Tạp chí Công nghệ Sinh học **15**(3): 451-460, 2017

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

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

[1]*Trường Đại học Sư phạm Thành phố Hồ Chí Minh*
[2]*Viện Công nghệ sinh học, Viện Hàn lâm Khoa học và Công nghệ Việt Nam*

✉ Người chịu trách nhiệm liên lạc. E-mail: tnhai@ibt.ac.vn

Ngày nhận bài: 19.6.2017
Ngày nhận đăng: 15.9.2017

TÓM TẮT

Theo phân loại của CAZy, xylan 1-4 beta xylosidase thuộc họ glycoside hydrolase (GH) 1, 3, 31, 39, 43,51, 52, 54, 116, 120. Trong nghiên cứu này, probe được xây dựng dựa trên các trình tự axit amin của enzyme này từ mỗi họ GH đã được nghiên cứu trong thực nghiệm. Các trình tự thu thập để xây dựng probe đảm bảo cùng có nguồn gốc từ vi khuẩn, có các thông tin chi tiết về hoạt tính enzyme, nhiệt độ và pH hoạt động tối ưu. Kết quả các họ GH1, GH3 chỉ tìm được duy nhất 01 trình tự, GH52 tìm được 3 trình tự và GH43 tìm được 11 trình tự đáp ứng các tiêu chí thu thập. Với kết quả tìm kiếm này, chỉ có duy nhấthọ GH43 có đủ số lượng trình tự đáng tin cậy để xây dựng probe. Probe của họ GH43 có độ dài là 507 amino acid, trong đó chứa toàn bộ 7 amino acid hoàn toàn giống nhau ở các trình tự, 32 vị trí giống nhau ở đa số các trình tự và có 30 vị trí giống nhau ở một số trình tự, với độ bao phủ thấp nhất là 60% và độ tương đồng thấp nhất là 30%, điểm tối đa là 17. Sử dụng probe này đã khai thác được 25 trình tự mã hóa xylan 1-4 beta xylosidase từ dữ liệu giải trình tự DNA metagenome của vi sinh vật sống tự do trong ruột mối. So sánh với kết quả chú giải gen dựa trên dữ liệu KEGG của BGI chỉ có 20 trình tự trùng nhau. Kết quả ước đoán cấu trúc bậc ba của các trình tự cho thấy tất cả các trình tự có chỉ số điểm tối đa khi so sánh với probe bằng BLASTP cao trên 100 đều có cấu trúc giống với cấu trúc của xylosidase đã được nghiên cứu. Như vậy, probe này có thể sử dụng để khai thác gen beta - xylosidase từ dữ liệu metagenome khi sử dụng chỉ số điểm tối đa trên 100.

Từ khóa: *BLASTP, ClustalW, Coptotermes gestroi, DNA metagenome, glycoside hydrolase (GH), probe, xylan 1-4 beta xylosidase,**X**bxs14*

MỞ ĐẦU

Mối *Coptotermes gestroi* thuộc mối bậc thấp trong họ *Rhinotermitidae* rất phổ biến ở Việt Nam cũng như một số quốc gia trên thế giới. Có ít nhất 16 họ GH của vi sinh vật cộng sinh trong ruột sau, bao gồm: GH2,3, 5, 7, 10, 11, 16,20, 26, 30, 42, 45, 47, 53, 77, 92(Scharf, Tartar, 2008; Franco Cairo *et al*, 2011). Ứng dụng kỹ thuật metagenomics theo hướng phân tích dữ liệu thu được từ việc giải toàn bộ trình tự metagenome, Do và đồng tác giả (2014) đã khai thác được 125.431 khung đọc mở (ORF) với tổng chiều dài lên tới 78.271.365 bp của hệ vi khuẩn sống tự do trong ruột mối, trong đó ước đoán gồm rất nhiều trình tự mã hóa cho enzyme xylan 1-4 beta xylosidase hay gọi tắt là enzyme beta xylosidase (Do *et al.*, 2014). Phương pháp dự đoán gen từ dữ liệu khổng lồ DNA metagenome của vi sinh vật trong ruột mối bước đầu thường dựa trên số liệu trình tự tương đồng với các loài đã được công bố trong Ngân hàng gen (Simon, Daniel, 2011). Nghiên cứu này quan tâm đến việc khai thác và lựa chọn một số gen mã hóa xylan 1-4 beta xylosidase từ nguồn dữ liệu rất lớn này để đưa vào biểu hiện hiệu quả trong thực nghiệm.

