TAP CHI SINH HOC 2019, 41(2): 49–60
DOI: 10.15625/0866-7160/v41n2.13683

SCREENING GENES ENCODING PROTEIN PROTEASE INHIBITOR FROM METAGENOME OF SPONGE-ASSOCIATED MICROORGANISMS IN QUANG TRI SEA, VIETNAM

Tran Thi Hong[1,2,*], Pham Viet Cuong[1], Nguyen Thi Kim Cuc[3]

[1]Mientrung Institute for Scientific Research, VAST, Vietnam
[2]Graduate University of Science and Technology, VAST, Vietnam
[3]Institute of Marine Biochemistry, VAST, Vietnam

Received 13 March 2019, accepted 5 May 2019

ABSTRACT

Using metagenomics-based method to isolate new compounds from the marine environment are getting more and more attention in recent years. Based on metagenome library, bioinformatics methods is a powerful tool for screening genes with new biological activities from uncultured microorganisms and become a breakthrough in research and application of biotechnology. In this study we selected and used the samples DNA QT2 which had high DNA content and purity from a total of 6 DNA samples of sponge-associated microorganisms collected in Quang Tri sea (Vietnam) for metagenomic sequencing (DNA concentration is 202.5 ng, A260/A280 value is 1.80). 16S rRNA metagenomic sequencing data of QT2 produced 44,117,722 reads, which were assembled into 120,236 contigs. ORF prediction using Prodigal produced 386,416 ORFs. Functional annotation was conducted based on 7 different databases (NR, COG, CAZy, Swissprot, GO, KEGG, Pfam), and there are 266,553 genes were annotated using Swiss-Prot. In addition, based on the obtained metagenomic data, 50 complete genes encoding protease inhibitor proteins were revealed and among them, 28 genes encoding protein (> 50%) belonged to the serine protease inhibitor family, and 22 genes genes encoding belonged to the Inter-alpha-trypsin inhibitor group. NCBI BLAST screening results that these proteins had higher 50% identity to protease inhibitors.

Keywords: Bioinformatics, metagenomics, protease inhibitor, sponge, sponge-associated microorganisms.

Citation: Tran Thi Hong, Pham Viet Cuong, Nguyen Thi Kim Cuc, 2019. Screening genes encoding protein protease inhibitor from metagenome of sponge-associated microorganisms in Quang Tri sea, Vietnam. *Tap chi Sinh hoc*, 41(2): 49–60. https://doi.org/10.15625/0866-7160/v41n2.13683.

[*]*Corresponding author email*: tranhongtrn@gmail.com

TAP CHI SINH HOC 2019, 41(2): 49–60
DOI: 10.15625/0866-7160/v41n2.13683

SÀNG LỌC GEN MÃ HÓA PROTEIN ỨC CHẾ PROTEASE TỪ METAGENOMICS CỦA VI SINH VẬT LIÊN KẾT VỚI HẢI MIÊN BIỂN QUẢNG TRỊ, VIỆT NAM

Trần Thị Hồng[1,2,*], Phạm Việt Cường[1], Nguyễn Thị Kim Cúc[2]

[1]Viện Nghiên cứu Khoa học miền Trung, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, Việt Nam
[2]Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, Việt Nam
[2]Viện Hóa sinh biển, Viện Hàn lâm Khoa học và Công nghệ Việt Nam, Việt Nam

Ngày nhận bài 13-3-2019, ngày chấp nhận 5-5-2019

TÓM TẮT

Trong những năm gần đây, phương pháp dựa trên metagenomics để phân lập các hợp chất mới từ môi trường biển đang ngày càng được chú ý. Từ thư viện metagenome, bằng phương pháp tin sinh học có thể sàng lọc các gen có hoạt tính sinh học mới từ vi sinh vật không thông qua nuôi cấy. Đây thực sự là bước đột phá trong nghiên cứu và ứng dụng của công nghệ sinh học. Trong nghiên cứu này, từ 6 mẫu DNA của vi sinh vật liên kết với hải miên thu thập tại biển Quảng Trị, chúng tôi đã lựa chọn được một mẫu DNA QT2 đạt hàm lượng và độ tinh sạch cao để giải trình tự metagenomics (nồng độ DNA: 202,5 ng/µl, độ tinh sạch A260/A280 đạt 1,80). Sau khi giải trình tự shortgun metagenome toàn bộ của mẫu QT2 đã nhận được 44.117.722 reads, từ đó sắp xếp được 120.236 contigs. Tổng số khung đọc mở dự đoán (predict ORFs theo Prodigal) là 386.416 và đã chú giải chức năng gen theo 7 cơ sở dữ liệu khác nhau (NR, COG, CAZy, Swiss-Prot, GO, KEGG, Pfam), trong đó dựa trên cơ sở dữ liệu Swiss-Prot đã chú giải được chức năng cho 266.553 gen. Bên cạnh đó, dựa vào số liệu metagenome nhận được, đã sàng lọc được 50 gen hoàn chỉnh mã hóa protein ức chế protease. Trong đó, 28 gen mã hóa protein (trên 50%) thuộc họ serpin (ức chế serine protease), còn lại 22 gen mã hóa cho các protein thuộc nhóm ức chế Inter-alpha-trypsin. Kết quả so sánh một số trình tự axit amin sàng lọc được trên ngân hàng NCBI cho thấy các protein này có độ tương đồng trên 50% với chất ức chế protease.

