TAP CHI SINH HOC 2018, 40(1): 39-50
DOI: 10.15625/0866-7160/v40n1.9200

XÂY DỰNG PROBE ĐỂ KHAI THÁC VÀ CHỌN GEN MÃ HÓA ENDO 1-4 XYLANASE TỪ DỮ LIỆU GIẢI TRÌNH TỰ DNA METAGENOME

Nguyễn Minh Giang[1], Đỗ Thị Huyền[2], Phùng Thu Nguyệt[2], Trương Nam Hải[2]*

[1]Đại học Sư phạm TP. Hồ Chí Minh
[2]Viện Công nghệ sinh học, Viện Hàn lâm KH & CN Việt Nam

TÓM TẮT: Theo phân loại của CAZY, endo 1-4 xylanase thuộc về các họ glycoside hydrolase (GH) 5, 8, 10, 11, 30, 51, 98. Tuy nhiên, chúng tôi chỉ tìm kiếm được ba trình tự thuộc GH8, 27 trình tự thuộc GH10, 18 trình tự thuộc GH11, một trình tự thuộc GH30 và một trình tự thuộc GH51 đã được nghiên cứu kỹ về hoạt tính sinh học, đặc điểm của enzyme. Dựa trên trình tự nhận được, hai probe cho endo 1-4 xylanase GH10 và GH11 đã được xây dựng. Probe cho GH10 dài 338, chứa 16 amino acid hoàn toàn giống nhau ở các trình tự, 13 amino acid giống nhau ở đa số các trình tự, 14 amino acid giống nhau ở một số trình tự với có điểm tối đa trên 189, độ bao phủ thấp nhất là 88%, độ tương đồng thấp nhất 39%. Probe cho họ GH11 dài 204 amino acid, trong đó có 54 amino acid hoàn toàn giống nhau ở các trình tự, 25 vị trí giống nhau ở đa số các trình tự, 24 vị trí giống nhau ở một số trình tự với có điểm tối đa trên 165, độ bao phủ thấp nhất là 84%, độ tương đồng thấp nhất 50%. Sử dụng hai probe trên chúng tôi lọc được duy nhất một trình tự GL0018509 mã hóa endo 1-4 xylanase GH10 từ dữ liệu giải trình tự DNA metagenome của vi khuẩn trong ruột mối. Kết quả ước đoán cấu trúc không gian bằng Phyre2 và Swiss Prot cho thấy GL0018509 tương đồng cao (95% và 93,4%) với endo 1-4 xylanase và được ước đoán là endo 1-4 xylanase với độ chính xác là 100%.

Từ khóa: Coptotermes gestroi, BLASTP, DNA metagenome, ClustalW, endo 1-4 xylanase, glycoside hydrolase (GH), probe.

MỞ ĐẦU

Enzyme endo 1-4 xylanase (EC 3.2.1.8) là nhóm enzyme phổ biến nhất thuộc nhóm hemicellulase, giúp phân cắt xylan thành đường xylobiose và các chuỗi polysaccharide ngắn, đóng vai trò quan trọng trong sản xuất cồn sinh học thế hệ thứ hai từ các nguồn sinh khối lignocellulose. Trong công nghệ sản xuất giấy và bông vải sợi, xylanase làm cho cấu trúc sợi xốp và dễ thấm hơn, nên dễ dàng loại bỏ được lignin, giúp tăng cường cho việc tẩy trắng. Mặt khác, xylanase có thể làm tăng khả năng tách sợi của lignin bằng cách giảm số lượng các tổ hợp lignin-carbohydrate có trong các sợi bột giấy. Việc sử dụng endo 1-4 xylanase chịu nhiệt cho quá trình tiền xử lý và thủy phân đã mang lại lợi ích lớn cho ngành công nghiệp này (Wang, 2013). Tác dụng phân giải xylan - một thành phần có nhiều trong thức ăn của vật nuôi, đã làm giảm độ nhớt của thức ăn, giúp cho vật nuôi tiêu hóa và hấp thụ chất dinh dưỡng tốt hơn, cải thiện hệ vi sinh vật đường ruột theo hướng có lợi (He et al., 2010). Ngoài ra enzyme này được sử dụng gạn lọc chất xơ trong công nghiệp nước hoa quả và rượu vang, hóa lỏng chất nhầy trong quá trình tạo cà phê lỏng, tách chiết hương liệu và chất màu, dầu thực vật và tinh bột (Kamble & Jadhav, 2012). Endo 1-4 xylanase đã được tìm thấy trong rất nhiều các loài sinh vật như vi khuẩn, nấm, xạ khuẩn, tảo biển, thực vật ở cạn (Rawashdeh et al., 2005). Trong đó nấm sợi là nguồn cung cấp xylanase phong phú nhất. Endo 1-4 xylanase của nấm đa số hoạt động ở nhiệt độ tối ưu từ 45^{o}C đến 55^{o}C, trong môi trường axit đến trung tính và ít tìm thấy hoạt động ở môi trường kiềm (Subramaniyan & Prema, 2002). Các nghiên cứu đầu tiên về xylanase chịu kiềm đã được công bố của vi khuẩn *Bacillus* sp. C-59-2, sau đó nhiều các nghiên cứu đã được tìm thấy ở *B. halodurans* C-125, *Bacillus* sp. AR-009, *Bacillus* sp. 41M-1 và *B. pumilus* 13a hoạt động tốt ở pH 9-10. Trong thực tế, vi khuẩn có khả năng sinh tổng hợp xylanase bền với nhiệt và thích hợp với môi trường từ pH trung tính đến pH kiềm (Subramaniyan & Prema, 2002). Tìm kiếm nguồn enzyme này có khả năng chịu được môi trường kiềm và nhiệt độ cao rất cần thiết

trong sản xuất nhiên liệu sinh học và các ngành công nghiệp khác.

