些疏水的芳香族氨基酸的残基,与疏水的晶体纤维素分子表面结合[8],但是不具有催化活性。

1953 年,DNA 双螺旋结构被发现[9],标志生物科学进入了分子水平研究的时代。许多生物的全基因组被测序,许多 CBHI 编码基因和蛋白质也已测序成功,大量的研究阐明了 CBHI 蛋白质的结构和功能。GenBank 数据库中收录的 CBHI 编码基因已有 10 000 多条,但又缺乏系统的研究和分析[10]。本研究在各基因组和核苷酸序列中搜索并鉴定 CBHI 编码基因并据此构建系统发育树,从而探索 CBHI 编码基因在生物中的进化关系。

1　数据来源与研究方法

1.1　数据来源

数据材料:GH7 蛋白结构域氨基酸序列(Pfam:PF00840);CBM1 蛋白结构域氨基酸序列(Pfam:PF00734);GenBank 核苷酸序列数据库。

软件:BLAST,Pfam,MEGA5.2,Family。

1.2　研究方法

以 Pfam 数据库中注释的 GH7 蛋白结构域氨基酸序列和 CBM1 蛋白结构域氨基酸序列作为问询序列,使用 BLAST 和隐马尔科夫模型(Hidden markov model,HMM)对各基因组和核苷酸序列同源搜索,设置阈值 E 为 0.1[11-12]。将获得的可能具有 GH7-CBHI 结构域的 CBHI 完整编码序列重新在 Pfam 数据库中验证,设定 E 值为默认值。在 GenBank 中搜寻注释为 CBHI 完整编码序列的核苷酸序列,放入 Pfam 数据库中预测功能结构域。

用 MEGA5.2 软件的 Clustal W 功能对获得的 GH7-CBM1-CBHI 完整编码序列进行多重序列比对,参数设置默认,对比对结果进行手工矫正[13-14]。用核苷酸构成功能(Nucleotide composition)计算使用计算平均长度和 G+C 含量;使用核苷配对频率功能(Nucleotide pair frequencies)计算各个位点的转换值、颠换值及转换颠换比。以最大似然法为算法,使用寻找最优替代模型(Find Best DNA/Protein Models)功能寻找最适合的核苷替代模型[15]。使用距离功能(Compute pairwise distances)并采用 Maximum composite likelihood 模型和 JC 矫正[16-17],构建序列遗传距离矩阵。用邻接法构建系统进化树[18],参数默认以自展法(Bootstrap)检测,共循环 1 000 次,将 GenBank 中注释为 CBHI 完整编码序列的氨基酸序列整合进 GH7-CBM1-CBHI 完整编码序列,构建系统进化树,依据系统进化树对 GH7-CBM1-CBHI 完整编码基因手工分组,计算各组的同义突变 Ka 值,非同义突变 Ks 值,Ka/Ks 值[19-20]及遗传距离Pi 值。

2　研究结果

2.1　CBHI 基因的鉴定

通过 BLAST 和 HMM 搜寻,在 GenBank 核苷酸序列数据库中鉴定出其中有 37 条完整的编码序列,统计长度及 G+C 含量(见表 1)。将 GenBank 上获得的注释为 CBHI 完整编码序列的核苷酸序列在 Pfam 上推测结构域,据此得到 5 条不含有 CBM1 的 CBHI 编码基因序列(见表 2)。分析结果表明,目前收集到的 CBHI 编码基因均来源于真菌,根据安斯沃斯真菌分类系统对各真菌进行系统分类[11],可见 CBHI 真菌主要分布于散囊菌纲、伞菌纲以及粪壳菌纲。在植物基因组和动物基因组中,未鉴定出编码 GH7-CBM1 蛋白质的氨基酸序列。由表 1 可知,序列的平均 GC 含量为 57.64%,平均长度为 1 776 bp。