Các nghiên cứu bước đầu đã lựa chọn các trình tự gen mong muốn bằng nhiều phần mềm dự đoán cấu trúc và chức năng của protein. Đầu tiên là việc sử dụng công cụ BLASTP để xác định chức năng bằng cách so sánh độ tương đồng và mức độ bao phủ của trình tự đã ước đoán chức năng ban đầu từ DNA metagenmone, với các trình tự có trên Ngân hàng gen của NCBI (https://www.ncbi.nlm.nih.gov/). Kết quả BLASTP giúp cho việc ước đoán rất nhiều các thông số của enzyme như vị trí các vùng bảo tồn, họ

enzyme, nguồn vi sinh vật chứa enzyme…. Trước khi đưa vào thực nghiệm, các gen này tiếp tục được dự đoán về tính chất như khả năng chịu kiềm/acid (Lin *et al.*, 2013) và khả năng chịu nhiệt của enzyme. Mặc dù dữ liệu trình tự nucleotide/amino acid của NCBI vô cùng lớn, nhưng có rất nhiều trình tự nucleotide mã hóa cho enzyme xylan 1-4 beta xylosidase chưa được nghiên cứu tính chất, mà chỉ dựa trên sự so sánh tương đồng đối với trình tự có sẵn trong NCBI. Do đó số liệu dự đoán của BLAST và các phần mềm khác sử dụng những dữ liệu của NCBI chưa được nghiên cứu chi tiết sẽ trở nên kém tin cậy khi đưa vào thực nghiệm. Đây là một trong những khó khăn trong việc lựa chọn gen từ số liệu DNA metagenome của ruột mối nếu chỉ sử dụng các công cụ trên. Do đó việc tìm kiếm phương pháp để có thể khai thác, lựa chọn được gen mã hóa xylan 1-4 beta xylosidase từ dữ liệu này và có thể thực nghiệm hiệu quả là rất cần thiết.

Thông qua nghiên cứu số liệu về các trình tự amino acid trong NCBI của xylan 1-4 beta xylosidase, có rất nhiều trình tự chỉ so sánh với các số liệu có sẵn có thể chưa được nghiên cứu thực nghiệm. Để đảm bảo số liệu đáng tin cậy, chỉ các trình tự amino acid của xylan 1-4 beta xylosidase đã được nghiên cứu kỹ tính chất mới được thu thập. Probe được xây dựng dựa trên việc so sánh tương đồng của nhiều trình tự và sử dụng để tìm kiếm tất cả các gen tiềm năng mã hóa cho xylan 1-4 beta xylosidase từ dữ liệu DNA metagenome. Probe sẽ là cơ sở quan trọng cho việc lựa chọn các gen đưa vào thực nghiệm với khả năng thành công cao và tiết kiệm thời gian khai thác gen từ DNA metagenome. Trong thực tế probe đã được sử dụng trong rất nhiều 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 hóa cho protein trong genome (Mitsuhashi *et al.*, 1994). Trong nghiên cứu này,phương pháp xây dựng probe cho enzyme xylan 1-4 beta xylosidase được mô tả chi tiết về quá trình thiết kế, lựa chọn và thử nghiệm.

VẬT LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

Vật liệu

Dữ liệu metgenome DNA gồm 125.431 khung đọc mở (ORF) của vi sinh vật cộng sinh trong ruột mối *C. gestroi,* được giải trình tự bằng hệ thống giải trình tự HiSeq Illuminia (BGI, Hồng Kông) và được chú giải chức năng dựa trên KEGG (Do *et al.*, 2014).

Phương pháp nghiên cứu

Xác định các họ GH có chứa enzyme xylan 1-4 beta xylosidase theo CAZY (http://www.cazy.org)

CAZy (Carbohydrate-Active enZymes) 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 (Cantarel *et al.*, 2009; Lombard *et al.*, 2014). CAZy xác định các họ thủy phân các liên kết glycoside (Glycosyl Hydrolases: GH) có liên quan tiến hóa được giới thiệu bởi Henrissat (Henrissat, 1991*;* Henrissat, Davies, 1997) và đến năm 2015 CAZy chứa 135 họ GH với các đặc điểm nhận biết khác nhau. Từ dữ liệu phân loại của CAZy các họ GH chứa enzyme này được xác định, sau đó tìm hiểu về cấu trúc không gian, các loại amino acid cho điện tử và proton trong quá trình hoạt động của enzyme. Kết quả sẽ được tổng hợp thành bảng phân loại danh sách các họ GH chứa enzyme này để tìm hiểu các thông tin sâu hơn.

Tìm kiếm các trình tự amino acid của enzyme xylan 1-4 beta xylosidase đã được nghiên cứu đặc tính

Dựa trên các thông tin phân loại của xylan 1-4 beta xylosidase trong các họ GH từ CAZy, nghiên cứu tiếp tục tiến hành tìm kiếm các trình tự amino acid có nguồn gốc từ vi khuẩn và các đặc tính như khả năng chịu nhiệt, chịu kiềm/acid, cấu trúc không gian, trung tâm hoạt động xylan 1-4 beta xylosidase đã được công bố. Hai nguồn thông tin chính để tìm kiếm trình tự amino acid của enzyme này là CAZy và NCBI.

Xác định mức độ tương đồng của các trình tự bằng ClustalW - PBIL (http://expasy.org/proteomics)

ClustalW - PBIL là phần mềm cho phép so sánh nhiều trình tự nucleotide hoặc amino acid và đưa ra bức tranh tổng thể và tính toán điểm tương đồng giữa các trình tự khác nhau. 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 đặc trưng chỉ ra các vị trí mà amino acid giống nhau hoàn toàn hoặc một phần giữa các trình tự để người dùng dễ dàng quan sát và phân tích. Sau khi nhập dữ liệu ClustalW-PBIL sẽ tính toán và trả kết quả sau vài phút tùy vào số lượng trình tự cần so sánh. Tất cả trình tự amino acid thu thập được của

enzyme xylan 1-4 beta xylosidase sẽ được so sánh với nhau cho từng họ GH bằng ClustalW - PBIL.