Năm 2012, DNA metagenome của vi sinh vật ruột mối *Coptotermes gestroi* đã được phòng Kỹ thuật di truyền giải trình tự với tổng dung lượng là 5,6 GB, với 125.431 ORF đã được ước đoán dựa trên dữ liệu KEGG và eggNOG. Từ nguồn dữ liệu này 587 ORF mã hóa enzyme thủy phân lignocellulose đã được lọc ra. Chúng tôi đặc biệt quan tâm đến nhóm enzyme xylanase, vì vậy sẽ tập trung khai thác từ nguồn dữ liệu giải trình tự DNA metagenome của vi sinh vật trong ruột mối *C. gestroi* (Do et al., 2014). Các ORF từ dữ liệu trên tiếp tục được đánh giá theo bốn bước: (1) vùng trình tự bảo thủ dựa trên BlastP và BlastPSI (lựa chọn các trình tự có vùng hoạt tính rõ ràng, đặc hiệu và đặc thù cho enzyme cần chọn); (2) đánh giá so sánh tương đồng với các trình tự tương ứng trên ngân hàng gen và dựa vào cây phát sinh (lựa chọn các trình tự nằm trong nhóm enzyme cần lựa chọn có độ tin cậy cao); (3) truy nguyên nguồn gốc của gen (ưu tiên các gen có nguồn gốc từ vi khuẩn); (4): Lựa chọn trình tự đơn giản để dễ dàng biểu hiện gen trong *E. coli*. Trong thực tế 2 trong 4 trình tự được lựa chọn theo cách trên đã biểu hiện ở dạng không tan trong *E. coli* và không có hoạt tính sinh học. Nguyên nhân có thể do phương pháp lựa chọn gen dựa vào nguồn dữ liệu của NCBI thông qua so sánh tương đồng có những hạn chế do nhiều trình tự của NCBI chưa được nghiên cứu chứng minh bằng thực nghiệm. Vì vậy, việc tìm kiếm phương pháp để có thể lựa chọn nhanh được gen mã hóa enzyme mong muốn có hoạt tính sinh học từ metagenome rất cần thiết.

Sử dụng mẫu dò (probe) để tìm kiếm, sàng lọc các trình tự gen mong muốn từ các ngân hàng metagenome đã được nhiều nhóm nghiên cứu trên thế giới sử dụng (Kushwaha et al., 2015; Zhou et al., 2015; Baldwin et al., 2014; Akama et al., 2013). Ngoài ra, probe còn được dùng nhiều trong các nghiên cứu như nhận diện các bản sao hoặc sản phẩm RNA của gen, các sinh vật có quan hệ gần gũi với đối tượng nghiên cứu nhằm tìm kiếm gen chức năng được bảo tồn qua tiến hoá, tìm kiếm trình tự trọn vẹn của gen mã hoá cho protein trong genome (Mitsuhashi et al., 1994). Trong nghiên cứu này, chúng tôi xây dựng probe để tìm kiếm trình tự gen mã hóa enzyme endo 1-4 xylanase từ dữ liệu giải trình tự metagenome nói chung và của mối *C. gestroi* nói riêng. Probe được xây dựng dựa trên các trình tự đã được nghiên cứu kỹ về mặt thực nghiệm để chứng minh hoạt tính sinh học cũng như đặc điểm của enzyme. Do các họ enzyme (CAZy) được phân loại dựa trên cả trình tự và cấu trúc nên các probe dùng cho lựa chọn gen cũng sẽ được xây dựng riêng rẽ theo họ enzyme. Probe hứa hẹn sẽ tiết kiệm được thời được thời gian tìm kiếm, lựa chọn trình tự gen, đồng thời giúp biểu hiện protein/enzyme tan và hoạt tính trong thực nghiệm.

VẬT LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

Dữ liệu DNA metagenome của vi sinh vật sống tự do trong ruột mối đã được giải trình tự và ước đoán được 125.431 khung đọc mở (ORF) trong đó dự đoán có 27 ORF mã hóa cho endo 1-4 xylanase (Do et al. 2014). Dữ liệu này được dùng làm nguồn cho khai thác gen endo 1-4 xylanase bằng probe. Các trình tự endo 1-4 xylanase từ ngân hàng NCBI và CAZy.

Xác định các họ GH có chứa enzyme endo 1-4 xylanase theo CAZy và lựa chọn trình tự cho xây dựng probe

CAZy (Carbohydrate-Active enZYmes, http://www.cazy.org) là một hệ thống phân loại chứa cơ sở dữ liệu về enzyme tham gia vào quá trình tổng hợp, trao đổi và vận chuyển carbohydrate. CAZy cung cấp số liệu trực tuyến và cập nhật liên tục các dữ liệu của GenBank về chuỗi thông tin của gần 340.000 enzyme (Lombard et al., 2014; Cantarel et al., 2009) trong đó có phân loại rõ các trình tự đã được nghiên cứu kỹ về tính chất sinh hóa. Dựa trên sự tương đồng về trình tự, cấu trúc phân tử, CAZy xác định các họ thủy phân các liên kết glycosyl (Hydrolases Glycosyl: GH) có liên quan tiến hóa được giới thiệu bởi Henrissat (Henrissat, 1991) và đến năm 2015, dữ liệu CAZy chứa 135 họ GH với các đặc điểm nhận biết khác nhau. Chúng tôi sử dụng dữ liệu phân loại của CAZy để xác định có bao nhiêu họ GH chứa enzyme endo 1-4 xylanase. Từ các họ GH lọc được từ dữ liệu, chúng tôi tìm kiếm các trình tự đã được nghiên cứu kỹ về hoạt tính, tính chất sinh lý,

sinh hóa của enzyme và xếp chúng vào cùng họ với nhau. Dựa trên dữ liệu này chúng tôi sẽ xây dựng probe cho khai thác gen mã hóa endo 1-4 xylanase. Ngoài dữ liệu CAZy, chúng tôi cũng tìm kiếm thêm các trình tự từ dữ liệu NCBI cũng đã được nghiên cứu để làm phong phú dữ liệu và probe xây dựng được sẽ đại diện tốt hơn cho các trình tự.

Xây dựng probe và giá trị tham chiếu

Các trình tự của mỗi họ GH thu thập được ở trên được so sánh bằng phần mềm ClustalW - PBIL (http://expasy.org/proteomics). Kết quả so sánh của phần mềm này sẽ cho ra một trình tự được cho là bảo tồn cao nhất (ký hiệu *Prim.cons*), đồng thời cũng sử dụng màu sắc kết hợp các ký hiệu đặc trưng chỉ ra các vị trí mà amino acid giống nhau hoàn toàn trong các trình tự, hoặc giống nhau ở đa số các trình tự.