表 1 鉴定出含有 CBHI 编码基因的完整编码序列

Table 1 CBHI Gene Complete CDs

编号	登陆号	种名	纲名	长度(bp)	G+C(%)
1	KC478360. 1	*Phanerochaete chrysosporium*	Agaricomycetes	1 533. 0	63. 47
2	AY559102. 1	*Volvariella volvacea*	Agaricomycetes	1 710. 0	54. 85
3	AY864863. 2	*Aspergillus terreus*	Eurotiomycetes	1 843. 0	60. 17
4	XM_003653460. 1	*Thielavia terrestris*	Sordariomycetes	1 581. 0	65. 65
5	HQ141568. 1	*Neolentinus lepideus*	Agaricomycetes	1 587. 0	59. 11
6	XM_002562710. 1	*Penicillium chrysogenum*	Eurotiomycetes	1 590. 0	56. 04
7	AY973993. 1	*Penicillium chrysogenum*	Eurotiomycetes	1 590. 0	56. 04
8	L22656. 1	*Phanerochaete chrysosporium*	Agaricomycetes	1 772. 0	61. 63
9	AY861347. 1	*Chaetomium thermophilum*	Sordariomycetes	1 593. 0	60. 95
10	XM_003660741. 1	*Myceliophthora thermophila*	Sordariomycetes	1 934. 0	61. 27
11	JQ716928. 1	*Stereum hirsutum*	Agaricomycetes	1 545. 0	58. 64
12	AF411250. 1	*Lentinula edodes*	Agaricomycetes	1 551. 0	52. 87
13	AB002821. 1	*Aspergillus aculeatus*	Eurotiomycetes	1 848. 0	56. 49
14	AF420020. 1	*Aspergillus nidulans*	Eurotiomycetes	1 911. 0	56. 36
15	EU727171. 1	*Penicillium oxalicum*	Eurotiomycetes	1 638. 0	61. 05
16	GQ844299. 1	*Penicillium decumbens*	Eurotiomycetes	1 641. 0	61. 24
17	AB177377. 1	*Irpex lacteus*	Agaricomycetes	1 566. 0	59. 26
18	D63515. 1	*Humicola grisea var. thermoidea*	Eurotiomycetes	2 537. 0	62. 28
19	L29379. 1	*Fusarium oxysporum*	Sordariomycetes	1 675. 0	54. 99
20	AY342396. 1	*Gibberella zeae*	Sordariomycetes	1 539. 0	56. 66
21	AY368686. 1	*Trichoderma viride*	Sordariomycetes	1 746. 0	58. 59
22	AB540999. 1	*Flammulina velutipes*	Agaricomycetes	1 689. 0	53. 76
23	JN180490. 1	*Penicillium glabrum*	Eurotiomycetes	1 623. 0	57. 18
24	HM053612. 1	*Hypocrea lixii*	Sordariomycetes	1 518. 0	56. 59
25	HQ615690. 1	*Penicillium funiculosum*	Eurotiomycetes	1 590. 0	55. 03
26	AB103461. 1	*Athelia rolfsii*	Agaricomycetes	1 810. 0	54. 53
27	JQ238604. 1	*Hypocrea orientalis*	Sordariomycetes	1 680. 0	59. 11
28	AY690482. 1	*Penicillium occitanis*	Eurotiomycetes	3 032. 0	49. 44
29	FJ871063. 1	*Trichoderma viride*	Sordariomycetes	2 118. 0	55. 71
30	JN992645. 1	*Hypocrea lixii*	Sordariomycetes	1 467. 0	55. 69
31	GU385810. 1	*Aspergillus glaucus*	Eurotiomycetes	2 592. 0	56. 37
32	AB019377. 1	*Irpex lacteus*	Agaricomycetes	2 630. 0	53. 73
33	EU872026. 1	*Hypocrea virens*	Sordariomycetes	1 639. 0	52. 35
34	XM_003653709. 1	*Thielavia terrestris*	Sordariomycetes	1 446. 0	63. 69
35	XM_003663393. 1	*Myceliophthora thermophila*	Sordariomycetes	1 610. 0	65. 09
36	AF223252. 1	*Trichoderma harzianum*	Sordariomycetes	1 698. 0	54. 30
37	HQ843504. 1	*Penicillium oxalicum*	Eurotiomycetes	1 641. 0	61. 18
平均值				1 776. 0	57. 64

表 2　GenBank 核苷酸序列库中不含 CBM1 蛋白结构域的 CBHI 完整编码序列

Table 2　Complete CDs of GenBank CBHI genes without CBM1 Domain

编号	登录号	种名	纲名	长度(bp)	G+C(%)
1	AB298322. 1	*Polyporus arcularius*	Agaricomycetes	2 424. 0	58. 70
2	JF513053. 1	*Aspergillus niger*	Eurotiomycetes	1 530. 0	54. 12
3	AF478686. 1	*Thermoascus aurantiacus*	Agaricomycetes	2 424. 0	51. 61
4	AF156693. 2	*Volvariella volvacea*	Agaricomycetes	2 294. 0	55. 71
5	U25129. 1	*Cochliobolus carbonum*	Dothideomycetes	2 400. 0	53. 17
6	L43048. 1	*Cryphonectria parasitica*	Sordariomycetes	2 424. 0	52. 31

2. 2　CBHI 完整编码序列比对及序列位点分析

对材料进行多重序列比对,参数默认,GH7-CBM1-CBHI 完整编码基因有 513 个位点是保守的。核苷配对频率功能(Nucleotide pair frequencies)计算碱基配对频率(见表 3)。

由表 3 可知,在 GH7-CBM1-CBHI 编码基因中碱基变异类型以颠换为主。第三位点的碱基最保守,而第二位点和第一位点都很相对活跃所有位点。平均 R 值为 0. 71,第一位点 R 值为 0. 71,第二位点 R 值为 0. 79,第三位点 R 值为 0. 60,平均 R 值<2,说明 GH7-CBM1-CBHI 编码基因饱和度较高。

表 3　GH7-CBM1-CBHI 编码基因各位点碱基配对频率

Table 3　GH7-CBM1-CBHI Gene Nucleotide Pair Frequencies

密码子位点	ii	si	sv	R	TT	TC	TA	TG	CT	CC
平均值	995. 00	229. 00	324. 00	0. 71	189. 00	74. 00	27. 00	29. 00	66. 00	337. 00
第一位点	316. 00	83. 00	117. 00	0. 71	57. 00	32. 00	8. 00	8. 00	29. 00	139. 00
第二位点	316. 00	88. 00	112. 00	0. 79	56. 00	30. 00	9. 00	13. 00	28. 00	121. 00
第三位点	363. 00	57. 00	96. 00	0. 60	77. 00	12. 00	10. 00	8. 00	10. 00	77. 00

