Tạp chí Công nghệ Sinh học **14**(1): 39-47, 2016

SỬ DỤNG MỘT SỐ CÔNG CỤ TIN SINH KHAI THÁC GEN MÃ HÓA ENZYME PHÂN HỦY LIGNOCELLULOSE TỪ DỮ LIỆU METAGENOME CỦA VI SINH VẬT TRONG RUỘT MỐI *COPTOTERMES GESTROI*

Nguyễn Minh Giang[1], Đỗ Thị Huyền[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*

Ngày nhận bài: 10.10.2015
Ngày nhận đăng: 20.01.2016

TÓM TẮT

Trong nghiên cứu trước đây, chúng tôi đã thu nhận và giải trình tự DNA metagenome của khu hệ vi sinh vật ruột mối *Coptotermes gestroi* bằng máy giải trình tự thế hệ mới (Illumina) và đã nhận được dữ liệu DNA với hơn 5 Gb. Sử dụng phần mềm MGA (MetaGeneAnnotator) đã dự đoán được 125.431 khung đọc mở (ORF). Số lượng ORF có liên quan đến quá trình trao đổi carbohydrate là 8508, trong đó có 587 ORF mã hóa cho các enzyme tham gia vào quá trình thủy phân lignocellulose. Với mục đích khai thác được các trình các trình tự DNA từ dữ liệu metagenome mã hóa enzyme có khả năng chịu kiềm và đưa vào thực nghiệm thành công, chúng tôi đã tìm kiếm được một số phần mềm phù hợp để dự đoán chức năng, cấu trúc và đặc tính của enzyme với độ tin cậy cao. Alcapred để dự đoán khả năng chịu kiềm, công cụ Blastp để dự đoán vùng bảo thủ (conserved domain) của trình tự amino acid suy diễn từ ORF, công cụ Phyre2 để dự đoán cấu trúc không gian và vị trí gắn cơ chất của enzyme, công cụ của TBI để dự đoán khả năng chịu nhiệt của enzyme. Kết quả là đã khai thác được 6 ORF hoàn thiện mã hóa enzyme chịu kiềm cellulase (GL0101308, GL0038126) và hemicellulase (GL0120095, GL0074258, GL0112518, GL0067868) từ số liệu metagenome của vi sinh vật ruột mối *C gestroi*. Các ORF được lựa chọn từ kết quả của Blastp đều được dự đoán có độ bao phủ từ 90% trở nên và hệ số tương đồng từ thấp (44%) đến cao (99%), chứa vùng bảo tồn và vị trí gắn của enzyme vào cơ chất. Tỷ lệ tương đồng cấu trúc bậc hai của cellulase và hemicellulase với các protein đã được công bố khi dự đoán bằng Phyre2 tương tự như kết quả dự đoán của Blastp, với độ tin cậy từ 98% đến 100%. Trong 6 enzyme lựa chọn có 2 enzyme được dự đoán có khả năng chịu nhiệt trên 65℃, 3 enzyme chịu nhiệt từ 55℃~65℃ và chỉ có một enzyme chịu nhiệt dưới 55℃.

Từ khóa: *Cellulase, Coptotermes gestroi, hemicellulase, lignocellulose, metagenomic, metagenome, tin sinh học*

LỜI MỞ ĐẦU

Trong tự nhiên lignocellulose chủ yếu được phân hủy bởi các enzyme của vi sinh vật. Việc tìm kiếm mô hình phân giải lignocellulose của tự nhiên để giúp khai thác và ứng dụng hiệu quả các nguồn enzyme vào trong sản xuất. Trong các loài sinh vật thì mối đóng vai trò sinh thái quan trọng phân giải lignocellulose nhờ sự hỗ trợ tích cực của nhóm vi sinh vật trong đường tiêu hóa. Nhóm vi sinh vật này có khả năng tiết ra các enzyme thủy phân hoàn toàn lignocellulose. Do đó, hệ vi sinh vật ruột mối được coi là nguồn dự trữ phong phú và đa dạng các enzyme tham gia vào phân hủy lignocellulose (Scharf, Tartar, 2008).

Mối *C gestroi* thuộc mối bậc thấp trong họ *Rhinotermitidae* rất phổ biến ở Việt Nam cũng như một số quốc gia trên thế giới. Loài mối này được xem là đối tượng gây hại rất lớn cho các công trình bằng gỗ do khả năng sử dụng gỗ làm thức ăn nhờ hệ vi sinh vật cộng sinh phong phú trong đường tiêu hóa (Nimchua *et al*., 2012). Các nghiên cứu đã chỉ ra trong ruột mối có khoảng 10^6 đến 10^8 tế bào nhân sơ chủ yếu là vi khuẩn (90%). Việc tiêu hóa lignocellulose ở mối là sự cộng tác chặt chẽ giữa các enzyme của mối và vi sinh vật cộng sinh trong ruột mối tiết ra. Người ta đã chứng minh được các enzyme lignases, β-glucosidases (GH1), endoglucanases (GH9), và β-xylosidases (GH43) có trong tuyến nước bọt và ruột trước của mối; các enzyme feruloyl nằm chủ yếu ở ruột giữa, phong phú nhất là các enzyme nằm ở ruột sau. Có ít nhất 16 họ GHF 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 (João Paulo *et al.*, 2011; Scharf, Tartar, 2008). Ứng dụng kỹ thuật metagenomics theo hướng

phân tích dữ liệu thu được từ việc giải toàn bộ trình tự metagenome của hệ vi sinh vật cộng sinh trong ruột mối, hy vọng có thể khai thác được các enzyme thủy phân lignocellulose ứng dụng hiệu quả trong thực tiễn.