Dựa trên kết quả của việc so sánh ở trên và dựa vào trình tự bảo tồn cao nhất, các gốc amino acid giống nhau hoặc tương đối giống nhau sẽ ưu tiên lựa chọn để làm probe và các trình tự khác nhau quá nhiều sẽ được loại bỏ. Probe này sẽ được so sánh lại với các trình tự đã được sử dụng để xây dựng nên probe để xác định giá trị tham chiếu về điểm tối đa (max score), mức độ bao phủ và tương đồng của probe với trình tự đã sử dụng bằng BLASTP. Giá trị tham chiếu sử dụng để khai thác gen mã hóa endo 1- 4 xylanase từ dữ liệu giải trình tự DNA metagenome là giá trị điểm tối đa, mức độ bao phủ và độ tương đồng thấp nhất mà probe còn nhận biết được trình tự chuẩn dùng để xây dựng nên probe.

Khai thác trình tự mã hóa endo 1- 4 xylanase dữ liệu DNA metagenome của vi sinh vật trong ruột mối

Sau khi có các probe mã hóa cho endo 1- 4 xylanase thuộc các họ GH khác nhau, chúng tôi tiếp tục sử dụng BLASTP để so sánh probe với trình tự các amino acid của các ORF thuộc metagenome của vi sinh vật trong ruột mối và sử dụng ngưỡng phát hiện ở trên để lọc trình tự. Kết quả của việc sử dụng probe tìm kiếm các gen mã hóa cho enzyme sẽ được so sánh với kết quả dự đoán dựa trên con đường trao đổi chất KEGG do Viện Nghiên cứu hệ gen Bắc Kinh (Beijing Genomics Institute: BGI) thực hiện.

Ước đoán cấu trúc bậc ba của trình tự mã hóa endo 1- 4 xylanase được khai thác

Để khẳng định kết quả lọc gen bằng probe, chúng tôi có kiểm tra lại cấu trúc bậc ba của các trình tự. Do các trình tự ở đây đều là các trình tự mới, chúng tôi đã sử dụng hai phần mềm khác nhau là SWISSprot và Phyre2 để ước đoán cấu trúc.

KẾT QUẢ VÀ THẢO LUẬN

Các họ GH có hoạt tính endo 1- 4 xylanase

Bảng 1. Các họ GH chứa enzyme endo 1-4 xylananse theo CAZY

Mã E.C	GH	Clan	Mô hình cấu trúc không gian	Chất cho điện tử xúc tác	Chất cho proton xúc tác	Cơ chế xúc tác
3.2.1.8	5	GH-A	$(\beta/\alpha)_8$	Glu	Glu	Giữ nguyên
3.2.1.8	8	GH-M	$(\alpha/\alpha)_6$	Asp	Glu	Đảo ngược
3.2.1.8	10	GH-A	$(\beta/\alpha)_8$	Glu	Glu	Giữ nguyên
3.2.1.8	11	GH-C	β-jelly roll	Glu	Glu	Giữ nguyên
3.2.1.8	30	GH-A	$(\beta/\alpha)_8$	Glu	Glu	Giữ nguyên
3.2.1.8	51	GH-A	$(\beta/\alpha)_8$	Glu	Glu	Giữ nguyên
3.2.1.8	98		Chưa xác định	Asp	Glu	Đảo ngược

Trên cơ sở số liệu của CAZy, dựa trên sự bảo tồn cao về sự cuộn, gấp trong cấu trúc không gian các enzyme được sắp xếp vào các nhóm lớn (clan). Enzyme endo 1- 4 xylanase thuộc về 3 nhóm lớn là GH-A, GH-M, GH-C và được sắp xếp vào 7 họ là GH5, 8, 10, 11, 30, 51, 98 (bảng 1). Theo kết quả này bốn họ GH5, GH10, GH30 và GH51 thuộc về nhóm lớn GH-A giống nhau về cấu trúc không gian là $(\beta/\alpha)_8$, GH8 và

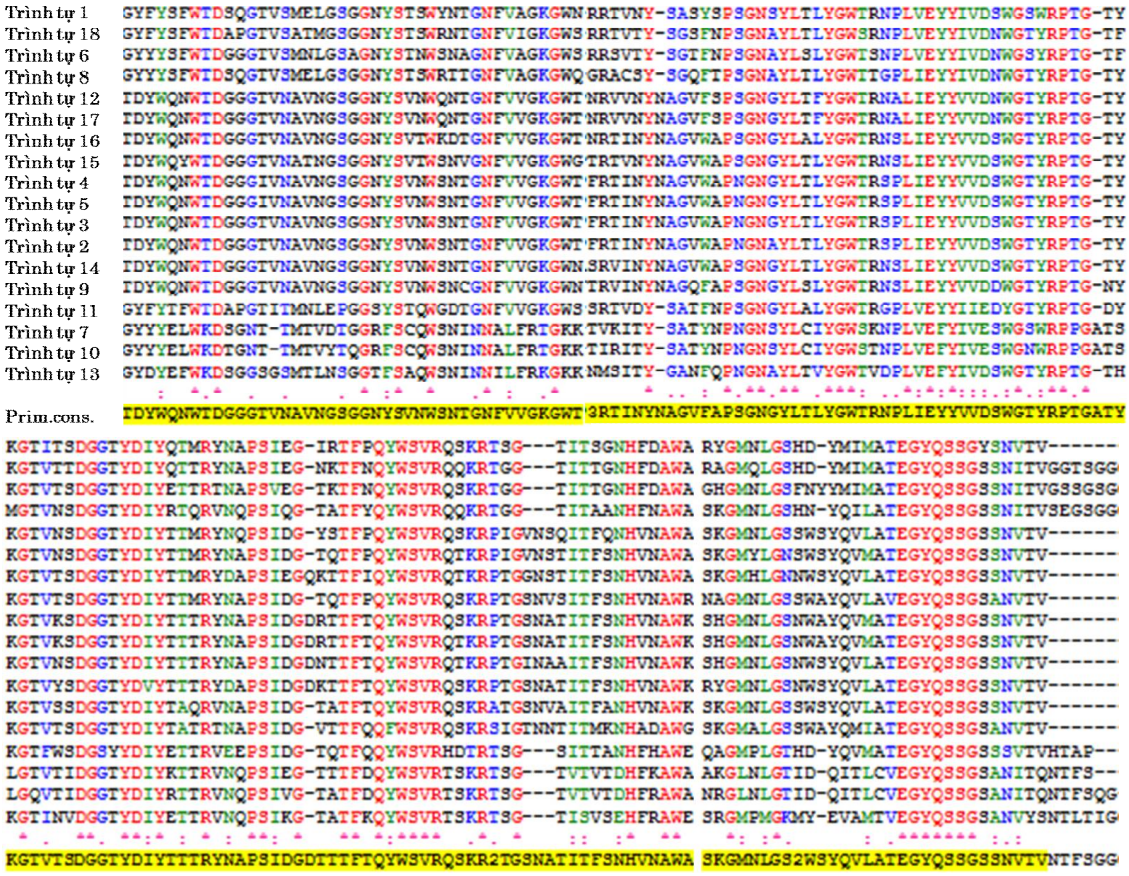
GH11 lần lượt thuộc về hai nhóm lớn là GH-M và GH-C, với mô hình cấu trúc không gian là $(\alpha/\alpha)_6$ và β-jelly roll, riêng GH98 vẫn chưa xác định được cấu trúc không gian nên chưa phân vào nhóm lớn. Các họ GH chứa endo 1- 4 xylanase đều có chất cho điện tử và proton trong quá trình hoạt động của enzyme là glutamate (Glu) trừ họ GH8 và GH98 chất cho điện tử là aspactate (Asp). Trong hai cơ chế phản ứng xúc tác thủy phân liên kết glycoside thường thấy nhất mà Koshland (1953) đưa ra gồm có cơ chế giữ nguyên và đảo ngược, họ GH8 và GH98 thuộc về cơ chế đảo ngược, còn lại đều thực hiện theo cơ chế giữ nguyên (Koshland, 1953).