密码子位点	CA	CG	AT	AC	AA	AG	GT	GC	GA	GG
平均值	46. 00	58. 00	29. 00	47. 00	207. 00	45. 00	28. 00	61. 00	44. 00	262. 00
第一位点	17. 00	22. 00	9. 00	18. 00	51. 00	10. 00	8. 00	26. 00	11. 00	69. 00
第二位点	12. 00	22. 00	9. 00	13. 00	49. 00	15. 00	11. 00	21. 00	15. 00	89. 00
第三位点	16. 00	14. 00	10. 00	15. 00	106. 00	19. 00	8. 00	14. 00	17. 00	103. 00

注:ii 为一致对,si 为转换对,sv 为颠换对,R 为 si/sv。

Notes: ii: instant pairs, si: transition pairs, sv: transversion pairs, R: si/sv.

2. 3　CBHI 编码基因序列最优替代模型

根据 Finding Best DNA/Protein Model 功能得出各个模型套用在 GH7-CBM1-CBHI 序列组上的贝叶斯得分[21],赤池得分及等级制似然比得分(hLRT)[22]。对 3 个指标进行综合评价,得出 GH7-CBM1-CBHI 编码基因序列的最优碱基替代模型为 GTR+G+I 模型[23]。

2. 4　序列遗传距离矩阵

通过 JC 单参数模型矫正计算得到序列间遗传距离。综合各数据得到 GH7-CBM1-CBHI 完整编码序列的平均遗传距离为 0. 424,GH7-CBM1-CBHI 完整编码序列的遗传距离从 0 到 0. 697 不等。由此可知,一些不同物种间的 GH7-CBM1-CBHI 具有比较大的相似性差异。

2. 5　系统发育树构建

图 1 为基于 37 条 GH7-CBM1-CBHI 编码序列构建的 NJ 系统发育树,各分支上数值代表可信度。照图对该系统进化树分组,分组情况标注见图 1。

对各个真菌来源进行分类学鉴定,则可以很明显地发现,G1、G2 组中除了 *Aspergillus glaucus*(GenBank:GU385810. 1)都为木霉,G3 组和 G13 组属于肉座菌亚纲的半知菌,G4 组除了 *Humicola grisea* var. thermoidea(GenBank:D63515. 1)不能具体分类外,都属于粪壳菌目,毛壳菌科。G5 组属于曲霉,G6、G7、G8 组以及与 G8 组关系较近,但未被分进组的 *Penicillium glabrum*(GenBank:JN180490. 1)都为散囊菌纲发菌科的真菌,即主要为曲霉和青霉。G9 组为伞目的两种真菌,分属于光柄菇科和小皮伞科。与 G9

亲缘关系较近,但比 G9 组更早分化出去的 *Flammulina velutipes*(GenBank:AB540999.1)也属于伞目,但是它属于泡头菌科。G10、G11、G12 组都属于伞菌纲,但分属于不同的目。整棵系统进化树可分为三个大分类簇,G1 至 G4 为一大簇,均属于粪壳菌纲;G5 至 G8 为一大簇,都属于散囊菌纲;而 G9 至 G11 为一大簇,都属于伞菌纲。G13 组的两种菌虽然与 G3 组的两种物种分别是一样的,但是更早以前就分化出去了,可能 G13 组的这两条序列编码的 CBHI 演变出了不同的结构。总体看来,伞菌纲的 CBHI 编码序列最早分化,其后散囊菌纲和粪壳菌纲的 CBHI 编码序列发生分化。由此可见,GH7-CBM1-CBHI 编码基因表现为趋异进化, GH7-CBM1-CBHI 编码基因与物种进化有着内在的联系。

图 2 为整合了 GH7-CBM1-CBHI 完整编码序列和 GenBank 上不具有 CBM1 结构域的 CBHI 的完整编码序列后构建的 NJ 系统进化树。相比图 1,图 2 的系统进化树许多分支置信度很低,尤其是 *Aspergillus niger*(GenBank:JF513053.1)、*Thermoascus aurantiacus*(GenBank: AF478686.1)和 *Polyporus arcularius*(GenBank:AB298322.1)。

2.6 GH7-CBM1-CBHI 编码基因的替代速率

利用 Family 软件计算各组的 Ka、Ks、Ka/Ks 以及 Pi 值,计算结果如表 4 所示。其中 G3,G5,G9,G13 的 Ka/Ks 大于 1。在 G2,G6,G7,G8,G10,G11,G12 等组中,Ka/Ks 大于 1;在 G1,G4 中,Ka/Ks 约等于 1。各组的 Pi 值较低,表明在组内基因的同源性较高。