Xây dựng probe và giá trị tham chiếu

Việc phân tích kết quả của ClustalW - PBIL sẽ giúp cho việc xây dựng các probe cho từng nhóm GH. Dựa trên kết quả của *Prim.cons,* các vị trí giống nhau hoặc tương đối giống nhau sẽ được 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 chính các trình tự đã nghiên cứu để xác định giá trị tham chiếu về mức độ bao phủ và tương đồng của probe với trình tự đã sử dụng bằng BLASTP. Trên cơ sở giá trị tham chiếu của probe để lựa chọn các gen mã hóa cho xylan 1-4 beta xylosidase từ số liệu DNA metagenome của vi sinh vật trong ruột mối.

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

Sau khi có các probe mã hóa cho xylan 1-4 beta xylosidase thuộc các họ GH khác nhau, công cụ BLASTP được sử dụng để so sánh probe với trình tự amino acid của các ORF thuộc metagenome của vi sinh vật trong ruột mối. Kết quả của việc sử dụng probe tìm kiếm các gen mã hóa cho enzyme sẽ được so sánh với kết quả dự đoán ban đầu của BGI. Sau đó các trình tự có chỉ số điểm tối đa (max score) giá trị tương đồng và bao phủ lớn hơn hoặc bằng giá trị tham chiếu sẽ được lựa chọn.

Ước đoán cấu trúc bậc ba của trình tự mã hóa xylan 1-4 beta xylosidase được khai thác

Cấu trúc bậc ba của phân tử được ước đoán dựa trên trình tự và với các protein khung đã được nghiên cứu về cấu trúc bằng phần mềm Swiss Prot.

KẾT QUẢ VÀ THẢO LUẬN

Xác định các họ GH có chứa enzyme xylan 1-4 beta xylosidase theo CAZy

Trên cơ sở số liệu của CAZy enzyme xylan 1-4 beta xylosidase được phân vào 10 họ GH (Bảng 1).

Bảng 1. Các họ GH chứa enzyme xylan 1-4 beta xylosidase theo CAZy

Mã E.C	GH	Clan	Mô hình cấu trúc không gian	Chất nhận điện tử	Chất cho điện tử	Cơ chế xúc tác
3.2.1.37	1	GH-A	$(\beta / \alpha)_8$	Glu	Glu	Giữ nguyên
3.2.1.37	3	Chưa xác định	Chưa xác định	Asp	Glu	Giữ nguyên
3.2.1.37	31	GH-A	$(\beta / \alpha)_8$	Glu	Glu	Giữ nguyên
3.2.1.37	39	GH-A	$(\beta / \alpha)_8$	Glu	Glu	Giữ nguyên
3.2.1.37	43	GH-F	5-fold β-propeller	Asp	Glu	Đảo ngược
3.2.1.37	51	GH-A	$(\beta / \alpha)_8$	Glu	Glu	Giữ nguyên
3.2.1.37	52	Chưa xác định	Chưa xác định	Asp	Glu	Giữ nguyên
3.2.1.37	54	Chưa xác định	Chưa xác định	Chưa xác định	Chưa xác định	Giữ nguyên
3.2.1.37	116	Chưa xác định	Chưa xác định	Asp	Glu	Giữ nguyên
3.2.1.37	120	Chưa xác định	Chưa xác định	Asp	Glu	Giữ nguyên

Theo kết quả này, 05 họ GH3, GH52, GH54, GH116 và GH120 chưa được phân loại vào các nhóm lớn, do đó chưa xác định được cấu trúc không gian. Các họ GH1, GH30, GH39, GH51 đều thuộc GH-A giống nhau trong cấu trúc không gian là $(\beta/\alpha)_8$, chỉ có họ GH43 thuộc nhóm lớn GH-F với mô hình cấu trúc không gian gồm phiến β tạo thành 5 cánh (5-fold β-propeller). Enzyme thuộc các nhóm lớn ("clan") dựa trên sự bảo tồn cao về sự cuộn, gấp trong cấu trúc không gian so với trình tự amino acid của protein. Các họ GH chứa xylan 1-4 beta xylosidase đều có chất cho điện tử và proton trong quá trình hoạt động của enzyme là glutamate (Glu) trừ họ GH3, GH43 và GH52 chất cho điện tử là aspartate (Asp), riêng GH54 vẫn chưa xác định được các thông số cụ thể. Hai cơ chế phản ứng xúc tác thủy phân liên kết glycoside thường thấy nhất mà Koshland (1953) đưa ra là giữ nguyên và đảo ngược. Kết quả cho thấy chỉ có họ GH43 thuộc về cơ chế đảo ngược, còn lại đều thực hiện theo cơ chế giữ

nguyên (Koshland, 1953).