Khai thác các trình tự amino acid của enzyme endo 1- 4 xylanase đã được nghiên cứu đặc tính

Bảng 2. Tổng hợp dữ liệu đã được nghiên cứu chi tiết về endo 1-4 xylanase

STT	Mã số trong GENBANK	Vi khuẩn	Số amino acid	pH tối ưu	Nhiệt độ tối ưu (^oC)	Nguồn http://www.ncbi.nlm.nih.gov/pubmed
		GH10				
1	ACY69980.1	*Alicyclobacillus* sp. A4	338	5.5	70	19916085
2	ADK91076.1	*Alicyclobacillus* sp. A4	411	6.2	55	20169343
3	ACY69979.1	*Anoxybacillus* sp. E2(2009)	328	7.8	60	DOI:10.1007/s11274-009-0254-5
4	AAQ83581.1	*Bacillus firmus*	396	4-11	70	15184083
5	AFE82288.1	*Bacillus* sp. HJ2	329		35	22534297
6	CAA84631.1	*Bacillus* sp.	331	8	40	7793963
7	AAB70918.1	*Bacillus* sp. NG-27	405	8,4	70	10831448
8	AGA16736.1	*Bacillus* sp. SN5	338	7	40	22864505
9	CBH32823.1	*Bacteroides xylanisolvens*B1A	378	6	37	20532756
10	AAA96979.1	*Dictyoglomus thermophilum*	352	6.5	85	8534104
11	AEO12683.1	*Paenibacillus xylanilyticus* J03	344	7.4	40	23462014
12	ACN76857.1	*Glaciecola mesophila* KMM241	423	7	30	19506861
13	ACR61562.1	*Sphingobacterium* sp. TN19	384	6.5	45	19554324
14	AAD32560.1	*Streptomyces avermitilis*	438	7.5	60	18645964
15	AAY98787.1	*Flavobacterium* sp. MSY2	371		30	16450065
16	AEO96821.1	*Geobacillus* sp. 71	407	7	75	22806019
17	ACX42569.1	*Geobacillus* sp. TC-W7	407	8.2	75	23075790
18	AAZ74783.1	*Geobacillus* sp. MT-1	331	7	70	16607523
19	AEP39603.1	*Geobacillus thermodenitrificans*	407	6	70	23156689
20	AEW07375.1	*Geobacillus thermoleovorans*	407	8,5	80	22212694
21	ACJ73932.1	*Kocuria* sp. MN22	389	8.5	55	19809242
22	AFE82289.1	*Lechevalieria* sp. HJ3	367	6	70	22430498
23	CBA13561.1	*Paenibacillus barcinonensis*	320	9.5	60	20218604
24	ACJ06666.1	*Paenibacillus* sp.	332	11.0	50	20493247
25	BAB40957.1	*Acidobacterium capsulatum*	405	5	65	9692186
26	AGA16736.1	*Bacillus* sp. SN5	338	7	40	22864505
27	CBH32823.1	*Bacteroides xylanisolvens*XB1A	378	6	37	20532756
		GH11				
1	BAK09352.1	*Actinomadura* sp. S14	228	6	80	21269876
2	CAD60654.1	*Bacillus* sp. BP-7	213	6	60	15057452
3	ACT79298.1	*Bacillus subtilis*	213	7	50-55	20075612
4	AAZ17393.1	*Bacillus subtilis*	213	9	50-60	16907724
5	BAH28803.1	*Bacillus subtilis*	213	6	40-50	19332293
6	AFH35005.1	Vi khuẩn gram âm	341	5	50	22487213

7	AFO70072.1	*Caldicellulosiruptor sp.* F32	357	5	75	25576604
8	CAJ57849.1	*Cellulomonas flavigena*	332	6.5	55	20231092
9	ACF57947.1	*Streptomyces* sp. S9	340	6.5	60	18521591
10	AAC46361.1	*Dictyoglomus thermophilum*	360	6.5	75-80	9572948
11	ACY70399.1	*Nesterenkonia xinjiangensis*	241	7	55	19838860
12	AHC74025.1	*Paenibacillus arcinonensis*	210	6.5	50	24549767
13	AEI54132.1	*Paenibacillus campinasensis*	377	7.5	60	22580312
14	ADK47978.1	*Paenibacillus polymyxa*	211	7	40	21161608
15	ABI96991.1	*Paenibacillus* sp.	211	6	60	18051350
16	AEB00656.1	*Paenibacillus* sp. ICGEB2008	204	7	50	21642416
17	BAE93061.1	*Paenibacillus* sp.	211	7	50	16348410
18	ACF57947.1	*Streptomyces* sp. S9	360	6.5	60	18521591



Hình 1. Kết quả so sánh sự tương đồng các trình tự amino acid của endo 1- 4 xylanase thuộc họ GH11

Trình tự Prim. Cons. được tô màu là trình tự sẽ được dùng làm probe. Mức độ bảo thủ của các gốc amino acid được đánh dấu từ dấu * đến dấu ":" dấu "." và không được đánh dấu.