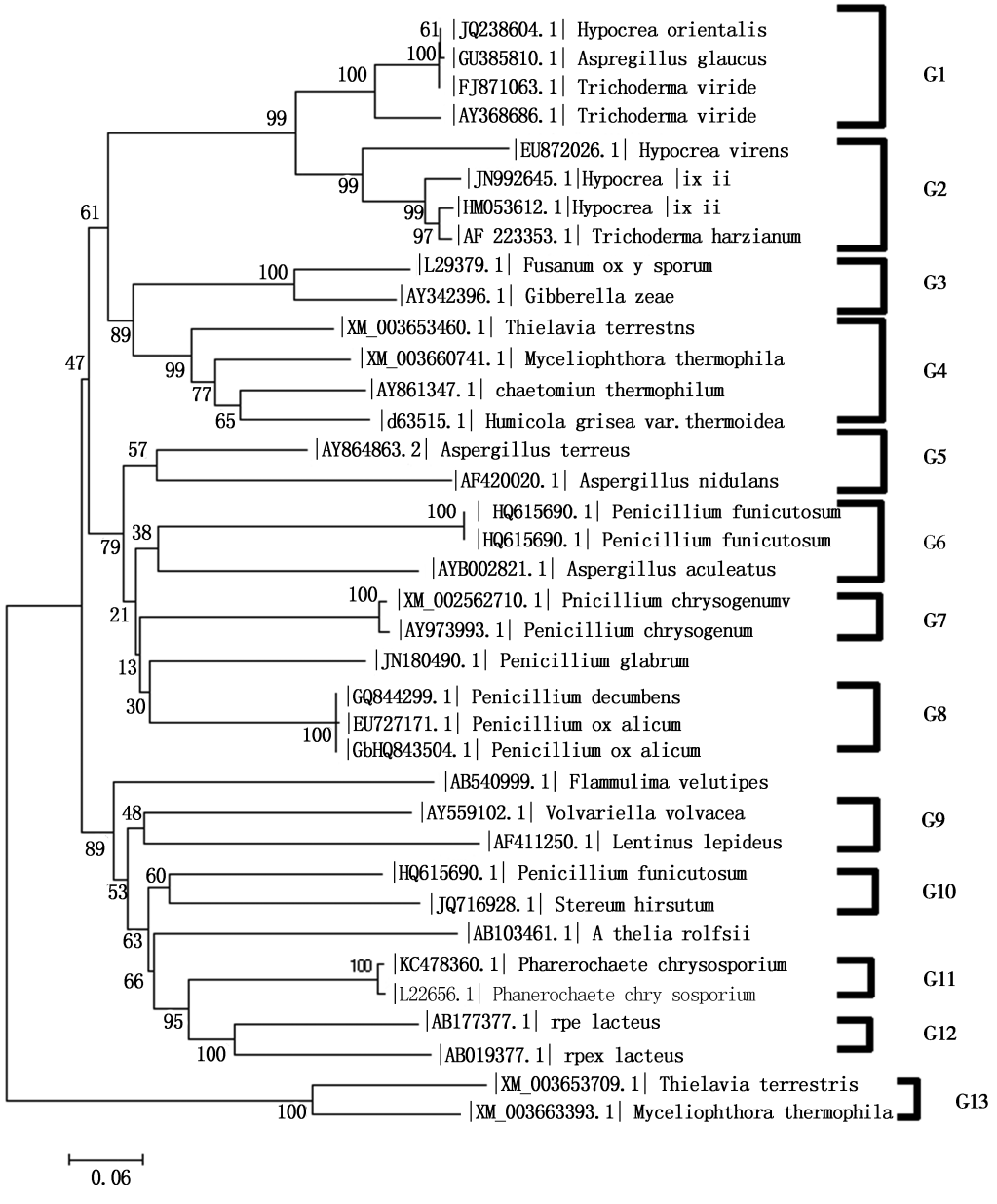

图 1 基于 GH7-CBM1-CBHI 编码序列构建的 NJ 进化树

Fig. 1 Neighbourhood-Joining Phylogeny Tree Based on GH7-CBM1-CBHI CDs

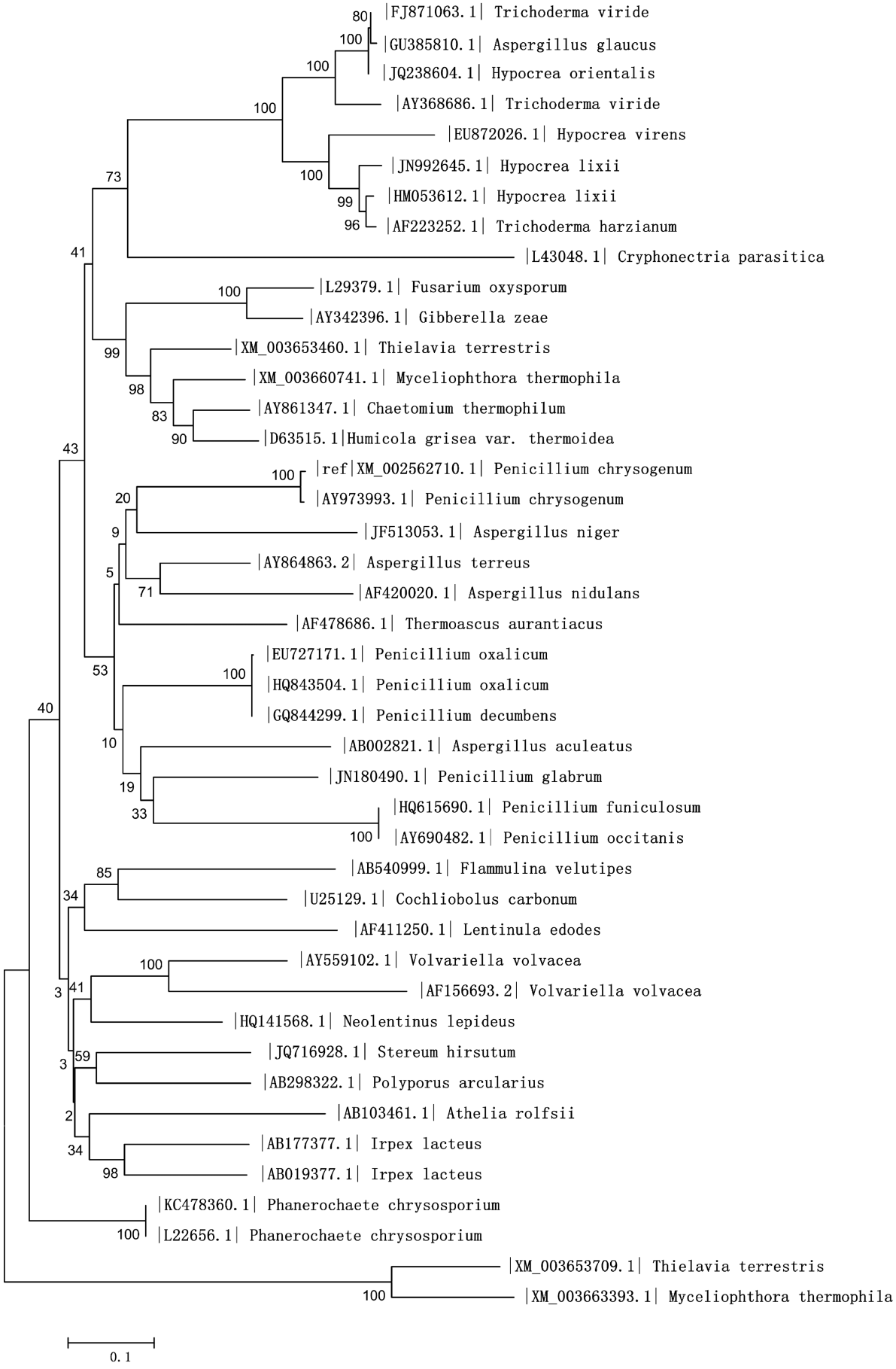

图 2　整合 CBHI 完整编码序列构建的 NJ 进化树

Fig. 2　Neighbourhood-Joining Phylogeny Tree Based on CBHI Complete CDs

表 4 GH7-CBM1-CBH1 各分组的平均 Ka 值,Ks 值,Ka/Ks 值和 Pi 值

Table 4 Average Ka, Ks, Ka/Ks, Pi value of GH7-CBM1-CBHI Groups

组名	Ka 值	Ks 值	Ka/Ks	Pi
G1	0.053 1	0.095 2	1.116 1	0.065 4
G2	0.099 1	0.153 6	0.548 5	0.114 8
G3	0.187 5	0.085 5	2.191 5	0.174 8
G4	0.233 9	0.311 3	1.060 0	0.250 9
G5	0.388 4	0.285 0	1.362 9	0.359 5
G6	0.201 8	0.880 1	0.229 0	0.304 6
G7	0.003 3	0.039 7	0.084 3	0.012 0
G8	0.000 5	0.005 0	0.091 1	0.001 6
G9	0.483 4	0.300 6	1.608 4	0.444 8
G10	0.246 0	0.905 8	0.271 6	0.357 9
G11	0.005 1	0.006 0	0.856 3	0.007 2
G12	0.242 3	0.409 7	0.591 5	0.281 9
G13	0.372 7	0.189 0	1.972 1	0.317 2