Tìm kiếm các trình tự amino acid của enzyme xylan 1-4 beta xylosidase đã được nghiên cứu đặc tính

Số liệu xây dựng probe phải từ các nghiên cứu thực nghiệm, do đó chỉ các trình tự mã hóa xylan 1-4 beta xylosidase đã xác định được chi tiết về khả năng biểu hiện, giá trị nhiệt độ và pH hoạt động tối ưu của enzyme (T_{opt}, pH_{opt}) mới được thu thập. Số liệu DNA metagenome theo ước đoán ban đầu chủ yếu từ vi khuẩn, nên khi tìm kiếm số liệu chỉ thu thập các trình tự amino acid của enzyme này từ vi khuẩn để đảm bảo độ tin cậy của kết quả cuối cùng. Hai nguồn công bố dữ liệu từ CAZy và từ NCBI được sử dụng để thu thậpvề trình tự amino acid đã được nghiên cứu và được tổng hợp chi tiết ở Bảng 2.

Sau khi tiến hành thu thập trình tự của nhóm enzyme này dựa trên các nghiên cứu công bố, chúng tôi đã tìm kiếm được mỗi họ GH1 và GH3 chỉ có 01 trình tự, họ GH52 có 03 trình tự và 11 trình tự thuộc họ GH43, các họ còn lại vẫn chưa tìm thấy công bố nào (Bảng 2). Từ dữ liệu về số các trình tự mã hóa cho enzyme xylan 1-4 beta xylosidase đã được nghiên cứu tính chất của enzyme các vùng tương đồng nhau sẽ tiếp tục tìm ra trong bước tiếp theo.

Bảng 2. Tổng hợp dữ liệu đã được nghiên cứu chi tiết về xylan 1-4 beta xylosidase

TT	Mã số trong GENBANK	Vi khuẩn	Số amino acid	pH_{opt}	T_{opt} (°C)	Nguồn thu thập số liệu
GH1						
1	ADT62000.1	gut microbiota of the termite (Reticulitermes santonensis).	464	6	40	http://www.ncbi.nlm.nih.gov/pubmed/12089151
GH3						
2	CAD48309.1	*Clostridium stercorarium*	715		50	http://www.ncbi.nlm.nih.gov/pubmed/21331632
GH43						
3	CAA29235.1	*Bacillus pumilus*	535	7	40	http://www.ncbi.nlm.nih.gov/pubmed/1765080
4	AAC97375.1	*B. pumilus* PLS	535	6	45	http://www.ncbi.nlm.nih.gov/pubmed/9114518
5	AAC27699.1	*Bacillus* sp. KK-1	533		55	http://www.ncbi.nlm.nih.gov/pubmed/9678255
6	BAA02527.1	*C. stercorarium*	473	7	65	http://www.ncbi.nlm.nih.gov/pubmed/7763495
7	BAC87941.1	*C. stercorarium*	497	3.5	80	http://www.ncbi.nlm.nih.gov/pubmed/15056894
8	AFZ78871.1	*Enterobacter* sp. nf1B6	536	6	40	http://www.ncbi.nlm.nih.gov/pubmed/23838121
9	AAT98625.1	*Geobacillus stearothermophilus* T-6 (XynB3)	535	6.5	60	http://www.ncbi.nlm.nih.gov/pubmed/15628881
10	ABC75004.1	*G. thermoleovorans* IT-08	511	5	70	http://www.academia.edu/16973683/Cloning_Sequencing_and_Characterization_of_the_Xylan_Degrading_Enzymes_from_Geobacillus_thermoleovurans_IT-08
11	ADV16404.1	*Paenibacillus woosongensis*	474	6-7-	30-45	http://www.ncbi.nlm.nih.gov/pubmed/21193828
12	AEF28829.1	*Thermobifida fusca* TM51	550	4.5	60	http://www.ncbi.nlm.nih.gov/pubmed/22893016
13	BAF98235.1	*Vibrio* sp. XY-214	535	7	35	http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2223237/
GH52						
14	BAA74507.1	*Aeromonas caviae*	729	8	60	http://www.ncbi.nlm.nih.gov/pubmed/11330658
15	AGE34479.1	*G. stearothermophilus*	705	5.5	70	http://www.ncbi.nlm.nih.gov/pubmed/24122394
16	ABI49956.1	*G. stearothermophilus*	705	6.3	65	http://www.ncbi.nlm.nih.gov/pubmed/11322943

pH_{opt}, T_{opt} là pH và nhiệt độ tối ưu cho hoạt động của enzyme

Xác định mức độ tương đồng của các trình tự bằng phần mềm ClustalW

Hình 1. So sánh sự tương đồng các trình tự amino acid của xylan 1-4 beta xylosidase thuộc họ GH43. 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.

Để có thể xây dựng được probe, yêu cầu số lượng trình tự mã hóa xylan 1-4 beta xylosidase càng nhiều thì độ tin cậy của probe sẽ càng cao. Do đó dựa trên số liệu thu thập các trình tự thuộc về các họ mỗi GH khác nhau và chỉ GH43 có số lượng trình tự đủ lớn để có thể so sánh tìm vùng tương đồng và thiết kế probe. Kết quả phân tích sau khi sử dụng 11 trình tự thuộc GH43 bằng phần mềm ClustalW - PBIL được thể hiện ở Hình 1.