Việc ứng dụng metagenomics kết hợp với kỹ thuật giải trình tự gen thế hệ mới trong khai thác nguồn gen đã tạo ra dữ liệu khổng lồ về DNA. Để khai thác hiệu quả các dữ liệu này cần có những công cụ tin sinh học chuyên biệt dùng trong dự đoán chức năng gen, protein và dự đoán cấu trúc protein. Hiện nay trên mạng đang có rất nhiều phần mềm dùng cho dự đoán cấu trúc, chức năng và đặc tính của các protein phục vụ cho nghiên cứu cơ bản và nghiên cứu ứng dụng. Tuy nhiên, việc lựa chọn và sử dụng các công cụ tin sinh phù hợp với mục đích nghiên cứu cụ thể là rất cần thiết để phân tích dữ liệu metagenome.

Trong nghiên cứu trước đây, chúng tôi đã thu nhận và giải trình tự DNA metagenome của khu hệ vi sinh vật ruột mối bằng máy giải trình tự thế hệ mới (Illuminia) và đã nhận được dữ liệu DNA với hơn 5 Gb (Do TH *et al.*, 2014). Trong nghiên cứu này chúng tôi sử dụng các công cụ tin sinh học khác nhau để dự đoán các chức năng của nhóm enzyme thủy phân lignocellulose từ dữ liệu DNA metagenome nhận được. Đây là nhóm enzyme đang rất được quan tâm trong việc xử lý các sản phẩm phế thải có nguồn gốc từ thực vật, giải quyết các vấn đề về môi trường và sản xuất nhiên liệu sinh học.

VẬT LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

Vật liệu nghiên cứu

Dữ liệu metgenome DNA có kích thước 5,4 Mb của vi sinh vật cộng sinh trong ruột mối *C gestroi,* được giải trình tự bằng hệ thống giải trình tự HiSeqIllumina (Illumina, San Diego, Hoa Kỳ). Từ 5,4 Mb dữ liệu đã khai thác được 125.431 khung đọc mở (ORF) với tổng chiều dài lên tới 78.271.365 bp. Sau đó sử dụng công cụ BLASTall, các ORF này đã được so sánh với: dữ liệu eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) và sắp xếp gen vào các nhóm chức năng và dữ liệu KEGG (Kyoto Encyclopedia of Genes and Genomes) để phân loại gen vào các con đường chuyển hóa khác nhau.

Phương pháp nghiên cứu

Dự đoán các ORF bằng phần mềm MGA (MetaGeneAnnotator)

Từ 5,4 Mb dữ liệu đã khai thác được 125.431 khung đọc mở (ORF) với tổng chiều dài lên tới 78.271.365 bp. Sau đó sử dụng công cụ BLASTall, các ORF này được so sánh với: dữ liệu eggNOG (evolutionary genealogy of genes: Non-supervised Orthologous Groups) để sắp xếp gen vào các nhóm chức năng và dữ liệu KEGG (Kyoto Encyclopedia of Genes and Genomes) để phân loại gen vào các con đường chuyển hóa khác nhau. Sau khi dự đoán được chức năng và con đường chuyển hóa của các ORF, chúng tôi tiến hành tìm kiếm các công cụ tin sinh học hỗ trợ cho việc khai thác các ORF mã hóa các enzyme chịu kiềm có khả năng thủy phân sinh khối thực vật và dự đoán các thông số liên quan trước khi thực nghiệm. Các phần mềm đã sử dụng đều được cung cấp địa chỉ và khi nhập dữ liệu vào phần mềm sẽ chạy với các thông số mặc định.

Dự đoán khả năng chịu kiềm/acid (http://lin.uestc.edu.cn/server/AcalPred)

Với mục đích tìm ra các enzyme có khả năng chịu được môi trường kiềm nên chúng tôi đã tìm được phần mềm để dự đoán là AcalPred. Đây là hệ thống phân tích miễn phí được phát triển để phân biệt giữa các enzyme khả năng chịu được môi trường axit hay môi trường kiềm. Phần mềm này dựa trên các thông tin theo thứ tự tổ hợp nhiều chỉ số khác nhau của các protein đã nghiên cứu thực nghiệm bao gồm: thành phần các amino acid, chỉ số GO, nhóm các amino acid được bảo tồn, giá trị của điện tích,… Các chỉ số này sẽ là cơ sở để thiết kế vector SVM (support Vector Machine) làm chỉ số tham chiếu với mẫu phân tích (Fan *et al.*, 2013; Lin *et al.*, 2013). Khi ta có một trình tự các amino acid nhập vào trong phần mềm sẽ tự động tính toán và cho ra kết quả cuối cùng về khả năng chịu kiềm/axit của protein sau vài phút.