Từ khóa: Hải miên, metagenomics, serpin, tin sinh học, vi sinh vật liên kết hải biên.

**Địa chỉ liên hệ email:* tranhongtrn@gmail.com

MỞ ĐẦU

Hải miên là vật chủ của cộng đồng vi sinh vật đa dạng, tính đến thời điểm hiện tại, nhiều hợp chất có hoạt tính sinh học đã được tách ra từ hải miên. Tuy nhiên, ngày càng có nhiều nghiên cứu xác nhận những hợp chất có hoạt tính sinh học này là do vi sinh vật cộng sinh với hải miên tổng hợp ra. Cách tiếp cận bằng phân lập và nuôi cấy vi sinh vật trong môi trường nhân tạo bị hạn chế rất nhiều, đặc biệt là vi sinh vật liên kết với các cơ thể khác bởi mối tương tác giữa chúng khá phức tạp. Hơn nữa, nếu khai thác hải miên để tách chiết các hoạt chất thì nguồn nguyên liệu có hạn này sẽ bị mất đi nhanh chóng, khó phục hồi và gây hủy hoại môi trường (Karuppiah & Li., 2017; Slaby et al., 2017; Thomas et al., 2017).

Sử dụng metagenomics có thể phân lập các cụm gen sinh tổng hợp mà cuối cùng có thể được khai thác để phát triển các nguồn sản phẩm tự nhiên bền vững bằng cách biểu hiện dị hợp (Gurgui & Piel, 2010). Hơn nữa,

metagenomics cũng có thể giúp làm sáng tỏ cấu trúc cộng đồng, cũng như sự trao đổi chất và chức năng của một cộng đồng vi khuẩn phức tạp cộng sinh với hải miên. Mặc dù vậy, việc nghiên cứu theo phương pháp metagenomics vẫn gặp nhiều khó khăn như DNA tách chiết dễ bị phân hủy một phần dẫn đến không đảm bảo hàm lượng theo yêu cầu hay trong sản phẩm tách DNA vẫn còn chứa nhiều sản phẩm lẫn tạp khác, vì vậy khó xác định được gen mục tiêu giữa nhiều gen tương đồng (Hyatt et al., 2010; Karuppiah & Li, 2017).

Sử dụng phương pháp chọn lọc chức năng từ thư viện metagenomics hải miên *Discodermia calyx* của Nhật Bản đã phân lập được cyclodipeptides (He et al., 2013). Tương tự đã nhận dạng được một nhóm các hợp chất hữu cơ dị vòng gồm 4 tiểu đơn vị porphyrin liên kết với nhau và 3 axít béo ß-hydroxyl có hoạt tính kháng khuẩn từ hải miên này (He et al., 2012). Ozturk et al. (2013) đã thiết kế thư viện cDNA để xác định sự đa dạng của các gen halogenase và 17 trình tự cDNA được cho là gen mã hóa cho tryptophan đã được nhận dạng mà phần lớn chúng ít quan hệ với các gen mã hóa chohalogenase đã được biết, biểu thị tiềm năng hệ vi sinh vật của hải miên Crambe crambe sản sinh ra các hợp chất hoạt tính sinh học mới (Ozturk et al., 2013).

Dựa trên trình tự của thư viện metagenomic từ vi sinh vật biển không nuôi cấy, đã sàng lọc được gen ức chế serine protease mới (serpin) gọi là Spi1C. Gen có vùng ORF là 642 bp, mã hóa cho polypeptide có 214 amino acid với khối lượng phân tử dự đoán 28,7 kDa. Protein Spi1C có hoạt tính ức chế một loạt các serine proteases như α-chymotrypsin và trypsin (Jiang et al., 2011). Tuy nhiên, chưa có nghiên cứu nào về sử dụng phương pháp này để khai thác gen ức chế protease từ metagenome của vi sinh vật liên kết với hải miên. Vì vậy, nghiên cứu này của chúng tôi đã mở ra một hướng đi đầy triển vọng cho việc khai thác gen ức chế protease phục vụ cho việc tìm kiếm các chất ức chế protease tái tổ hợp mới có hiệu quả điều trị cao trong y học.