Theo kết quả tính toán khi so sánh 11 trình tự amino acid thu thập được có 7 amino acid hoàn toàn giống nhau, 32 vị trí giống nhau ở đa số các trình tự và có 30 vị trí giống nhau ở một số trình tự, còn lại là khác nhau (Hình 1). Kết quả này cho thấy số lượng trình tự mã hóa enzyme xylan 1-4 beta xylosidase thuộc họ GH43 của vi khuẩn bảo tồn rất cao.

Xây dựng probe và giá trị tham chiếu

Sau khi so sánh, các vị trí giống nhau sẽ được lựa chọn để làm probe và các vị trí khác nhau quá nhiều sẽ được loại bỏ (Hình 2).

IXNPVLKGFNPDPSICRVGEDYYIAXSTFEWFPGVXIXHSKDLVNWRLXXRPLXRXSQLDMKGNPDSGGVWAPCLSYXDGKFWLIYTDVKVV
DGPWKDGHNYLVTADDIDGPWSPEPXPLNSSGFDPSLFHDDDGXKYLLNMLWGHREGHHSFGGIVXQEFSVAEKKLVGERKIIFXGTPXKLT
EAPHLYKIXGYYYLLTAEGGTRYEHAVTVARSKXIXGPYEVHPDNPILTSXHXPEXPLQKXGHASIVXTHTGEWYLAHLTGRPIPTPPGYRG
YCPLGRETAIQKLEWKDDGWPYVGGGKELEVEXPRYIXEHPXXXTYXEVDEFDDXTLNXHFQTLRIPFTEQGSLTERPGHLRLYGREKSLTS
FTQAFVARRWQHFXFTAETAFKPTFQQSAGLVNYYNTENWLQVTYDETAGGRXLVCDNLVSQPLDDIYVYLRVVDRETYKYSYSFDGEWKIP
VPLESTKLSDYIRGGFFTGAFVGLXCXYDXSGXGLPADFDYFXYEEX

Hình 2. Trình tự probe cho enzyme xylan 1-4 beta xylosidase thuộc họ GH43. X: gốc amino acid không xác định

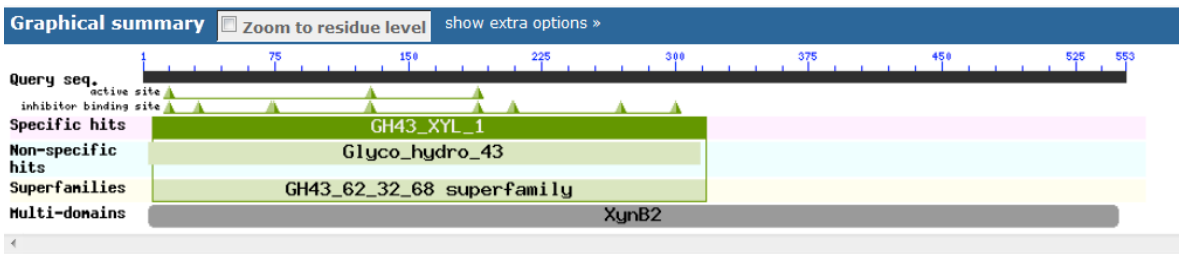
Bảng 3. So sánh tương đồng giữa probe với các trình tự thuộc GH43.

Trình tự	Điểm tối đa	Tổng điểm	Độ bao phủ	Giá trị E	Độ tương đồng
Trình tự 2	608	608	100%	0.0	72%
Trình tự 1	603	603	99%	0.0	71%
Trình tự 6	444	444	99%	5e-155	56%
Trình tự 3	596	596	99%	0.0	71%
Trình tự 8	231	248	99%	2e-73	37%
Trình tự 7	547	547	99%	0.0	66%
Trình tự 11	351	351	99%	4e-119	49%
Trình tự 10	358	377	98%	9e-122	51%
Trình tự 5	101	134	83%	3e-27	30%
Trình tự 4	25.8	76.2	61%	0.003	33%
Trình tự 9	17.3	113	60%	1.2	46%

Để tìm giá trị tham chiếu cho việc sử dụng probe trong khai thác gen, probe được so sánh tương đồng với từng trình tự đã sử dụng để xây dựngban đầu. Kết quả (Bảng 3) cho thấy để khai thác triệt để các trình tự mã hóa xylan 1-4 beta xylosidase GH43 phải có điểm tối đa trên 17, độ bao phủ và độ tương đồng tối thiểu 60% và 30% so với probe. Tuy nhiên, nhìn vào kết quả Bảng 3 cho thấy giá trị điểm tối đa, độ tương đồng và độ bao phủ của probe với các trình tự khác nhau. Probe này phù hợp nhất với các trình tự số 1, 2, 3, 6, 7, 8, 10, 11 (chỉ số điểmtối đa trên 101, độ tương đồng trên 37%) và không đại diện tốt cho các trình tự số 4, 5, 9 vì chỉ số điểm tối đa thấp và độ tương đồng thấp. Như vậy, khi sử dụng probe để khai thác gen, các trình tự có điểm tối đa cao trên 200 và độ tương đồng cao trên 37% sẽ được ưu tiên lựa chọn