Dự đoán chức năng của các ORF bằng BLAST (http://blast.ncbi.nlm.nih.gov/Blast)

BLASTall (Basic Local Alignment Search Tool) là một tập hợp các chương trình tìm kiếm được thiết kế để khám phá tất cả các cơ sở dữ liệu trình tự protein và DNA có sẵn. BLASTall có rất nhiều loại tìm kiếm khác nhau phục vụ cho nhiều mục đích nghiên cứu (Madden, 2013). Trong nghiên cứu này chúng tôi quan tâm đến BLASTp để tìm kiếm tất cả các trình tự protein tương đồng với trình tự protein cần phân tích trong cơ sở dữ

liệu protein. Kết quả Blastp có thể xác định được chức năng, nguồn gốc và đặc biệt là tính mới của gen (thông qua hệ số tương đồng tối đa). Để chạy, BLASTp chúng tôi cung cấp một chuỗi amino axit đang quan tâm (chuỗi truy vấn) và so sánh với cơ sở dữ liệu của NCBI. BLASTp sẽ tìm kiếm các chuỗi con trong chuỗi truy vấn mà giống với các chuỗi con trong cơ sở dữ liệu, sau đó sẽ cho ra kết quả sau một thời gian ngắn. Trong nghiên cứu này, chúng tôi quan tâm đến mức độ tương đồng, độ bao phủ, vùng bảo tồn, vùng xúc tác của chuỗi đích so với các trình tự có sẵn trong NCBI.

Dự đoán cấu trúc không gian và vị trí gắn cơ chất của enzyme *bằng phần mềm Phyre2 (www.sbg.bio.ic.ac.uk/phyre2)*

Phyre2 là phần mềm dựa trên các nguyên tắc tương đồng ở các vùng bảo tồn cao của protein, cho phép dự đoán cấu trúc, chức năng, phân loại, tiến hóa,.. và giải quyết các cấu trúc tinh thể protein. Các trình tự axit amin của một protein sẽ được xử lý bằng cách quét đối với cơ sở dữ liệu các trình tự protein và tìm sự tương đồng trong các vùng cấu trúc, từ đó xuất ra kết quả (Kelley *et al.*, 2015). Để dự đoán cấu trúc bậc cao của protein người dùng sẽ nhập trình tự chuỗi amino acid, và chờ từ 30 phút đến vài giờ (tùy thuộc vào các yếu tố như chiều dài chuỗi, số lượng trình tự tương đồng và tần số và độ dài của chèn và xóa) phần mềm sẽ đưa ra một dự báo để hoàn thành. Thông tin tóm tắt của kết quả dự báo sẽ chuyển đến email đăng ký. Bảng kết quả chính trong Phyre2 cung cấp mức độ tin cậy của ước tính, hình ảnh và các liên kết đến các mô hình ba chiều dự báo và thông tin thu được từ một trong hai cấu trúc theo cơ sở dữ liệu Protein (Scop) hoặc Protein Ngân hàng dữ liệu (PDB) tùy thuộc vào nguồn gốc của các mẫu phát hiện.

Dự đoán khả năng chịu nhiệt của enzyme (www.tbi.org.tw/tools/)

Dựa trên thành phần và trình tự của các amino acid, liên kết hydrogen, liên kết Van der waals, tương tác kỵ nước và đặc điểm của các enzyme từ các sinh vật sống trong điều kiện môi trường có nhiệt độ cao (Ebrahimi *et al.*, 2011). Phần mềm của TBI (www.tbi.org.tw/tools) xây dựng trên số liệu của 150.000 protein chịu nhiệt độ khác nhau trong ngân hàng NCBI để dự đoán khả năng chịu nhiệt, dựa trên nguyên tắc tương đồng. Khả năng chịu nhiệt được dự đoán ở ba mức là trên 65^{o}C, 55 - 65^{o}C và dưới 55^{o}C. Để dự đoán khả năng chịu nhiệt của enzyme quan tâm, chỉ cần nhập số liệu trình tự amino acid và phần mềm sẽ trả kết quả sau vài phút.

KẾT QUẢ VÀ THẢO LUẬN

Thống kê số liệu trình tự DNA mã hóa enzyme, protein từ metagenome tham gia vào chuyển hóa lignocellulose

Trong công bố trước đây (Do *et al.*, 2014), toàn bộ sinh vật trong ruột mối được thu nhận và tiến hành tách chiết DNA và giải trình tự thu được bộ dữ liệu metagenome.