VẬT LIỆU VÀ PHƯƠNG PHÁP NGHIÊN CỨU

Vật liệu

Các mẫu hải miên QT2, QT3, QT4, QT5, QT6 và QT7 được thu thập bằng thiết bị thở dưới nước khép kín (SCUBA: self contained underwater breathing apparatus) tại vùng biển Quảng Trị ở độ sâu 5–10 m; tọa độ 107°07’06,0”E; 17°04’50,2”N. Các mẫu được chứa trong các lọ đựng mẫu cùng với nước biển trong đó có 30% glycerol, bảo quản trong đá, vận chuyển về phòng thí nghiệm và giữ ở -25°C, trong vòng 1 tuần cho đến khi tách chiết DNA của vi sinh vật liên kết hải miên.

Kít tách DNA ZR Soil Microbe DNA MiniPrep™ (Zymo Research Corp.) và một số hóa chất điện di DNA được mua từ Merck (Đức); Sigma (USA) đã được sử dụng cho nghiên cứu.

Phương pháp tách DNA của vi sinh vật liên kết hải miên

Tách chiết DNA tổng số của vi sinh vật liên kết với hải miên theo phương pháp của Abe et al. (2012) với một số cải tiến nhỏ cho phù hợp với điều kiện của Việt Nam. Các mẫu hải miên được rửa 3 lần bằng nước biển nhân tạo vô trùng. 10 g mẫu được cắt nhỏ và nghiền đến đồng nhất trong dung dịch đệm TE (10 mM Tris HCl, 1 mM ethylene diaminetetraacetic acid (EDTA), pH 8,0). Đầu tiên, lọc hỗn hợp qua hai lớp vải màn, sau đó ly tâm 250 g trong 1 phút để loại bỏ các mảnh vỡ của hải miên và các chất bẩn. Dịch phía trên được ly tâm tiếp 8.000 g trong 15 phút để thu tế bào vi sinh vật. Rửa tế bào thu được bằng dung dịch TE50 (10 mM Tris–HCl, 50 mM EDTA, pH 8,0). DNA tổng số được tách bằng ZR Soil Microbe DNA MiniPrep™ (Zymo Research Corp.) theo hướng dẫn của nhà sản xuất.

Phân tích metagenomics của vi sinh vật liên kết với hải miên QT2

Đánh giá và tiền xử lý dữ liệu

Dữ liệu giải trình tự thô được đánh giá chất lượng bằng phần mềm FastQC, sau đó được tinh sạch nhằm loại bỏ những đoạn trình tự có chất lượng thấp và độ dài ngắn, sử dụng

phần mềm Trimmomatics (Bolger et al., 2014). Trong nghiên này tất cả những đoạn trình tự có điểm chất lượng nhỏ hơn 30 (QC < 30) và độ dài nhỏ hơn 70 bp đều được loại bỏ.

Lắp ráp DE NOVO metagenome

Dữ liệu sau tinh sạch được dùng để lắp ráp *de novo* metagenome sử dụng phần mềm SPAdes (Bankevich et al., 2012) với k-mer biến thiên từ 21 đến 55. Để chọn được tham số k-mer tối ưu, chúng tôi sử dụng phần mềm QUAST để đánh giá dựa trên các tiêu chí: kích thước hệ metagenome tổng số, độ dài contig lớn nhất, chỉ số N50 và tỷ lệ đoạn trình tự ánh xạ ngược lại (remapping) sử dụng phần mềm Bowtie2 (Langmead & Salzberg, 2012) và Qualimap (Garcia et al., 2012). Tất cả những contigs có kích thước nhỏ hơn 1.000 bp đều bị loại bỏ để thu được hệ metagenome cuối cùng.

Dự đoán gen

Hai phần mềm Prodigal (Hyatt et al., 2010) và MetageneMark (Zhu et al., 2010) với tham số mặc định được sử dụng để dự đoán gen trên hệ metagenome thu được. Để chọn ra tập gen chung nhất, hai tập gen dự đoán thu được từ hai phần mềm Prodigal và MetageneMark được phân cụm (clustering) bằng phần mềm CD-HIT (Li & Godzik, 2006) với mức độ tương đồng 90%. Điều này có nghĩa là, hai gen dự đoán từ hai phần mềm phải có mức độ tương đồng từ 90% trở lên thì mới được chọn làm gen dự đoán cuối cùng. Sau đó, loại bỏ các gen > 250 bp.