Khai thác trình tự mã hóa cho xylan 1-4 beta xylosidase 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 46 trình tự có hoạt tính xylan 1-4 beta xylosidase thuộc họ GH43. Khi sử dụng probe, nếu sử dụng chỉ số điểm tối đa trên 17, độ bao phủ và độ tương đồng tối thiểu 60% và 30% sẽ có 25 trình tự được lựa chọn, trong đó có 20 trình tự trùng với dự đoán BGI (Bảng 4). Nếu dựa trên điểm tối đa trên 200 và độ tương đồng trên 37% thì chỉ 8 trình tự được khai thác là đảm bảo yêu cầu (Bảng 4). Khảo sát lại vùng bảo thủ bằng BLASTP cho thấy các trình tự đều chứa các vùng đặc thù cho xylosidase (specific hit) và có vị trí hoạt động (active site) (Hình 3).



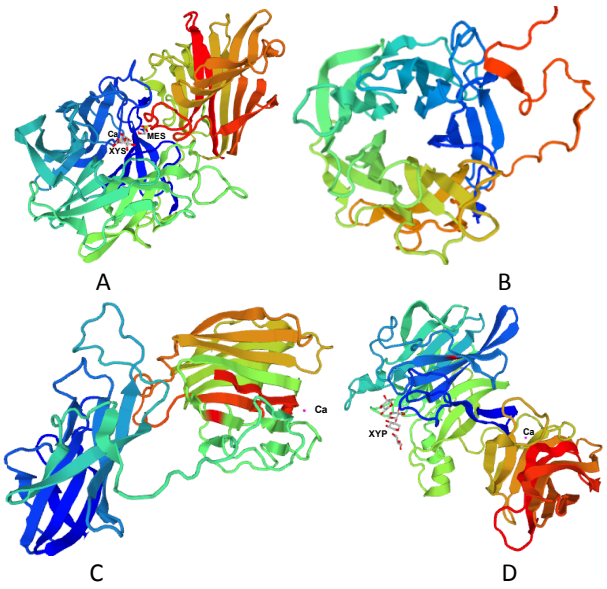
Hình 3. Dự đoán tương đồng đặc hiệu trình tự và các gốc hoạt động của các trình tự được lựa chọn bằng probe GH43. GH43_XYL: họ glycosyl 43 có khả năng phân hủy cấu trúc beta-D-xyloside; XynB2: enzyme beta-xylosidase.

Bảng 4. 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 đa	Tổng điểm	Độ bao phủ	Giá trị E	Độ tương đồng
GL0104795	GL0104795	382	382	99%	1e-130	48%
GL0020896	GL0020896	285	285	62%	3e-96	54%
GL0117445	GL0117445	285	285	62%	3e-96	54%
GL0076106	GL0076106	243	259	85%	2e-77	38%
GL0006405	GL0006405	227	271	58%	7e-75	57%
GL0044986	GL0044986	225	256	87%	7e-72	37%
GL0077826	GL0077826	216	265	67%	9e-70	48%
GL0112518	GL0112518	216	265	67%	9e-70	48%
GL0001262	GL0001262	214	266	82%	8e-67	35%
GL0095948	GL0095948	173	173	63%	3e-53	39%
GL0015489	GL0015489	161	182	67%	2e-49	42%
GL0088906	GL0088906	154	193	52%	2e-48	53%
GL0016592	GL0016592	130	160	54%	8e-40	49%
GL0021333	GL0021333	99.8	117	77%	2e-27	30%
GL0090776	GL0090776	54.7	119	53%	4e-12	27%
GL0083296	GL0083296	43.9	58.1	47%	4e-09	25%
GL0091901	GL0091901	40.4	54.3	54%	3e-08	25%
GL0079004	GL0079004	37.0	85.1	45%	5e-07	26%
GL0050001	GL0050001	35.8	112	67%	8e-07	26%
GL0106540	GL0106540	34.7	72.0	46%	2e-06	25%
	GL0119754	93.2	156	51%	3e-26	38%
	GL0039878	55.8	103	59%	2e-13	28%
	GL0122352	27.3	102	49%	4e-04	28%
	GL0068837	24.6	115	46%	0.002	25%
	GL0042431	22.7	84.3	51%	0.011	21%

Khảo sát cấu trúc bậc 3 của các trình tự xylan 1-4 beta xylosidase được khai thác bằng probe

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 do đó nghiên cứu đã sử dụng phần mềm Swiss Prot để ước đoán. Kết quả (Bảng 5) cho thấy hầu hết các trình tự đều được ước đoán có cấu trúc tương đồng cao với beta xylosidase, trong khi đó có một số trình tự có hoạt tính tương tự như xylanase và arabinfuranosidase. Các trình tự có hoạt tính khác này là các trình tự có giá trị điểm tối đa (Max score) thấp dưới 100 (Bảng 4). Các trình tự có điểm tối đa cao trên 200 đều có cấu trúc dự đoán tương đồng với enzyme beta-D-xylosidase và thường có vị trí liên kết với ion Ca^{++} nằm ở vùng liên kết giữa các phân tử polymer để tạo dạng tetramer và có vùng liên kết với xylopyranose đặc thù trên phân tử cơ chất của xylosidase, chứng tỏ chúng có hoạt tính phân cắt xylose thành đường (Hình 4). Điều này hoàn toàn phù hợp giữa ước đoán cấu trúc và chức năng của phân tử.