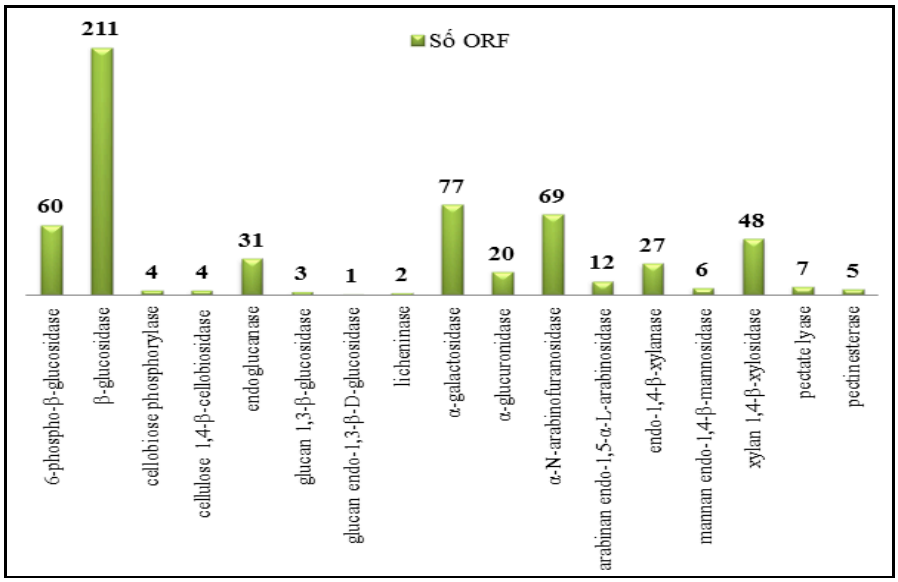
Bằng phần mềm MGA (MetaGeneAnnotator), 125.431 khung đọc mở (ORF) đã được khai thác với tổng chiều dài là 78.271.365 bp. Kích thước trung bình mỗi ORF là 624,02 bp. Trong số đó, số lượng các ORF hoàn chỉnh là 37.545 (chiếm 29.9%); còn số ORF mất một hoặc hai đầu 3’ hoặc/và 5’ là 87.886 (chiếm 70,1%). Khoảng 85.443 ORF đã được phân loại vào các nhóm chức năng (68,12%) và tham gia vào các con đường chuyển hóa (55,51%). Số lượng ORF có liên quan đến quá trình trao đổi carbohydrate là 8508, trong đó có 587 ORF (6,9%) mã hóa cho các enzyme tham gia vào quá trình thủy phân lignocellulose. Các ORF này mã hóa cho 17 nhóm enzyme tham gia vào các quá trình tiền xử lý (2), hemicellulase (7) và cellulase (8) (Hình 1).

Trong 587 ORF mã hóa cho các enzyme tham gia vào quá trình thủy phân lignocellulose được khai thác từ dữ liệu metagenome hệ vi khuẩn trong ruột mối *C. gestroi* thì chỉ có 99 ORF (16,87%) là chứa trọn vẹn gen (hoàn thiện), bao gồm 55 ORF mã hóa cho 4 nhóm cellulase và 44 ORF thuộc 6 nhóm hemicellulase, còn lại 488 ORF là không hoàn thiện**.** Trong phân tích số liệu các gen mã hóa lignocellulase, chúng tôi ưu tiên lựa chọn các ORF hoàn chỉnh đã được dự đoán là có cả đầu 3’ – 5’ và có vùng bám của ribosome trong quá trình dịch mã, để thuận lợi cho các nghiên cứu thực nghiệm biểu hiện gen sau này (Bảng 1).

Công cụ tin sinh hỗ trợ khai thác enzyme phân giải lignocellulose

Dự đoán khả năng chịu kiềm/acid (http://lin.uestc.edu.cn/server/AcalPred)

Trong nghiên cứu với mục đích tìm ra các enzyme có khả năng chịu được môi trường kiềm nên chúng tôi đã tìm được phần mềm để dự đoán là AcalPred. Khi ta có một trình tự các amino acid nhập vào trong phần mềm sẽ tự động tính toán và cho ra kết quả cuối cùng về khả năng chịu kiềm hay acid của protein.



Hình 1. Các ORF mã hóa enzyme lignocellulase hệ vi khuẩn ruột mối *C. gestroi.*

Bảng 1. ORF mã hóa enzyme thủy phân lignocellulose từ metagenome của vi sinh vật ruột mối *C gestroi.*

Enzyme	Mã E.C	ORF hoàn thiện	ORF mất đầu 5'	ORF mất đầu 3'	ORF mất 2 đầu	Tổng
Cellulase						
6-phospho-β-glucosidase	3.2.1.86	19	20	15	6	60
β-glucosidase	3.2.1.21	29	40	55	87	211
cellobiose phosphorylase	2.4.1.20	0	1	1	2	4
cellulose 1,4-β-cellobiosidase	3.2.1.91	0	1	0	3	4
Endoglucanase	3.2.1.4	5	8	6	12	31
glucan 1,3-β-glucosidase	3.2.1.58	2		1		3
glucan endo-1,3-β-D-glucosidase	3.2.1.39	0	1	0	0	1
Licheninase	3.2.1.73	0	0	2	0	2
Hemicellulase						
α-galactosidase	3.2.1.22	9	20	22	26	77
α-glucuronidase	3.2.1.13	3	4	5	8	20
α-N-arabinofuranosidase	3.2.1.55	15	11	14	29	69
arabinan endo-1,5-α-L-arabinosidase	3.2.1.99	0	2	3	7	12
endo-1,4-β-xylanase	3.2.1.8	3	9	5	10	27
mannan endo-1,4-β-mannosidase	3.2.1.78	1	3	1	1	6
xylan 1,4-β-xylosidase	3.2.1.37	12	9	16	11	48
Enzyme tiền xử lý						
pectate lyase	4.2.2.2	0	0	0	7	7
Pectinesterase	3.1.1.11	1	2	1	1	5
Tổng		**99**	**131**	**147**	**210**	**587**

Bảng 2. ORF hoàn thiện mã hóa cellulase và hemicellulase chịu kiềm.