3 讨论

GH7 结构蛋白属于糖基水解酶家族(,GH),此家族是一类能够水解糖苷键的蛋白质的酶的总和。根据糖基水解酶家族蛋白质的序列和构象分类,能将其分为 100 多个家族[24-25]。GH7 家族含有 CBHI 和内切葡聚糖酶 I(EGI)。CBM1 结构蛋白属于糖类结合域家族(,CBM)[26],它不具有催化活性,由其表面的疏水残基负责将其吸附到纤维素上以帮助 CD 完成催化作用。根据氨基酸序列可以将 CBM 分成相当数量的家族,目前在 CAZy 数据库有 64 个 CBM 家族[27]。CBM1 蛋白由 36 个氨基酸组成。虽然有许多氨基酸序列被预测具有结合纤维素的结构域,但实际它们并没有结合纤维素的活性[28]。虽然 CBHI 的 GH7 结构域和 CBM1 结构域不能单独作为 CBHI 的特征结构域,但已知 EGI 的 CBD 并不属于 CBM1 家族,所以联合 GH7 蛋白和 CBM1 蛋白在 Pfam 上的 HMM 校准氨基酸序列,可以作为问询序列获得编码 GH7-CBM1-CBHI 的核苷酸序列。在实际操作过程中,在 GenBank 上搜到了相当多的注释为 CBHI 完整编码序列的核苷酸序列,在 Pfam 中推测其结构域,发现有些序列只含有 GH7 结构域而并不含有 CBM1 结构域,同时在 BLAST 过程中也搜索到大量注释为 EGI 编码序列但是具有 GH7 结构域不具有 CBM1 结构域的核苷酸序列,因此这些序列的保留或者摈弃成为研究的一个问题。作者保留了 6 条注释为 CBHI 编码序列但是只具有 GH7 结构域的核苷酸序列,整合在 GH7-CBM1-CBHI 编码核苷酸序列中构建系统进化树,相比基于 GH7-CBM1-CBHI 编码核苷酸序列构建的系统进化树,很明显构成的系统进化树,如图 2 所示,许多分支置信度很低。因此,不含有 CBM1 结构域的 CBHI 编码序列需要排除在外,不仅因为容易与 EGI 混淆,也因为它与 GH7-CBM1-CBHI 编码序列的差异太大,同源性低,将其放在系统进化树的分析对象中,会使系统进化树的可靠性降低。

当前生物学界对真菌分子鉴定,系统进化分析大多采用 18S rDNA 或核糖体基因转录间隔区(,ITS)作为分子标记,存在广泛的异种同源性[29]本实验构建的系统进化树与分类学进化树有着很大的相似性,伞菌纲的 GBHI 编码基因分化最早,其次是粪壳菌纲与散囊菌纲。在系统分类学中伞菌纲属于担子菌门,粪壳菌纲和散囊菌纲属于子囊菌门,因此伞菌纲分化时间更早,与另两种纲的分化时间可视作子囊菌门和担子菌门的分化时间,CBHI 编码基因可以作为系统分类学的一个参考。Kasuga 在 2002 年计算得出散囊菌纲真菌的 DNA 核苷酸替换速率为每年每个位点 0.9×10^{-9} 到 16.7×10^{-9}[30],据此速率以及 1980 年日本的 Kimura 提出的算法 $K\approx2kt$(其中 k 是单位时间内碱基替换的总频率,K 为 DNA 距离)和各 GH7-CBM1-CBHI 序列的构建的系统进化树的分支长度[31],大致计算得出伞菌纲 CBHI 编码基因分化时间为 13~253 Ma 前,散囊菌纲和粪壳菌纲的 CBHI 编码基因分化时间为 13~239 Ma 前。1999 年,Taylor 在化石中发现了粪壳菌纲,将粪壳菌纲和散囊菌纲的分化时间定为大约 400 Ma 前[32]。2001 年,Berbee 等人通过建立 SSU rRNA 的系统进化树,并以动物和真菌的分化时间为 965 Ma 作为依据,计算出粪壳菌纲和散囊菌纲的分化时间为 310 Ma前,子囊菌门和担子菌门的分化时间为 500 Ma。Heckman 使用了真菌-动物-植物的分化时间 1 576 Ma 作为参考,再分析这三者蛋白质的差异,得出结论认为粪壳菌纲和散囊菌纲在 670 Ma 已完成了分化,子囊菌门和担子菌门的分化时间为1 200~1 400 Ma 之间[33]。将计算出的 GH7-CBM1-CBHI 编码基因分化时间同两者的分类学分化时间相比较,则 CBHI 编码基因的分化时间晚于分类学分化时间。

在进化学分析中,同义替代和非同义替代的速率以及它们的速率比具有很重要的生物学意义。当一个基因没有受到自然选择压力的时候,它的同义突变和非同义突变的速率应该是相等的,一般情况下,由于大多数非同义替代是有害的,所以非同义替

代的速率低于同义替代的速率,而一旦新的有利于自然选择的非同义突变发生,则非同义突变的速率会大大高于同义突变的速率,此称为正选择效应。Ka/Ks 是用于判断是否有选择压力作用于一个基因的重要指标,当 Ka/Ks>1,则认为有正选择效应,Ka/Ks=1,则认为存在中性选择;Ka/Ks<1 时,则认为有纯化选择作用[20];Ka/Ks 计算结果表明,木霉、青霉和伞菌纲中除了伞目以外的各物种的 CBHI 受到了自然选择的净化作用;肉座菌亚纲、曲霉和伞目各物种的 CBHI 获得了正选择,推测他们的 CBHI 编码基因发生对自然选择的有利突变,而粪壳菌目的 CBHI 编码基因的 Ka/Ks 为 1.06,接近于 1。我们可以认为受到了中性选择或者效果不太明显的正选择作用,据此推测粪壳菌目的这几个物种的生存已经不太依赖于纤维素的代谢。