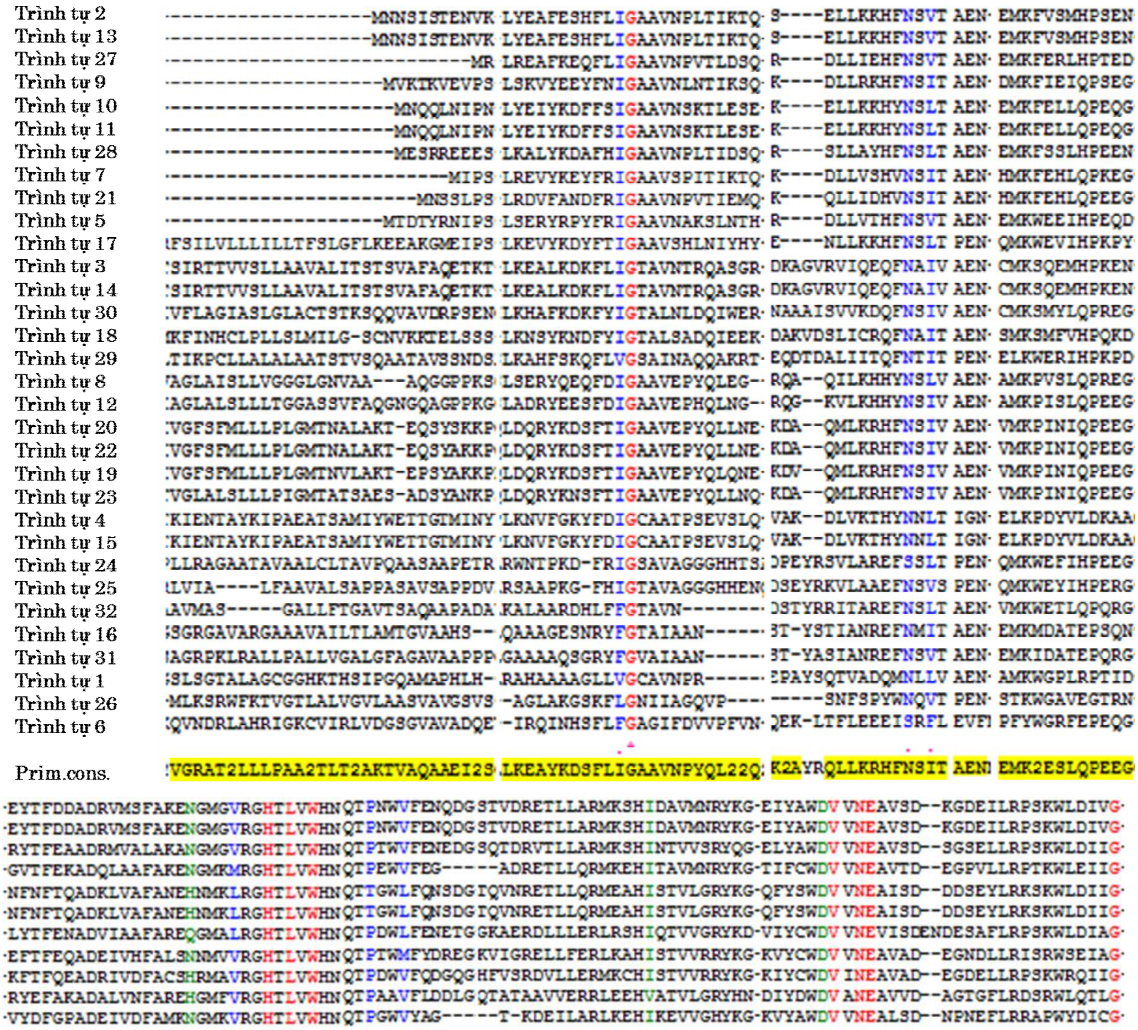
Số liệu xây dựng probe phải từ các nghiên cứu thực nghiệm, do đó, chỉ các trình tự có hoạt tính endo 1-4 xylanase đã xác định được chi tiết về khả năng biểu hiện, giá trị nhiệt độ và pH hoạt động tối ưu của enzyme mới được thu thập. Số liệu DNA metagenome chủ yếu từ vi khuẩn, nên khi tìm kiếm số liệu chúng tôi chỉ thu thập các trình tự amino acid của enzyme có nguồn gốc từ vi khuẩn để đảm bảo kết quả đáng tin cậy.

Kết quả chúng tôi tìm kiếm được ba trình tự thuộc họ GH8, 30 trình tự thuộc GH10, 18

trình tự thuộc GH11, một trình tự thuộc GH30, một trình tự thuộc GH51. Riêng họ GH5 và GH98 chúng tôi không tìm kiếm được trình tự nào thỏa mãn các yêu cầu đề ra. Tuy nhiên, do số lượng các trình tự thuộc các họ GH khác hạn chế, không đủ để xây dựng probe nên chúng tôi chỉ thống kê số liệu của họ GH10, GH11 dùng để xây dựng probe ở bảng 2. Từ dữ liệu về số các trình tự mã hóa cho enzyme endo 1-4 xylanase đã được nghiên cứu tính chất của enzyme chúng tôi sẽ tiếp tục tìm các vùng tương đồng nhau.

Xây dựng probe từ các trình tự

Kết quả phân tích sau khi sử dụng 27 trình tự thuộc GH10 và 18 trình tự thuộc GH11 bằng phần mềm ClustalW - PBIL được trình bày trên hình 1 và hình 2. Kết quả cho thấy, khi so sánh 27 trình tự amino acid thu thập được của GH10 có 16 amino acid hoàn toàn giống nhau, 13 vị trí giống nhau ở đa số các trình tự và có 14 vị trí giống nhau ở một số trình tự, còn lại là khác nhau (hình 1). Khi so sánh 18 trình tự amino acid của họ GH11 chúng tôi thấy có 54 amino acid hoàn toàn giống nhau, 25 vị trí giống nhau ở đa số các trình tự và có 24 vị trí giống nhau ở một số trình tự, còn lại là khác nhau. Kết quả này cho thấy trình mã hóa enzyme endo 1- 4 xylanase thuộc họ GH11 bảo tồn cao hơn trình tự thuộc GH10.



Hình 2. Kết quả so sánh tương đồng các trình tự amino acid của endo 1- 4 xylanase thuộc họ GH10

Trình tự Prim. Cons. được tô màu là trình tự sẽ được dùng làm probe. Mức độ bảo thủ của các gốc amino acid được đánh dấu từ dấu * đến dấu ":" dấu "." và không được đánh dấu.