STT	Mã gen	Enzyme	Chỉ số chịu acid	Chỉ số chịu kiềm
		Cellulase		
1	GL0036080	alkaline enzyme	0.329652	0.670348
2	GL0055814	alkaline enzyme	0.135497	0.864503
3	GL0062475	alkaline enzyme	0.405365	0.594635
4	GL0068982	alkaline enzyme	0.391706	0.608294
5	GL0089695	alkaline enzyme	0.277601	0.722399
6	GL0101308	alkaline enzyme	0.102305	0.897695
7	GL0023880	alkaline enzyme	0.024886	0.975114
8	GL0038126	alkaline enzyme	0.006766	0.993234
9	GL0057320	alkaline enzyme	0.169742	0.830258
10	GL0071848	alkaline enzyme	0.014432	0.985568
11	GL0071911	alkaline enzyme	0.010604	0.989396
12	GL0085197	alkaline enzyme	0.024500	0.9755
13	GL0109414	alkaline enzyme	0.004207	0.995793
14	GL0105118	alkaline enzyme	0.006272	0.993728
15	GL0029085	alkaline enzyme	0.006841	0.993159
16	GL0066724	alkaline enzyme	0.479200	0.5208
17	GL0003694	alkaline enzyme	0.163982	0.836018
18	GL0079178	alkaline enzyme	0.125535	0.874465
19	GL0095893	alkaline enzyme	0.262734	0.737266
20	GL0113116	alkaline enzyme	0.004207	0.995793
21	GL0033071	alkaline enzyme	0.117700	0.8823
Hemicellulase				
22	GL0050278	alkaline enzyme	0.479788	0.520212
23	GL0125198	alkaline enzyme	0.150251	0.849749
24	GL0120095	alkaline enzyme	0.050839	0.949161
25	GL0070950	alkaline enzyme	0.036672	0.963328
26	GL0080470	alkaline enzyme	0.044444	0.955556
27	GL0074258	alkaline enzyme	0.022364	0.977636
28	GL0076016	alkaline enzyme	0.324861	0.675139
29	GL0079057	alkaline enzyme	0.402239	0.597761
30	GL0107923	alkaline enzyme	0.256011	0.743989
31	GL0021085	alkaline enzyme	0.423751	0.576249
32	GL0024829	alkaline enzyme	0.24851	0.75149
33	GL0028245	alkaline enzyme	0.008125	0.991875
34	GL0072752	alkaline enzyme	0.341394	0.658606
35	GL0074257	alkaline enzyme	0.050608	0.949392
36	GL0075126	alkaline enzyme	0.458958	0.541042
37	GL0075711	alkaline enzyme	0.125509	0.874491
38	GL0024062	alkaline enzyme	0.114523	0.885477
39	GL0076106	alkaline enzyme	0.128208	0.871792
40	GL0112518	alkaline enzyme	0.015478	0.984522
41	GL0067868	alkaline enzyme	0.215684	0.784316

Kết quả dự đoán khả năng chịu kiềm của các ORF hoàn chỉnh mã hóa enzyme thủy phân cellulose và hemicellulose như bảng 2. Theo dự đoán của công cụ này, nếu chỉ số dự đoán cao hơn 0,5 đến 1 thì enzyme đó có khả năng chịu kiềm, còn thấp hơn 0,5 thì enzyme đó có khả năng chịu được acid. Kết quả dự đoán với ORF hoàn thiện mã hóa cellulase chịu kiềm là 21 và hemicellulase chịu kiềm là 20.

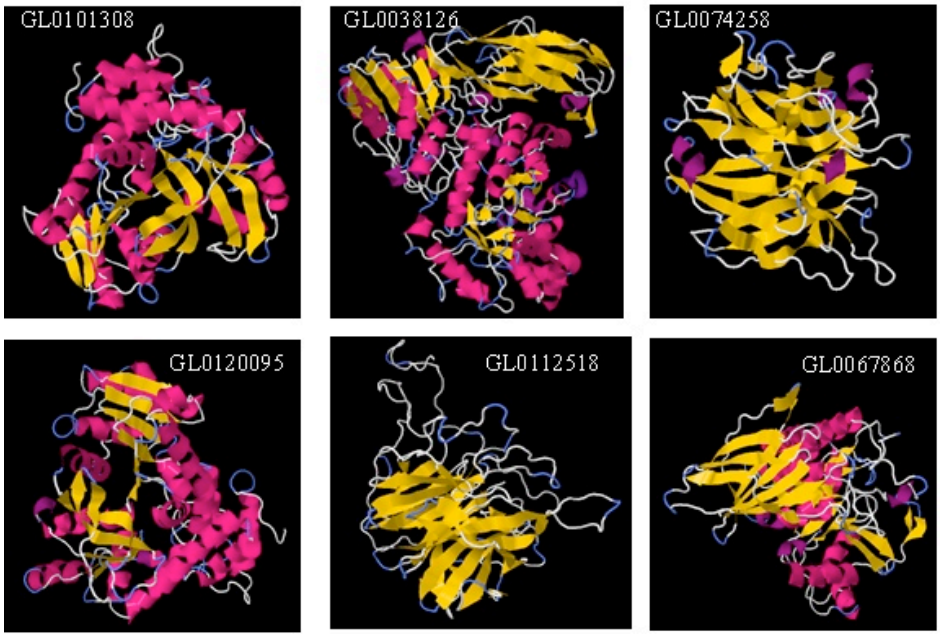
Trong 99 ORF hoàn chỉnh có 41 ORF mã hóa enzyme chịu kiềm với chỉ số dự đoán từ 0,52 đến 0,98. Khả năng chịu kiềm của các enzyme này khá cao phù hợp với các nghiên cứu trước đây về môi trường ruột mối có thể đạt đến giá trị pH là 12 (Brune *et al.*, 1995; Nimchua *et al.*, 2012).