在 GenBank 上收集到的注释为 CBHI 编码基因有很多,其中有一部分不含有 CBD,因而没有作为分析材料,但是它们都含有 GH7 结构域,产的酶也具有外切葡聚糖酶的活性。其原因是 GH7 结构域表面具有识别纤维素的位点,同时具有 GH7 结构域和 CBM1 结构域的 CBHI 在与纤维素作用的时候,在 GH7 和 CBM1 上分别有一处位点与纤维素结合。很多实验表明,将 CBHI 的 CBD 移除,CBHI 的对结晶纤维素的活性降低了 50%到 80%,对非结晶的纤维素的活性不变,因为 CBM1 在锚定纤维素的时候在局部放大了底物浓度,根据酶学动力学,反应的速率会大大加快,同时它也破坏纤维素的晶体结构[34-37]。在 Pfam 数据库的注释中,GH7 蛋白家族只分布于真菌中,因此在真菌以外的物种的基因中无法找到含有 GH7 蛋白家族的基因。CBM1 在 Pfam 中即注释为真菌的 CBD,而 CBHI 本身就是定位于真菌产的外切葡聚糖酶,因此 CBHI 是一种高度特异性的酶。目前的研究表明它只存在于真菌中,但是产外切葡聚糖酶的生物还是有可能存在的,如细菌降解纤维素的外切葡聚糖酶则被称为 Cex,例如 *Cellulomonas fimi* 所产生的一种 Cex 的 CD 属于 GH10 家族[38-39],因此我们希望在其他物种中寻找到 CBHI 的同工酶。

参考文献(References)

[1] Klemm Dieter, Heublein Brigitte, Fink Hans-Peter, Bohn Andreas. Cellulose: fascinating biopolymer and sustainable raw material[J]. Angewandte Chemie International Edition, 2005,44(22):3358-3393.

[2] Goro Takada, Takashi Kawaguchi, Jun-Ichi Sumitani, Motoo Arai. Cloning, nucleotide sequence, and transcriptional analysis of *Aspergillus aculeatus* no. F-50 cellobiohydrolase I (cbhI) gene[J]. Journal of Fermentation and Bioengineering, 1998,85(1):1-9.

[3] Brian K. Barr, Yin-Liang Hsieh, Bruce Ganem, David B. Wilson. Identification of two functionally different classes of exocellulases[J]. Biochemistry, 1996,35(2):586-592.

[4] Jaana M. Uusitalo, K M Nevalainen, Anu M. Harkki, Jonathan K. C. Knowles, Merja E. Penttilä. Enzyme production by recombinant *Trichoderma reesei strains*[J]. Journal of biotechnology, 1991,17(1):35-49.

[5] J. Jia, P. S. Dyer, J. A. Buswell, J. F. Peberdy. Cloning of the cbhI and cbhII genes involved in cellulose utilisation by the straw mushroom *Volvariella volvacea*[J]. Molecular and General Genetics MGG, 1999,261(6):985-993.

[6] Fatma Bhiria, Ali Gargourib, Mamdouh Ben Alic, Hafedh Belghithb, Monia Blibecha, Semia Ellouz Chaabouni. Molecular cloning, gene expression analysis and structural modelling of the cellobiohydrolase I from *Penicillium occitanis*[J]. Enzyme and Microbial Technology, 2010,46(2):74-81.

[7] S. Shoemaker, V. Schweickart, M. Ladner, D. Gelfand, S. Kwok, K. Myambo, M. Innis. Molecular cloning of exo-cellobiohydrolase I derived from *Trichoderma reesei* strain L27[J]. Nature Biotechnology, 1983,1(8):691-696.

[8] M Linder, Tuula T. Teeri. The cellulose-binding domain of the major cellobiohydrolase of *Trichoderma reesei* exhibits true reversibility and a high exchange rate on crystalline cellulose[J]. Proceedings of the National Academy of Sciences, 1996,93(22):12251-12255.

[9] J. D. Watson, F. H. C. Crick. A structure for deoxyribose nucleic acid[J]. Nature, 1953,171:737-738.

[10] Dennis A Benson, Karsch-Mizrachi Ilene, David J Lipman, James Ostell, David L Wheeler. GenBank[J]. Nucleic acids research, 2012,40(D1):D48-D53.

[11] Alex Bateman, Ewan Birney, Richard Durbin, Sean R. Eddy, Kevin L. Howe, Erik L. L. Sonnhammer. The Pfam protein families database[J]. Nucleic acids research, 2004,32(suppl 1):D138-D141.

[12] Sean R Eddy. Hidden markov models[J]. Current opinion in structural biology, 1996,6(3):361-365.

[13] Tamura Koichiro, Peterson Daniel, Peterson Nicholas, Stecher Glen, Nei Masatoshi, Kumar Sudhir. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods[J]. Mol Biol Evol, 2011,28(10):2731-2739.

[14] Ramu Chenna, Hideaki Sugawara, Tadashi Koike, Rodrigo Lopez, Toby J. Gibson, Desmond G. Higgins, Julie D. Thompson. Multiple sequence alignment with the Clustal series of programs[J]. Nucleic acids research, 2003,31(13):3497-3500.

[15] F Rodríguez, J L Oliver, A Marín, J R Medina. The general stochastic model of nucleotide substitution[J]. Journal of theoretical biology, 1990,142(4):485-501.

[16] Cristiano Varin, Paolo Vidoni. A note on composite likelihood inference and model selection[J]. Biometrika, 2005,92(3):519-528.

[17] T.H. Jukes, C.R. Cantor. Evolution of protein molecules[J]. In: Munro HN (ed), Mammalian protein metabolism, II. New York, Academic Press,1969: 21.