Chú giải chức năng gen

Tập gen dự đoán cuối cùng được so sánh với các cơ sở dữ liệu sinh học khác nhau bao gồm: CAZy (Cantarel et al., 2009) (sử dụng phần mềm DBCAN (Yin et al., 2012)), GO (Ashburner et al., 2006), COG (Tatusov et al., 2001), Swiss-Prot (Bairoch et al., 2000), KEGG (Kanehisa el al., 2011) và NR (Pruitt et al., 2007) (blast, evalue < 1.e-3, max_target_seqs 20). Trong đó, NR: cơ sở dữ liệu các trình tự protein không lặp lại từ các cơ sở dữ liệu GenPept, Swiss-Prot, PIR, PDF, PDB và RefSeq; CAZy: là cơ sở dữ liệu Carbohydrate Active enzyme; GO: Gen Ontology, Dự án Gen Ontology được xây dựng nhằm đưa ra những mô tả, định nghĩa những sản phẩm của gen. Dự án GO được phát triển bao gồm: Structured, controlled vocabularies (ontologies) nhằm mô tả các chức năng của gen liên quan đến các chu trình sinh học, thành phần tế bào và chức năng phân tử của 1 loài sinh vật độc lập; COG: Cluster of Orthologous Groups: Là cơ sở dữ liệu những trình tự protein được tạo ra bởi NCBI. Cơ sở dữ liệu này được tạo nên dựa trên mối quan hệ tiến hóa của hệ thống protein giữa vi khuẩn, tảo và sinh vật nhân chuẩn. Trình tự protein có thể được chia vào 1 loại của COG và mỗi loại của COG được tạo nên bởi những trình tự tương đồng và hình thành chức năng của protein; KEGG: Kyoto Encyclopedia of Gens and Genomes với cơ sở dữ liệu chính là KEGG PATHWAY. Cơ sở KEGG PATHWAY chia con đường sinh học thành 8 phần chính và mỗi phần được hình thành từ nhiều phần nhỏ khác nhau, mỗi phần được chú giải bởi các gen liên quan. Bằng việc sử dụng chú giải trên KEGG, chúng ta có thể tìm ra những gen liên qua đến những gen đã được chú giải một cách dễ dàng; UniProtKB/Swiss-Prot là một phần của UniProt Knowledgebase được chú giải và đánh giá thủ công. Nó là một cơ sở dữ liệu của các trình tự protein không lặp lại có chất lượng chú giải được kiểm chứng bằng thực nghiệm; Pfam là cơ sở dữ liệu tập hợp các họ của protein. Các trình tự protein được tạo bởi một hoặc nhiều vùng chức năng, thông thường là các domain. Sự kết hợp khác nhau sẽ làm tăng tính đa dạng của protein được tìm thấy trong tự nhiên.

KẾT QUẢ VÀ THẢO LUẬN

Kết quả tách DNA của vi sinh vật liên kết hải miên biển Quảng Trị

Từ 6 mẫu hải miên QT2, QT3, QT4, QT5, QT6 và QT7 đã tách được DNA của vi sinh vật liên kết theo phương pháp đã mô tả. Kết quả điện di trên gel agarose 1% và kết quả đo nồng độ, độ tinh sạch của DNA trên máy Nanodrop 2000 spectrophotometer nhận được cho thấy tất cả các mẫu đều đã tách được DNA tổng số, tuy nhiên lượng DNA tách được ở mỗi mẫu khác nhau (hình 1). Mẫu

QT6 thu được nhiều DNA nhất (228,2 ng/µl), sau đó đến mẫu QT2 (202,5 ng/µl), còn các mẫu khác lượng DNA thu được khá ít từ 28,7–160,9 ng/µl. Tuy nhiên, xét về độ tinh sạch thì QT2 có độ tinh sạch cao hơn so với QT6 (QT2 có độ tinh sạch A260/A280= 1,80 còn QT6 có độ tinh sạch A260/A280=1,70. Theo hướng dẫn chuẩn bị mẫu để giải metagenomic của IGA Tech (IGA Tech: Metagenomics sample preparation guidelines) thì nồng độ DNA nên đạt 200 ng/mẫu và độ tinh sạch A260/A280 ≥ 1,8. Vì vậy, chúng tôi lựa chọn DNA tổng số của vi sinh vật liên kết hải miên QT2 để giải trình tự metagenomics. Ngoài ra, từ kết quả trên cũng cho thấy, phương pháp tách chiết DNA mà chúng tôi thực hiện đã không làm ảnh hưởng đến chất lượng DNA từ các mẫu tách chiết.