Probe được xây dựng chủ yếu dựa trên các trình tự bảo tồn cao và được tô màu trên hình 1 và hình 2. Kết quả, probe của họ GH10 được xây dựng bao gồm có 338 amino acid (hình 3), trong đó chứa toàn bộ 16 amino acid hoàn toàn giống nhau, 13 vị trí giống nhau ở đa số các trình tự và còn lại là vị trí giống nhau ở một số trình tự.

Probe của họ GH11 bao gồm 204 amino acid (hình 4) trong đó chứa toàn bộ 54 amino acid hoàn toàn giống nhau 25 vị trí giống nhau ở đa số các trình tự và còn lại là vị trí giống nhau ở một số trình tự.

VGRATXLLLPAAXTLTXAKTVAQAAEIXSLKEAYKDSFLIGAAVNPYQLXXQKXAQLLKRHFNSITA
ENEMKXESLQPEEGKXNFEQADRIVAFAKKNGMAVRGHTLVWHSQTPGWFFXDEEGTVSXELLLX
RMKEHIKTVVGRYKGKIYAWDVVNEAVSDSGXGXLRKSKWLQILGEDYIAKAFEYAXEADPNXAK
LFYNDYNEEVPPAKREAIYKLVKSLKXKGVPIDGIGLQAHWNLXWPSLDEXIRAAIERFASLGLXXQI
TELDVSXFGWPDARTDLDAPTEEEMLEXQAERYDQLFQLFLXYSDKITSVTFWGVADDYTWLDDFP
VRGRKGKDWPFLFDENYQPKPAYWAXIDLANXK

Hình 3. Trình tự probe cho enzyme endo 1- 4 xylanase thuộc họ GH10. X: gốc amino acid không xác định

MFKFKKXFLXVLLAALMSIXLFAATXSAATDYWQNWTDGGGTVNAVNGSGGNYSVNWSNTGNFV
VGKGWTTGPXRTINYNAGVFAPSGNGYLTLYGWTRNPLIEYYVVDSWGTYRPTGATYKGTVTSDG
GTYDIYTTTRYNAPSIDGDTTTFTQYWSVRQSKRXTGSNATITFSNHVNAWASKGMNLGSXWSYQV
LATEGYQSSGSSNVTV

Hình 4. Trình tự probe cho enzyme endo 1- 4 xylanase thuộc họ GH11. X: gốc amino acid không xác định

Xác định giá trị ngưỡng cho việc sử dụng probe trong khai thác gen

Bảng 3. So sánh tương đồng giữa probe với các trình tự thuộc GH11

Trình tự	Điểm tối đa	Tổng điểm	Độ bao phủ	Giá trị E	Độ tương đồng
Trình tự 4	337	337	100%	1.00E-122	88%
Trình tự 5	336	336	100%	3.00E-122	88%
Trình tự 2	332	332	100%	1.00E-120	87%
Trình tự 3	328	328	100%	7.00E-119	87%
Trình tự 14	324	324	100%	2.00E-117	85%
Trình tự 12	314	314	100%	2.00E-113	82%
Trình tự 17	313	313	100%	5.00E-113	83%
Trình tự 16	312	312	97%	7.00E-113	84%
Trình tự 15	310	310	98%	1.00E-111	85%
Trình tự 9	293	293	99%	3.00E-105	74%
Trình tự 18	237	272	84%	2.00E-81	71%
Trình tự 6	233	246	100%	1.00E-79	61%
Trình tự 8	231	246	95%	3.00E-79	64%
Trình tự 1	228	244	87%	1.00E-79	71%
Trình tự 11	202	202	96%	3.00E-69	55%
Trình tự 7	183	203	95%	2.00E-60	50%
Trình tự 13	175	207	86%	9.00E-57	53%
Trình tự 10	162	184	86%	3.00E-52	51%

Để tìm giá trị ngưỡng phát hiện cho việc sử dụng probe trong khai thác gen, chúng tôi đã so sánh tương đồng giữa probe với từng trình tự đã sử dụng để xây dựng nên chúng. Kết quả (bảng 3, 4) cho thấy, để khai thác triệt để các trình tự mã hóa endo 1-4 xylanase, đối với probe GH10 điểm số tương đồng tối đa phải đạt tối thiểu 207, độ bao phủ và độ tương đồng tối thiểu 88% và 39%. Tuy nhiên, bảng 3 chỉ ra probe của GH10 không phù hợp với các trình tự số 3, 21. Vì vậy, ngưỡng được xem là tốt nhất cho việc khai thác các gen bằng probe GH10 là điểm tối đa đạt tối thiểu 200 điểm, độ bao phủ đạt trên 80%. Đối với probe GH11, điểm điểm tối đa đạt từ 162 trở lên, độ bao phủ và độ tương đồng tối thiểu 84% và 50%. Như vậy, khi sử dụng probe để khai thác gen, các trình tự có điểm tối đa cao trên các chỉ số của các trình tự trên cho từng probe sẽ được ưu tiên lựa chọn.

Bảng 4. So sánh tương đồng giữa probe với các trình tự thuộc GH10

Trình tự	Điểm tối đa	Tổng điểm	Độ bao phủ	Giá trị E	Độ tương đồng
Trình tự 22	426	426	89%	3e-153	67%
Trình tự 10	423	423	93%	5e-152	64%
Trình tự 16	415	415	92%	7e-149	63%
Trình tự 8	412	412	91%	8e-148	62%
Trình tự 7	412	412	91%	8e-148	62%
Trình tự 6	411	411	93%	2e-147	63%
Trình tự 4	410	410	90%	3e-147	63%
Trình tự 23	400	400	90%	9e-143	63%
Trình tự 18	363	363	98%	2e-127	53%
Trình tự 15	355	355	98%	2e-124	51%
Trình tự 17	355	355	98%	3e-124	52%
Trình tự 14	354	354	98%	7e-124	52%
Trình tự 5	346	346	92%	1e-120	53%
Trình tự 12	337	337	94%	6e-118	52%
Trình tự 2	335	335	91%	2e-117	53%
Trình tự 9	330	330	92%	2e-114	52%
Trình tự 11	311	311	96%	2e-107	45%
Trình tự 25	295	295	100%	4e-101	43%
Trình tự 24	290	290	92%	2e-98	46%
Trình tự 27	242	242	93%	2e-80	41%
Trình tự 1	242	242	93%	2e-80	41%
Trình tự 19	218	218	81%	1e-70	42%
Trình tự 20	209	209	89%	7e-68	40%
Trình tự 26	207	207	88%	3e-66	39%
Trình tự 21	97,1	97,1	77%	7e-26	27%
Trình tự 3	90,9	90,9	76%	3e-23	27%