Hình 4. Cấu trúc bậc ba của các xylan beta xylosidase khuôn (template) 2exk.1.A (A), 1yrz.1.A (B), 1yi7.1.A (C), 3c7g.1.A (D) họ GH43 tương đồng với các trình tự được khai thác bằng probe sử dụng Swiss prot. B3P: 2-[3-[[1,3-dihydroxy-2-(hydroxymethyl)propan-2-yl]amino]propylamino]-2-(hydroxymethyl)propane-1,3-diol; XYS: xylopyranose; CA: Ca^{++}; MES: 2-morpholin-4-ium-4-ylethanesulfonate.

Bảng 5. Ước đoán cấu trúc bậc ba của các trình tự bằng Swiss Prot

Mã gen	Pfam	Khuôn	Hoạt tính	Độ bao phủ	Độ tương đồng	Phương pháp	Cấu trúc	Phối tử
GL0104795	PF04616	3c2u.1.A	Xylosidase/arabinosidase	0.95	54.77	X-ray, 1.3Å		4 x B3P
GL0117445	PF04616	2exj.1.A	beta-D-xylosidase	0.96	55.34	X-ray, 2.2Å		4 x XYS-XYS, 4 x CA, 3 x MES
GL0076106	PF04616	2exk.1.A	beta-D-xylosidase	0.94	37.55	X-ray, 2.2Å	homo-tetramer	4 x XYS-XYS, 4 x CA, 3 x MES
GL0006405	PF04616	2exk.1.A	beta-D-xylosidase	0.98	59.53	X-ray, 2.2Å		4 x XYS-XYS, 4 x CA, 3 x MES
GL0044986	PF04616	2exj.1.A	beta-D-xylosidase	0.94	40.29	X-ray, 2.2Å		4 x XYS-XYS, 4 x CA, 3 x MES
GL0077826	PF04616	1yrz.1.A	beta-D-xylosidase	0.96	57.85	X-ray, 2.0Å	monomer	Không
GL0112518	PF04616	1yrz.1.A	beta-D-xylosidase	0.93	58.86	X-ray, 2.0Å	monomer	Không
GL0001262	PF04616	2exk.1.A	beta-D-xylosidase	0.94	36.73	X-ray, 2.2Å	homo-tetramer	4 x XYS-XYS, 4 x CA, 3 x MES
GL0095948	PF04616	1yi7.1.A	beta-D-xylosidase	0.93	46.52	X-ray, 1.9Å	homo-tetramer	4 x CA
GL0015489	PF04616	2exh.1.A	beta-D-xylosidase	0.98	46.56	X-ray, 1.9Å	homo-tetramer	4 x CA, 3 x MES
GL0088906	PF04616	1yrz.1.A	beta-D-xylosidase	0.96	64.1	X-ray, 2.0Å	monomer	Không
GL0016592	PF04616	1yi7.1.A	beta-D-xylosidase	0.98	48.9	X-ray, 1.9Å	homo-tetramer	4 x CA
GL0021333	PF04616	1yrz.1.A	beta-D-xylosidase	0.96	31.65	X-ray, 2.0Å	monomer	Không
GL0090776	PF04616	2exj.1.A	beta-D-xylosidase	0.42	24.22	X-ray, 2.2Å	homo-tetramer	4 x XYS-XYS, 4 x CA, 3 x MES
GL0083296	PF04616, PF03422	3c7g.1.A	Endo-1,4-beta-xylanase	0.93	33.88	X-ray, 2.0Å	monomer	4 x XYP, 1 x CA
GL0091901	PF04616	3c7g.1.A	Endo-1,4-beta-xylanase	0.91	31.94	X-ray, 2.0Å	monomer	4 x XYP, 1 x CA
GL0079004	PF04616	1yrz.1.A	xylan beta-1,4-xylosidase	0.94	18.62	X-ray, 2.0Å	monomer	Không
GL0050001	PF04616	3qee.1.A	Beta-xylosidase/alpha-L-arabinfuranosidase	0.99	56.51	X-ray, 1.6Å	monomer	1 x CA
GL0106540	PF04616	4nov.1.A	Xylosidase/arabinofuranosidase	0.93	55.74	X-ray, 1.3Å	monomer	1 x CA
GL0119754	PF04616	2exh.1.A	beta-D-xylosidase	0.81	42.77	X-ray, 1.9Å	homo-tetramer	4 x CA, 3 x MES
GL0039878	0	1yrz.1.A	xylan beta-1,4-xylosidase	0.94	31.09	X-ray, 2.0Å	monomer	Không
GL0122352	PF04616	3kst.1.A	Endo-1,4-beta-xylanase	0.9	36	X-ray, 1.7Å	monomer	1 x CA
GL0068837	PF04616	2exk.1.A	beta-D-xylosidase	0.89	25.68	X-ray, 2.2Å	homo-tetramer	4 x XYS-XYS, 4 x CA, 3 x MES
GL0042431	PF04616	3c7f.1.A	Endo-1,4-beta-xylanase	0.82	32.08	X-ray, 1.5Å	monomer	1 x CA