Hình 1. Hình thái hải miên QT2 và điện di đồ trên gel agarose 1% DNA của vi sinh vật liên kết với các mẫu hải miên biển Quảng Trị

Kết quả phân tích dữ liệu metagenomics mẫu QT2

Kết quả tiền xử lý và lắp ráp reads của DNA metagenome

Sau khi giải trình tự shortgun metagenome toàn bộ của mẫu QT2, dữ liệu thô thu được bao gồm 2 tập tin (R1 và R2) (hình 2). Sau quá trình tinh sạch loại bỏ tất cả những trình tự không bắt cặp với nhau ở 2 tập tin (những trình tự chất lượng thấp và ngắn sử dụng phần mềm trimmomatic), tổng số hơn 44 triệu đoạn trình tự paired reads được dùng để lắp ráp *de novo* metagenome sử dụng phần mềm SPAdes. Tổng kích thước hệ lắp ráp thu được là khoảng 418 Mb bao gồm 102.236 contigs. Contig có kích thước dài nhất là hơn 855 kb, contigs nhỏ nhất là 1.000 bp, độ dài trung bình là 4.089 bp. Gần 90% số đoạn trình tự có thể ánh xạ ngược lại với hệ gen lắp ráp (bảng 1). Điều này chứng tỏ rằng tất cả thông tin đã được chuyển đến tổ hợp lắp ráp. Kết quả nhận được cho thấy các contigs chủ yếu phân bố trong khoảng từ 1.000 đến 100.000 bp. Tỷ lệ GC% trong hệ gen của mẫu QT2 là 61,82%. Nhìn chung, bộ gen của các vi sinh vật liên kết với hải biên có hàm lượng GC cao. Theo kết quả phân tích metagenomics của hải miên Địa Trung Hải cho thấy tỉ lệ GC của hệ gen là 36–70%. Hàm lượng GC tương đối cao là một đặc điểm của metagenomic hải miên (Horn et al., 2016).

Bảng 1. Kết quả lắp ráp DNA metagenome của mẫu QT2

Chỉ số	QT2	Chỉ số	QT2
Tổng số reads (paired-end)	44.117.722	Trung bình của contig (nt)	4.089
Phạm vi độ dài reads (nt)	70–126	N50 (nt)	6.929
Số lượng contigs	102.236	N75 (nt)	1.718
Độ dài tổng số của contigs (nt)	418.103.634	Lượng GC (%)	61,82
Contig lớn nhất (nt)	855.566	% mapped reads	89,88
Contig ngắn nhất (nt)	1.000		

Hình 2. Kết quả tinh sạch dữ liệu

Kết quả dự đoán gen

Dự đoán gen trong giải trình tự metagenomics vẫn là một vấn đề khó khăn. Một số phần mềm không đảm bảo có thể lắp ráp được hết các bộ gen riêng lẻ trong một mẫu đại diện điển hình, do đó, các chuỗi chạy tạo ra một số lượng lớn các chuỗi ngắn mà không rõ nguồn gốc chính xác. Vì các chuỗi này thường nhỏ hơn độ dài trung bình của gen nên các thuật toán phải đưa ra dự đoán dựa trên rất ít dữ liệu. Trong số các phần mềm dự đoán gen hiện nay thì Prodigal và MetageneMark được đánh giá là có thể dự đoán các gen ngắn với độ chính xác cao (Hyatt et al., 2010). Kết quả dự đoán gen bằng phần mềm Prodigal và MetageneMark nhận được lần lượt là khoảng 366 Mb (386.416 ORFs) và 361 Mb (380.886 ORFs). Kết quả dự đoán gen của hai phần mềm khá tương đồng nhau, với gen lớn nhất có kích thước là 66.639 bp, độ dài trung bình là 864 bp và tỷ lệ GC là khoảng hơn 62%. Sau khi loại bỏ tất cả những gen có kích nhỏ hơn 250 bp, sử dụng phần mềm CD-HIT với mức độ tương đồng 90%, thu được tập gen cuối cùng có tổng kích thước gần 360 Mb bao gồm 372.732 unified genes, trong đó có 262.159 gen hoàn chỉnh (chiếm 70,33%) (gen có đủ mã mở đầu và mã kết thúc); 53.162 (14,26%) gen thiếu mã kết thúc 3’; 49.569 (13,3%) gen thiếu mã mở đầu 5’ và số lượng gen thiếu cả mã mở đầu và mã kết thúc chỉ có 7.842 gen, chiếm 2,10%. Phân bố độ dài cho thấy gen dự đoán chủ yếu có kích thước từ khoảng 250 bp đến khoảng 2.000 bp (bảng 2).