Bảng 3. Kết quả Blast ORF hoàn thiện mã hóa cellulase và hemicellulase chịu kiềm.

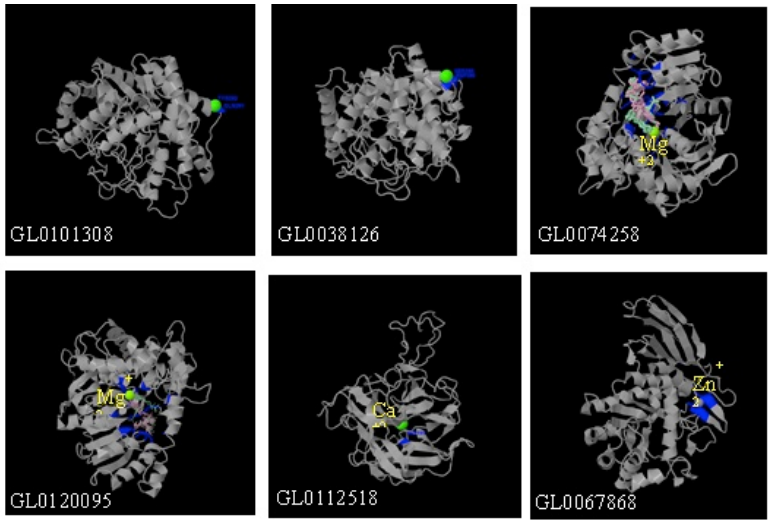
STT	Mã gen	Enzyme	Số axit amin	Độ bao phủ (%)	Độ tương đồng (%)
		Cellulase			
1	GL0036080	6-phospho-beta-glucosidase	578	100	90
2	GL0055814	6-phospho-beta-glucosidase	474	99	77
3	GL0062475	6-phospho-beta-glucosidase	477	100	99
4	GL0068982	6-phospho-beta-glucosidase	484	100	77
5	GL0089695	6-phospho-beta-glucosidase	484	100	99
6	GL0101308	6-phospho-beta-glucosidase	471	100	71
7	GL0023880	aryl-phospho-beta-D-glucosidase	478	100	99
8	GL0038126	beta-glucosidase	846	98	44
9	GL0057320	beta-glucosidase	695	100	85
10	GL0071848	beta-glucosidase	478	100	99
11	GL0071911	beta-glucosidase	132	72	49
12	GL0085197	beta-glucosidase	762	95	58
13	GL0109414	beta-glucosidase	883	95	58
14	GL0105118	beta-D-glucoside glucohydrolase	763	100	97
15	GL0029085	Xylosidase	763	94	45
16	GL0066724	xylosidase	762	95	48
17	GL0003694	beta-xylosidase	761	93	44
18	GL0079178	endo-1,4-D-glucanase	353	99	47
19	GL0095893	Cellulase	362	100	90
20	GL0113116	cellulase	540	100	85
21	GL0033071	Cellulase	362	100	90
		Hemicellulase			
22	GL0050278	alpha-galactosidase	742	99	63
23	GL0125198	alpha-galactosidase	685	99	55
24	GL0120095	alpha-galactosidase	446	99	81
25	GL0070950	alpha-glucuronidase	672	99	52
26	GL0080470	alpha-glucuronidase	256	99	28
27	GL0074258	alpha-N-arabinofuranosidase	315	98	67
28	GL0107923	alpha-N-arabinofuranosidase	690	100	98
29	GL0021085	alpha-N-arabinofuranosidase	730	42	72
30	GL0024829	alpha-N-arabinofuranosidase	245	97	49
31	GL0028245	alpha-N-arabinofuranosidase	405	99	67
32	GL0072752	Alpha-L-arabinofuranosidase	477	100	55
33	GL0074257	alpha-N-arabinofuranosidase	508	99	69
34	GL0075126	alpha-N-arabinofuranosidase	485	98	67
35	GL0075711	alpha-N-arabinofuranosidase	519	95	52
36	GL0076106	Beta-1,4-xylosidase	553	100	97
37	GL0076016	Beta-1,4-xylosidase	553	100	97
38	GL0112518	xylan 1,4-beta-xylosidase	358	90	68
39	GL0071848	xylan 1 4-beta-xylosidase	619	94	57
40	GL0024062	Beta-1,4-beta-xylanase	455	82	41
41	GL0067868	Endo1,4 – xylanase	560	99	47