Khai thác trình tự mã hóa cho endo 1- 4 xylanase bằng probe từ số liệu DNA metagenome của vi sinh vật trong ruột mối

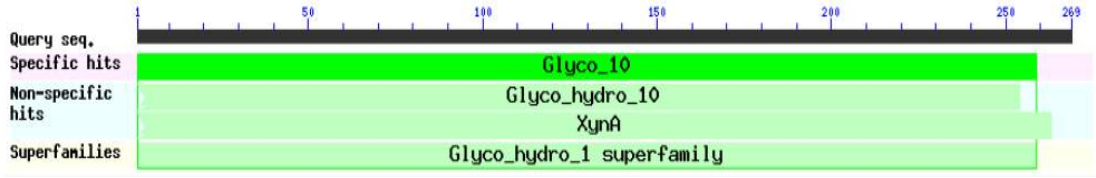
Dựa trên dữ liệu KEGG, công ty BGI đã chú giải 27 trình tự có hoạt tính endo 1- 4 xylanase (bảng 5). Khi sử dụng probe GH10 có điểm tối đa là 200, độ bao phủ và độ tương đồng tối thiểu 80%, chúng tôi chỉ lựa chọn được duy nhất một trình tự (GL0018509) (bảng 5). Sử dụng probe GH11 có điểm tối đa trên 165, độ bao phủ và độ tương đồng tối thiểu 84% và 50% chúng tôi không lựa chọn được trình tự nào từ dự đoán của BGI.

Khảo sát lại vùng bảo thủ bằng BLASTP trên các trình tự chúng tôi nhận thấy, các trình tự GL0018509 đều chứa các vùng đặc thù cho endo 1- 4 xylanase (XynA ở GH10) (hình 5).

Bảng 5. So sánh số lượng trình tự khai thác bằng probe với dự đoán của BGI

Kết quả dự đoán của BGI	Kết quả khai thác bằng Probe						
	Mã gen	Điểm tối	Tổng điểm	Độ bao	Giá trị	Độ tương	GH
GL0018509	GL001850	202	217	85%	9e-67	50%	10
GL0119674	GL011967	182	182	62%	1e-59	45%	10
GL0019299	GL001929	167	167	78%	3e-53	37%	10
GL0122083	GL012208	134	150	56%	6e-40	45%	10
GL0024062	GL002406	84,0	130	81%	3e-22	28%	10
GL0012670	GL001267	70,9	87,4	53%	4e-19	32%	10
GL0026972	GL002697	63,2	80,1	69%	1e-15	29%	10

GL0072419	GL007241	35,0	49,7	75%	1e-06	20%	10
GL0080679	GL008067	149	174	91%	2e-47	41%	11
GL0052827	GL005282	94,4	149	57%	3e-28	44%	11
GL0046968							
GL0084263							
GL0005720							
GL0024381							
GL0035995							
GL0047812							
GL0053063							
GL0054125							
GL0059925							
GL0066893							
GL0081111							
GL0085332							
GL0087399							
GL0092907							
GL0099528							
GL0100470							
GL0109190							

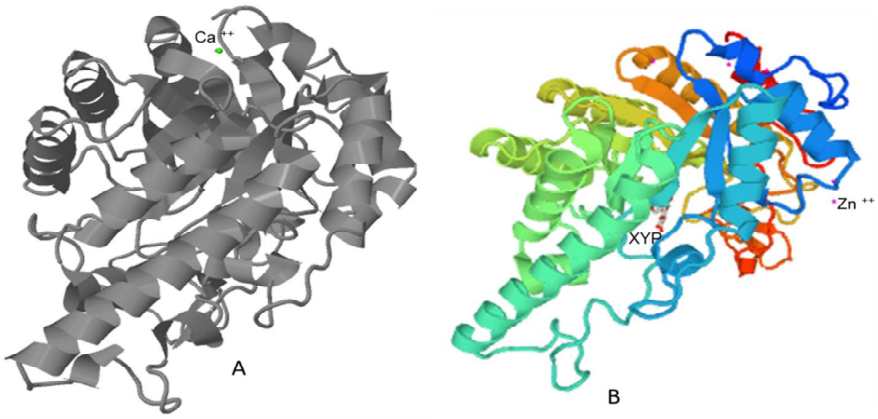


Hình 5. Kết quả dự đoán tương đồng đặc hiệu trình tự và các gốc hoạt động của trình tự mã gen GL0018509 được lựa chọn bằng probe GH10. Glyco_ hydro_10: họ GH10; XynA: xylanase

Khảo sát cấu trúc không gian của trình tự GL0018509 endo 1- 4 xylanase được khai thác bằng probe

Để chắc chắn hơn, chúng tôi tiến hành khảo sát cấu trúc không gian của enzyme mã hóa bởi gen mã số GL0018509. Cấu trúc bậc ba của phân tử cho phép ước đoán cụ thể hơn về trung tâm hoạt động, cấu hình phân tử và các phối tử liên quan đến hoạt tính sinh học của enzyme. Vì các trình tự khai thác được đều có độ tương đồng thấp với gen trên ngân hàng gen dựa trên BLASTP nên chúng tôi đã sử dụng phần mềm Swiss Prot và Phyre2 để ước đoán.

Kết quả cho thấy trình tự GL0018509 có cấu trúc tương đồng cao (95% và 93,4%) với endo 1- 4 xylanase của khuôn *2uwf_A* (hình 6A) theo ước đoán của Phyre2 và khuôn *1r87.1.A* theo Swiss Prot với độ tương đồng 93,4% và độ bao phủ 93,0% (hình 6B). Cả hai khuôn đều có cấu trúc tương tự nhau, tuy nhiên các phối tử của các khuôn có khác nhau. Vị trí liên kết với ion Ca^{2+} (Phyre2) hoặc Zn^{2+} (Swiss Prot) nằm ở vùng liên kết giữa các phân tử polymer để tạo dạng tetramer và có vùng liên kết với xylopyranose-một cơ chất đơn giản tương tự xylan. Kết quả ước đoán cấu trúc không gian hoàn toàn phù hợp với ước đoán chức năng của phân tử.