Bảng 2. Kết quả dự đoán gen và kiểm tra tính toàn vẹn của gen (mẫu QT2)

Chỉ số	Prodigal	Metagenemark	Cluster
Tổng gen dự đoán	386.416	380.886	372.732
Tổng độ dài gen dự đoán (nt)	366.878.679	361.181.676	359.967.498
Gen lớn nhất (nt)	66.639	66.639	66.639
Gen ngắn nhất (nt)	250	250	252
Độ dài trung bình của gen	864	864	965
Hàm lượng GC (%)	62,33	62,45	62,40
Tình trạng gen	Gen thống nhất giữa hai phần mềm		Phần trăm
Gen hoàn chỉnh	262.159		70,33
Thiếu đầu 3’	53.162		14,26
Thiếu đầu 5’	49.569		13,30
Thiếu cả 2 đầu	7.842		2,10

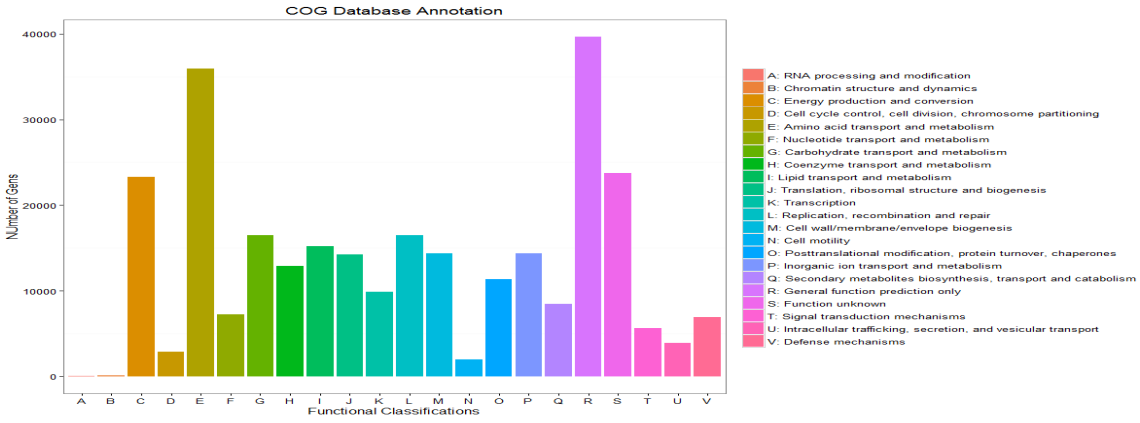
Kết quả chú giải và phân loại chức năng gen

Trong khi các nghiên cứu trước đây chủ yếu đánh giá đa dạng loài trong cộng đồng, thì ngày nay, nhiều nghiên cứu về metagenomic đã tập trung vào gen và chức năng của gen. Trong các nghiên cứu như vậy, các lần đọc trình tự ngắn được ánh xạ tới các cơ sở dữ liệu (ví dụ: COG, KEGG, Swiss-Protein…) để xác định các kết quả khớp với chức năng gen và protein đã biết và chú thích (Carr & Borenstein, 2014). Kết quả chú giải bằng các cơ sở dữ liệu cho thấy, với tổng số 372.732 trình tự gen (axit amin), có 360.564 (96,74%) gen được chú giải trên cở sở dữ liệu NR; 266.553 gen được chú giải trên Swissprot chiếm 71,51%; 274.632 gen chiếm 73,68% được chú giải trên cơ sở dữ liệu COG; chỉ có 11.974 (3,21%) gen được chú giải trên cơ sở dữ liệu CAZy; số gen được chú giải trên cơ so dữ liệu GO là 165.552 gen chiếm 44,42%, 244.436 gen được chú giải trên cơ so dữ liệu KEGG chiếm 65,58%; đối với cơ sở dự liệu Pfam, có 273.826 (73,46%) gen được chú giải (bảng 3).

Bảng 3. Tổng hợp kết quả chú giải chức năng gen (QT2)