Dự đoán chức năng của các ORF bằng BLAST (http://blast.ncbi.nlm.nih.gov/Blast)

Sau khi lựa chọn các gen mã hóa enzyme dự đoán có khả năng chịu kiềm, chức năng và tính mới được dự đoán bằng Blastp. Trình tự amino acid của các cellulase và hemicellulase được mã hóa bởi các ORF hoàn thiện chịu kiềm đã được sử dụng để tìm kiếm các trình tự protein tương đồng trên ngân hàng gen NCBI, qua đó có thể xác định được chức năng, nguồn gốc và đặc biệt là tính mới của gen (thông qua hệ số tương đồng tối đa). Trong số ORF hoàn thiện chịu kiềm mã hóa cho cellulase là 21 và hemicellulase là 20, có 10 ORF thuộc nhóm cellulase có độ tương đồng cao từ 85% – 100% so với các protein đã công bố trên ngân hàng gen, còn lại có độ tương đồng từ 44%-77%; Chỉ có 4 ORF mã hóa cho hemicellulase có độ tương đồng từ 80% - 100 %, các ORF còn lại có độ tương đồng khá thấp với các gen đã biết. Kết quả chi tiết trong bảng 3.



Hình 2. Dự đoán cấu trúc bậc hai của cellulase và hemicellulase bằng Phyre 2.



Hình 3. Dự đoán vị trí gắn cơ chất của cellulase và hemicellulase bằng Phyre2.

Dựa trên kết quả dự đoán khả năng chịu kiềm và chức năng của các ORF hoàn chỉnh, chúng tôi lựa chọn 2 ORF mã hóa cellulase (GL0101308, GL0038126) và 4 ORF mã hóa hemicellulase (GL0074258, GL0067868, GL0120095, GL0112518). Các ORF được lựa chọn từ kết quả của Blastp đều được dự đoán có độ bao phủ từ 90% trở nên và hệ số tương đồng từ thấp (44%) đến cao (99%), chứa vùng bảo tồn và vị trí gắn của enzyme vào cơ chất.

Dự đoán cấu trúc không gian và vị trí gắn cơ chất của enzyme *bằng phần mềm Phyre2 (www.sbg.bio.ic.ac.uk/phyre2).*

Phần mềm Phyre2 đã dự đoán cấu trúc bậc hai của các ORF mà chúng tôi lựa chọn với sự sắp xếp của cấu trúc xoắn α và gấp nếp β khác nhau. Kết quả cho thấy tỷ lệ tương đồng cấu trúc bậc hai của cellulase và hemicellulase với các protein đã được công bố tương tự như kết quả dự đoán của Blastp, với độ tin cậy từ 98% đến 100% (Hình 2).

Kết quả này giúp chúng tôi khẳng định lại việc lựa chọn các ORF có nhiều khả năng đúng với các dự đoán chức năng ban đầu và có khả năng thực nghiệm thành công. Tiếp tục tìm hiểu kỹ hơn về cách thức và khả năng hoạt động của các enzyme này, dữ liệu về cấu trúc bậc hai tiếp tục được gửi đến Phyre 2 để dự đoán vị trí trung tâm hoạt động của các enzyme, thành phần ion và các amino acid nằm trong trung tâm hoạt động. Kết quả phân tích các enzyme cellulase và hemicellulase như trong hình 3.

Trong kết quả này cho phép mô tả cấu trúc không gian ba chiều của enzyme, đặc biệt là vị trí gắn cơ chất của 6 ORF mã hóa enzyme cellulase và hemicellulase đều có các ion kim loại như Mg^{2+}, Ca^{2+}, Zn^{2+}… Bên cạnh đó mô hình dự đoán cũng chỉ ra các loại amino acid được bảo tồn cao trong trung tâm hoạt động của enzyme.

Dự đoán khả năng chịu nhiệt của enzyme (www.tbi.org.tw/tools/)

Khi đưa trình tự amino acid của các enzyme cellulase và hemicellulase đã lựa chọn vào phần mềm để dự đoán khả năng chịu nhiệt thu được kết quả như bảng 4. Theo kết quả dự đoán nếu giá trị Tm lớn hơn 1 thì enzyme đó có khả năng chịu nhiệt độ cao hơn 65℃, còn Tm nằm trong khoảng từ 0 đến 1, thì khả năng chịu nhiệt từ 55℃~65℃, và Tm nhỏ hơn 0, thì khả năng chịu nhiệt dưới 55℃. Như vậy trong 6 enzyme lựa chọn có 2 enzyme được dự đoán có khả năng chịu nhiệt trên 65℃, 3 enzyme chịu nhiệt từ 55℃~65℃ và chỉ có một enzyme chịu nhiệt dưới 55℃. Kết quả này sẽ giúp chúng tôi khi thực nghiệm có thể chọn mức nhiệt độ thích hợp theo kết quả dự đoán để kiểm tra hoạt tính enzyme.