[18] E. O. Wiley, Bruce S. Lieberman. Phylogenetics: theory and practice of phylogenetic systematics[M]. Wiley-Blackwell, 2011.

[19] L. D. Hurst. The Ka/Ks ratio: diagnosing the form of sequence evolution.[J]. Trends in genetics: TIG, 2002,18(9):486.

[20] Masatoshi Nei, Takashi Gojobori. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions.[J]. Molecular biology and evolution, 1986,3(5): 418-426.

[21] David Posada, Thomas R. Buckley. Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests[J]. Systematic biology, 2004,53(5):793-808.

[22] D.A.Williams. Improved likelihood ratio tests for complete contingency tables[J]. Biometrika, 1976,63(1):33-37.

[23] S Tavaré. Some probabilistic and statistical problems in the analysis of DNA sequences[J]. Lect. Math. Life Sci, 1986,17:57-86.

[24] B Henrissat. A classification of glycosyl hydrolases based on amino acid sequence similarities[J]. Biochem J, 1991,280 (Pt 2):309 -316.

[25] B Henrissat, I Callebaut, S Fabrega, P Lehn, J P Mornon, and G Davies. Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases [J]. Proceedings of the National Academy of Sciences, 1995,92(15): 7090-7094.

[26] N R Gilkes, B Henrissat, D G Kilburn, R C Miller Jr, R A Warren. Domains in microbial beta-1, 4-glycanases: sequence conservation, function, and enzyme families.[J]. Microbiological reviews, 1991,55(2):303-315.

[27] Brandi L. Cantarel, Pedro M. Coutinho, Corinne Rancurel, Thomas Bernard, Vincent Lombard, Bernard Henrissat. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics[J]. Nucleic acids research, 2009,37 (suppl 1):D233-D238.

[28] Yvette Roske, Anwar Sunna, Wolfgang Pfeil, Udo Heinemann. High-resolution Crystal Structures of Caldicellulosiruptor Strain Rt8B. 4 Carbohydrate-binding Module CBM27-1 and its Complex with Mannohexaose[J]. Journal of molecular biology, 2004,340 (3):543-554.

[29] B. Esteve-Zarzoso, C. Belloch, F. Uruburu, A. Querol. Identification of yeasts by RFLP analysis of the 5.8 S rRNA gene and the two ribosomal internal transcribed spacers [J]. International Journal of Systematic Bacteriology, 1999,49(1):329-337.

[30] Takao Kasuga, Tomas J. White, John W. Taylor. Estimation of nucleotide substitution rates in eurotiomycete fungi[J]. Molecular Biology and Evolution, 2002,19(12):2318-2324.

[31] Motoo Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences[J]. Journal of molecular evolution, 1980,16(2):111-120.

[32] T. N. Taylor, H. Hass, H. Kerp. The oldest fossil ascomycetes[J]. Nature, 1999,399(6737):648.

[33] Daniel S. Heckman, David M. Geiser, Brooke R. Eidell, Rebecca L. Stauffer, Natalie L. Kardos, S. Blair Hedges. Molecular evidence for the early colonization of land by fungi and plants[J]. Science, 2001,293(5532):1129-1133.

[34] Herman Van Tilbeurgh, Peter Tomme, Marc Claeyssens, Rama Bhikhabhai, Göran Pettersson. Limited proteolysis of the cellobiohydrolase I from Trichoderma reesei: Separation of functional domains[J]. FEBS letters, 1986,204(2):223-227.

[35] Peter Tomme, Herman Van Tilbeurgh, Göran Pettersson, Jozef Van Damme, Joel Vandekerckhove, Jonathan Knowles, Tuula Teeri, Marc Claeyssens. Studies of the cellulolytic system of Trichoderma reesei QM 9414 [J]. European Journal of Biochemistry, 1988,170(3):575-581.

[36] Göran Pettersson, Markus Linder, Tapani Reinikainen, Torbjörn Drakenberg, Maija-Liisa Mattinen, Arto Annila, Maarit Kontteli, Gunnar Lindeberg, Jerry Ståhlberg. Identification of functionally important amino acids in the cellulose - binding domain of Trichoderma reesei cellobiohydrolase I[J]. Protein Science, 1995, 4(6):1056-1064.

[37] Jerry Ståhlberg, Gunnar Johansson, Göran Pettersson. A new model for enzymatic hydrolysis of cellulose based on the two-domain structure of cellobiohydrolase I[J]. Nature Biotechnology, 1991,9 (3):286-290.

[38] Alasdair M. MacLeod, Thisbe Lindhorst, Stephen G. Withers, R. Antony J. Warren. The acid/base catalyst in the exoglucanase/ xylanase from Cellulomonas fimi is glutamic acid 127: evidence from detailed kinetic studies of mutants[J]. Biochemistry, 1994,33 (20):6371-6376.

[39] N.R. Gilkes, B Henrissat, D.G. Kilburn, R.C. Miller, Jr, R.A. Warren. Domains in microbial beta-1, 4-glycanases: sequence conservation, function, and enzyme families[J]. Microbiol Rev, 1991,55(2):303-315.