Hình 6. Cấu trúc không gian của các GL0018509 được khai thác bằng probe sử dụng Phyre2 (A) dựa trên khuôn 2uwf_A và Swiss Prot (B) dựa trên khuôn 1r87.1.A. XYP: xylopyranose

KẾT LUẬN

Dựa trên các trình tự đã được nghiên cứu kỹ về đặc điểm chức năng, hai probe dùng để khai thác endo 1-4 xylanase GH10 và GH11 đã được xây dựng. Kết quả sử dụng probe trên đã lựa chọn được một gen mã hóa endo 1-4 xylanase từ dữ liệu giải metagenome của vi khuẩn trong ruột mối. Trình tự này đã được kiểm chứng lại về chức năng bằng BlastP và cấu trúc không gian bằng hai phần mềm Phyre2 và Swiss Prot.

Lời cảm ơn: Công trình được thực hiện bằng nguồn kinh phí của Đề tài độc lập "Nghiên cứu metagenome của một số hệ sinh thái mini tiềm năng nhằm khai thác các gen mới mã hóa hệ enzyme chuyển hóa hiệu quả lignocelluloses" mã số ĐTĐLCN.15/14 và trang thiết bị của phòng Thí nghiệm trọng điểm Công nghệ gen.

TÀI LIỆU THAM KHẢO

Akama T., Kawashima A., Tanigawa K., Hayashi M., Ishido Y., Luo Y., Hata A., Fujitani N., Ishii N., Suzuki K., 2013. Comprehensive analysis of prokaryotes in environmental water using DNA microarray analysis and whole genome amplification. Pathogens, 2(4): 591-605.

Baldwin D. A., Feldman M., Alwine J. C., Robertson E. S., 2014. Metagenomic assay for identification of microbial pathogens in tumor tissues. mBio, 5(5): e01714-01714.

Cantarel B. L., Coutinho P. M., Rancurel C., Bernard T., Lombard V., Henrissat B., 2009. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucleic Acids Res., 37(Database issue): D233-D238.

Do T. H., Nguyen T. T., Nguyen T. N., Le Q. G., Nguyen C., Kimura K., Truong N. H. 2014. Mining biomass-degrading genes through Illumina-based de novo sequencing and metagenomic analysis of free-living bacteria in the gut of the lower termite Coptotermes gestroi harvested in Vietnam. J. Biosci. Bioeng., 118(6): 665-671.

He J., Yin J., Wang L., Yu B., Chen D., 2010. Functional characterisation of a recombinant xylanase from Pichia pastoris and effect of the enzyme on nutrient digestibility in weaned pigs. Br. J. Nutr., 103(10): 1507-1513.

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

Kamble R. D., Jadhav A. R., 2012. Isolation, purification, and characterization of xylanase produced by a new species of *Bacillus* in solid state fermentation. Int. J. Microbiol., 2012: e683193.

Koshland D. E., 1953. Stereochemistry and the mechanism of enzymatic reactions. Biol. Rev., 28(4): 416-436.

Kushwaha S. K., Manoharan L., Meerupati T.,

Hedlund K., Ahrén D., 2015. MetCap: a bioinformatics probe design pipeline for large-scale targeted metagenomics. BMC Bioinformatics, 16: 65.

Lombard V., Ramulu G. H., Drula E., Coutinho P. M., Henrissat B., 2014. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res., 42(D1): D490-D495.

Mitsuhashi M., Cooper A., Ogura M., Shinagawa T., Yano K., Hosokawa T., 1994. Oligonucleotide probe design - a new approach. Nature, 367(6465): 759-761.

Rawashdeh R., Saadoun I., Mahasneh A., 2005. Effect of cultural conditions on xylanase production by *Streptomyces* sp. (strain Ib 24D) and its potential to utilize tomato pomace. Afr. J. Biotechnol., 4(3): trang??.

Subramaniyan S., Prema P., 2002. Biotechnology of microbial xylanases: enzymology, molecular biology, and application. Crit. Rev. Biotechnol., 22(1): 33-64.

Wang Q., 2013. Bioprocessing technologies in biorefinery for sustainable production of fuels, chemicals, and polymers. Green Process. Synth., 2(6): 637-637.

Zhou J., He Z., Yang,Y., Deng Y., Tringe S. G., Alvarez-Cohen L., 2015. High-throughput metagenomic technologies for complex microbial community analysis: open and closed formats. mBio., 6(1): e02288-14.

PROBE DESIGN FOR MINING AND SELECTION OF GENES CODING ENDO 1-4 XYLANASE FROM DNA METAGENOME DATA

Nguyen Minh Giang[1,2], Do Thi Huyen[1], Phung Thu Nguyet[1], Truong Nam Hai[1]*

[1]Institute of Biotechnology, VAST
[2]Ho Chi Minh University of Pedagogy

SUMMARY

According to the CAZY classification, endo 1- 4 xylanase belongs to GH 5, 8, 10, 11, 30, 51, 98. However only 03 sequences of GH8, 27 sequences of GH10, 18 sequence of GH11, only one sequence of each GH30 and GH51 from CAZy and NCBI database were thouroughly experimentally studied for biological activity and characteristics of the enzyme. Through the collected sequences, two probes for endo 1-4 xylanase of GH10 and GH11 were designed, based on the sequence homology. The GH10 probe was 338 amino acids lenghth contained all the conserved amino acid residues (16 conserved residues in all sequences, 13 residues similar in almost sequences, 14 residues conserved in many sequences) with the lowest maxscore of 189, coverage of 88% and identity of 39%. The GH11 probe was 204 amino acids contained all the conserved amino acid residues (54 conserved residues were identity in all sequences, 25 residues similar in almost sequences, 24 residues conserved in many sequences) with the lowest maxscore of 165, coverage of 84% and identity of 50%. Using the two probes, we mined only one sequence (GL0018509) for endo 1-4 xylanase from metagenomic DNA data of free-living bacteria in *Coptotermes termite* gut. Prediction of three-dimention structure of GL0018509 sequence by Phyre2 and Swiss Prot showed that this sequence was high similarity (95% by Phyre2 and 93,4% by Swiss Prot) with endo 1-4 xylanase with the 100% confidence.
Keyword: Coptotermes gestroi, BLASTP, DNA metagenome, ClustalW, endo 1-4 xylanase, glycoside hydrolase (GH), probe.

Citation: Nguyen Minh Giang, Do Thi Huyen, Phung Thu Nguyet, Truong Nam Hai, 2018. Probe design for mining and selection of genes coding endo 1-4 xylanase from dna metagenome data. Tap chi Sinh hoc, 40(1): 39-50. DOI: 10.15625/0866-7160/v40n1.9200.

**Corresponding author:* tnhai@ibt.ac.vn

Received 7 Februaary 2017, accepted 20 December 2017