Chú thích: B3P: 2-[3-[[1,3-dihydroxy-2-(hydroxymethyl)propan-2-yl]amino]propylamino]-2-(hydroxymethyl)propane-1,3-diol; XYS: xylopyranose; CA: Ca^{++}; MES: 2-morpholin-4-ium-4-ylethanesulfonate

KẾT LUẬN

Phương pháp xây dựng probe dựa trên các trình tự mã hóa enzyme xylan 1-4 beta xylosidase là một hướng mới để ứng dụng trong việc tìm kiếm trình tự đích từ dữ liệu DNA metagnone. Việc lựa chọn các trình tự amino acid của enzyme này đã được nghiên cứu chi tiết về hoạt tính từ vi khuẩn để xây dựng 01 probe thuộc họ GH43 có thể hỗ trợ hiệu quả cho việc lựa chọn các gen mã hóa cho xylan 1-4 beta xylosidase từ dữ liệu khổng lồ DNA metagenome.

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 Đề tài cơ sở "Nghiên cứu lựa chọn gen mã hóa xylan beta xylosidase từ dữ liệu giải trình tự DNA metagenome của vi sinh vật ruột mối và biểu hiện gen tạm thời trên cây thuốc lá bằng phương pháp agroinfiltration", mã số CS16-01 và trang thiết bị của Phòng thí nghiệm Trọng điểm Công nghệ gen.*

TÀI LIỆU THAM KHẢO

Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* 37: D233–238.

Do TH, Nguyen TT, Nguyen TN, Le QG, Nguyen C, Kimura K, Truong NH (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: 665–671.

Franco Cairo JPL, Leonardo FC, Alvarez TM, Ribeiro DA, Büchli F, Costa-Leonardo AM, Carazzolle MF, Costa FF, Paes Leme AF, Pereira GA, Squina FM (2011) Functional characterization and target discovery of glycoside hydrolases from the digestome of the lower termite *Coptotermes gestroi. Biotechnol Biofuels* 4: 50.

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

Henrissat B, Davies G (1997) Structural and sequence-based classification of glycoside hydrolases. *Curr Opin Struct Biol* 7: 637–644.

Koshland DE (1953) Stereochemistry and the mechanism of enzymatic reactions. *Biol Rev* 28: 416–436.

Lin H, Chen W, Ding H (2013) AcalPred: a sequence-based tool for discriminating between acidic and alkaline enzymes. *PloS One* 8: e75726.

Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* 42: D490-495.

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

Scharf ME, Tartar A (2008) Termite digestomes as sources for novel lignocellulases. *Biofuels Bioprod Biorefining* 2: 540–552.

Simon C, Daniel R (2011) Metagenomic analyses: past and future trends. *Appl Environ Microbiol* 77: 1153–1161.

PROBE DESIGN FOR EXPLOITING GENES CODING XYLAN 1-4 BETA XYLOSIDASE FROM DNA METAGENOME OF FREE – LIVING BACTERIA IN THE GUT OF THE TERMITE, *COPTOTERMES GESTROI*

Nguyen Minh Giang[1], Do Thi Huyen[2], Phung Thu Nguyet[2], Nguyen Thi Trung[2], Truong Nam Hai[3]

[1]*Ho Chi Minh University of Pedagogy*
[2]*Institute of Biotechnology, Vietnam Academy of Science and Techology*

SUMMARY

According to the CAZy classification, xylan 1-4 beta xylosidase belongs to GH 1, 3, 31, 39, 43, 51, 52, 54, 116, 120. In this study, a probe was designed from bacterial sequences exhibiting xylan 1-4 beta xylosidase activity with experimented optimal pH and temperature. As the results, only one sequence of each GH1, GH3, eleven sequences of GH43, three sequences of GH52, particularly, none of GH31, GH51, GH116, GH120 were mined. Through the collected data, one probe for xylan 1-4 beta xylosidase of GH43 was designed based on the sequence homology. This probe was 507 amino acids in length, contained all the conserved amino acid residues (7 conserved residues in all sequences, 32 similar residues in almost sequences, and 30 residues conserved in many sequences) and homologus with all the reference sequences (11 sequences) with the lowest

coverage of 60% and identity of 30% respectively, and max score of 17. Using the probe, 25 sequences for xylan 1-4 beta xylosidase from metagenomic DNA data of free-living bacteria in gut of *Coptotermes gestroi* have been mined, while 20 of the 25 mined sequences were investigated by both probe and KEGG pathway by BGI. Three-dimentional prediction of all probe-based sequences by SWISS-PROT showed that the sequences had max score with probe when compared by BLASTP over than 100 also have structures of xylan 1-4 beta xylosidase. Thus the probe can be used for mining xylan 1-4 beta xylosidase GH43 from metagenome data using max score over than 100.

Keywords: *BLASTP, ClustalW, Coptotermes gestroi, DNA metagenome, glycoside hydrolase (GH), probe, xylan 1-4 beta xylosidase*