Bảng 4. Kết quả dự đoán khả năng chịu nhiệt của cellulase và hemicellulase.

STT	Mã gen	Tm	Nhiệt độ
1	GL0101308	1.2242989647072	> 65℃
2	GL0038126	1.399424567471	> 65℃
3	GL0120095	0.83571495201794	55℃~65℃
4	GL0074258	0.89572778424265	55℃~65℃
5	GL0112518	0.7027544545928	55℃~65℃
6	GL0067868	-0.060637515738974	< 55℃

KẾT LUẬN

Dữ liệu metagenome của vi sinh vật ruột mối *C gestroi* đã được phân tích hiệu quả dưới sự hỗ trợ của công cụ tin sinh. Việc xử lý và khai thác các gen mã hóa cho enzyme cần phân lập nhờ các phần mềm dự đoán chức năng và vùng bảo tồn của Blast, cấu trúc bậc hai và trung tâm hoạt động bởi Phyre2, khả năng chịu kiềm của Alcapred và khả năng chịu nhiệt của enzyme của TBI. Căn cứ trên các dự đoán đã chọn được 6 ORF mã hóa cho enzyme phân giải lignocellulose để tiến hành thực nghiệm, nghiên cứu biểu hiện và xác định hoạt tính với các đặc điểm nổi bật đó là khả năng chịu kiềm và chịu nhiệt khá đa dạng.

Lời cảm ơn: *Công trình được thực hiện bằng nguồn kinh phí của đề tài Nghị định thư với Nhật Bản giai đoạn 2012-2015.*

TÀI LIỆU THAM KHẢO

Brune A, Emerson D, Breznak JA (1995) The Termite gut microflora as an oxygen sink: microelectrode determination of oxygen and pH gradients in guts of lower and higher termites. *Appl Environ Microbiol* 61: 2681–2687.

Do TH, Nguyen TT, Nguyen TN, Le QG, Nguyen C, Kimura K, Truong NH (2014) Mining biomass-d egrading 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.

Fan GL, Li QZ, Zuo YC (2013) Predicting acidic and alkaline enzymes by incorporating the average chemical shift and gene ontology informations into the general form of Chou's PseAAC. *Process Biochemistry* 48: 1048–1053.

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

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

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE (2015) The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc* 10: 845–858.

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

Madden T, PhD. The BLAST Sequence Analysis Tool. Bookshelf ID: NBK153387. http://www.ncbi.nlm.nih.gov/books/NBK1763/.

Nimchua T, Thongaram T, Uengwetwanit T, Pongpattanakitshote S, Eurwilaichitr L (2012) Metagenomic analysis of novel lignocellulose-degrading enzymes from higher termite guts inhabiting microbes. *J Microbiol Biotechnol* 22: 462–469.

USING BIOINFORMATIC TOOLS IN EXPLOITED GENE ENCODING ENZYME TO DECOMPOSE LIGNOCELLULOSE FROM METAGENOME OF FREE - LIVING BACTERIA IN THE GUT OF THE LOWER TERMITE *COPTOTERMES GESTROI*

Nguyen Minh Giang[1], Do Thi Huyen[2], Truong Nam Hai[2,✉]

[1]*Ho Chi Minh City University of Education*
[2]*Institute of Biotechnology, Vietnam Academy of Science and Technology*

SUMMARY

Microbial metagenome DNA in the guts of *Coptotermes gestroi* has been extracted and sequenced by metagenomic techniques. In previous studies, we acquired and sequenced more than 5 Gb of DNA metagenome DNA of the termite gut microbiota by next-generation sequencing (Illumina). Software MGA (MetaGeneAnnotator) exploited 125.431 open reading frames with 8508 ORFs related to carbohydrate metabolism, including 587 ORFs coding for enzymes involved in the hydrolysis of lignocellulose. We identified software to reliably predict function, structure and characteristics of proteins corresponding to DNA sequences encoding alkaline enzymes from the metgenome of *C gestroi*. The online software Alcapred was used to predict alkaline enzymes, Blastp to predict conserved domains of amino acid sequences deduced from ORFs, Phyre2 to predict the three dimentional structure and substrate binding site of the enzymes, TBI to predict melting temperature of the enzyme. We identified 6 ORFs encoding alkaline cellulases (GL0101308, GL0038126) or alkaline hemicellulases (GL0120095, GL0074258, GL0112518, GL0067868). The amino acid sequences deduced from ORFs had 90% coverage and from 44% to 99% identity to the corresponding sequences in NCBI by BLASTP. All of them contained conserved domains with corresponding activities and binding sites of the enzyme to the substrate. The three dimentional structures of amino acid sequences were predicted by Phyre2 with reliability from 98% to 100% to the annotated activities. Among six selected amino acid sequences, two sequences of enzymes had the melting temperature above 65 °C, three sequences had melting temperature from 55°C to 65°C and only one below 55°C.

Keywords: *Cellulase, Coptotermes gestroi, hemicellulase, lignocellulose, metagenomic, metagenome, bioinformatics*

✉ *Author for correspondence*: E-mail: tnhai@ibt.ac.vn