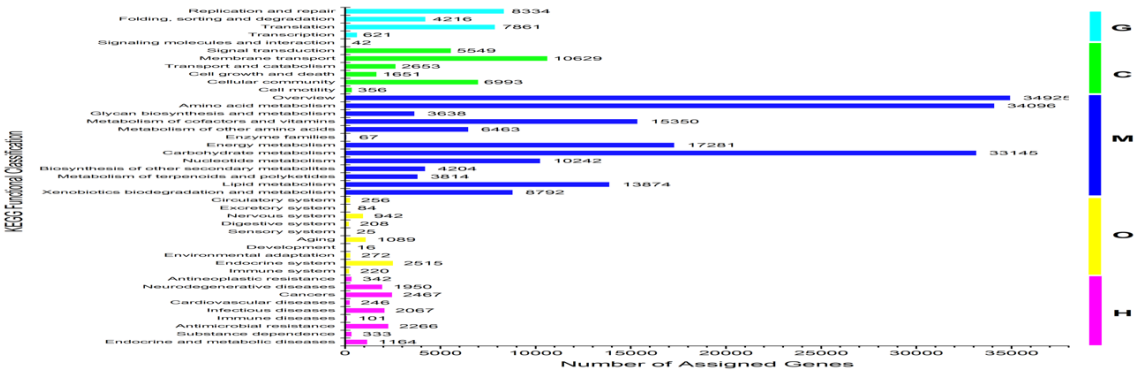
Dữ liệu	NR	Swiss-Prot	COG	CAZy	GO	KEGG	Pfam
Chú giải gen	360.564	266.553	274.632	11.974	165.552	244.436	273.826
%	96,74	71,51	73,68	3,21	44,42	65,58	73,46

Kết quả phân loại chức năng gen trên cơ sở dữ liệu COG cho thấy chủ yếu lượng gen được phân loại chức năng vào nhóm R: chức năng chung (Genral function prediction only), tiếp theo là nhóm E: Trao đổi và vận chuyển axít amin (Amino Acid Transport and Metabolism); theo sau là nhóm C: Chuyển hóa và sản xuất năng lượng (Energy Production and Conversion). Các nhóm chức năng còn lại có số lượng gene tương đối bằng bằng nhau. Riêng chỉ có nhóm A: Chỉnh sửa và xử lý RNA (RNA processing and modification) và nhóm B: Cấu trúc và động lực học của chất nhiễm sắc (Chromatin Structure and dyamics) là hầu như không có gen tương đồng (hình 3).



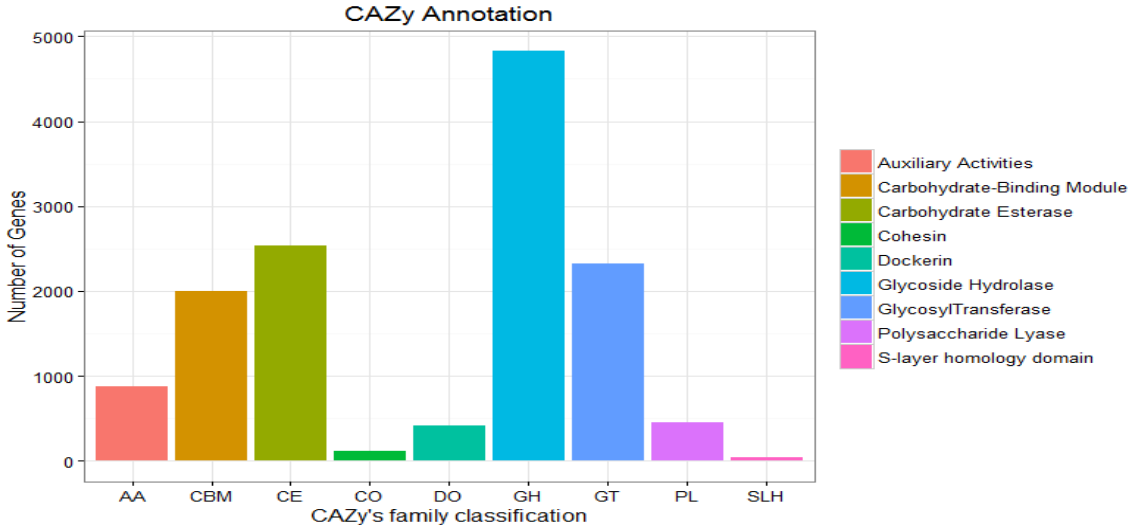
Hình 3. Phân loại chức năng gen trên CSDL COG

Kết quả phân loại trên cơ sở dữ liệu KEGG được trình bày trong hình 4. Kết quả chú giải cho thấy gen dự đoán chủ yếu có chức năng liên quan đến con đường trao đổi chất (M: Metabolism); tiếp theo đó là nhóm Cellular Process (C) và Xử lý thông tin di truyền (Gentic Information Processing). Và một phần nhỏ gen tham gia vào nhóm O: Organismal Systems và H: Human Diseases.



Hình 4. Kết quả phân loại trên cơ sở dữ liệu KEGG

Kết quả phân loại nhóm chức năng enzym cho thấy, dữ liệu gen chủ yếu thuộc vào nhóm GH (Glycoside Hydrolase) với khoảng gần 5.000 gen, tiếp theo sau là hai nhóm Carbohydrate Esterase (CE) và Glycosyl Transferase (GT) với khoảng 2.500 gen tương đồng. Số lượng đoạn trình tự thuộc nhóm chức năng Carbohydrate Binding Module (CBM) thấp hơn 1 chút, khoảng 2.000 trình tự. Các nhóm chức năng còn lại có số lượng gen tương đồng không đáng kể, khoảng dưới 1.000 trình tự gen (hình 5).



Hình 5. Phân nhóm chức năng của enzym trên CSDL CAZY (QT2)

Kết quả sàng lọc gen mã hóa ức chế protease

Dựa trên kết quả thu được từ chú giải chức năng gen, chúng tôi đã sàng lọc được 50 gen liên quan đến chất ức chế protease (bảng 4). Trong đó có 28 gen, chiếm 56% được chú giải thuộc họ serpin (serine protease inhibitor), 22 gen (44%) thuộc nhóm Inter-alpha-trypsin inhibitor. Gen ngắn nhất là 198 bp, mã hóa cho 66 axit amin; gen dài nhất là 2.406 bp, mã hóa cho 802 axit amin. Một số gen đã được xác định là gen có hoạt tính sinh học mới so với ở Việt Nam (bảng 5). Nhằm xác định lại độ tin cậy của kết quả chú giải trên, một số trình tự axít amin đã được lựa chọn để so sánh protein trên NCBI (hình 6). Kết quả sau so sánh cho thấy các axít amin này thuộc nhóm ức chế protease tương ứng với kết quả chú giải. Như vậy, kết quả chú giải trên có độ tin cậy cao.

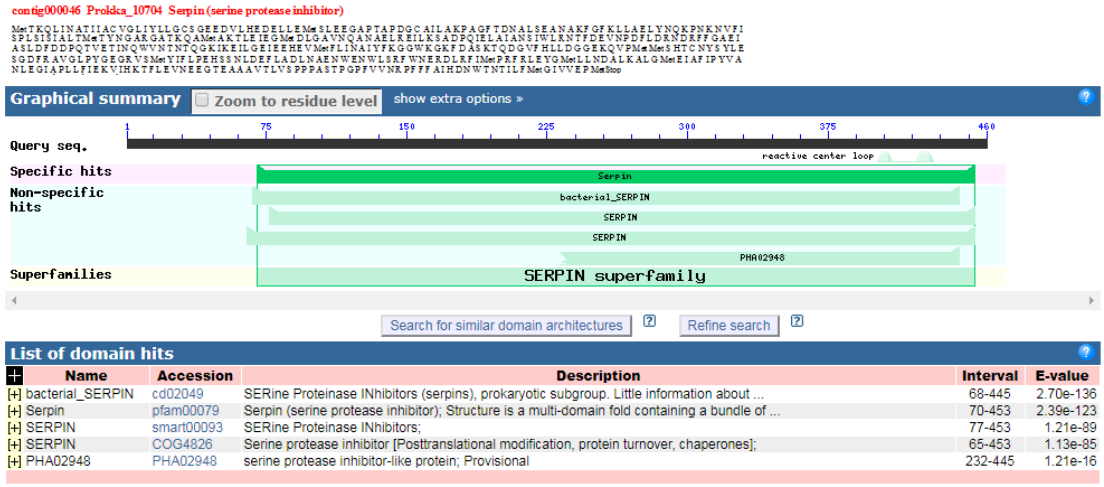
Bảng 4. Kết quả sàng lọc các gen có hoạt tính protease inhibitor mẫu QT2

STT	**Contig**	**Locus_tag**	**Acid amin**	**Uni_ accession_1**	**UniProtKB_ product**	**Uni_score**	**Uni_ evalue**
1	contig000016	Prokka_05808	442	Q5RB37	ITIH chain H3	89.4	1.00E-17
2	contig000019	Prokka_06418	323	O02668	ITIH chain H2	61.2	5.00E-09
3	contig000019	Prokka_06445	398	Q61703	ITIH chain H2	71.6	4.00E-12
4	contig000046	Prokka_10704	429	Q9D154	Serpin	184	6.00E-52
5	contig000046	Prokka_10705	405	Q5BIR5	Serpin	214	1.00E-63
6	contig000127	Prokka_20087	328	Q3T052	ITIH chain H4	62	3.00E-09
7	contig000172	Prokka_24210	400	Q8BJD1	ITIH chain H5	71.6	4.00E-12
8	contig000213	Prokka_27784	418	Q5BIR5	Serpin B8	216	5.00E-64
9	contig000213	Prokka_27785	405	Q99574	Neuroserpin	209	2.00E-61
10	contig000314	Prokka_34813	736	A6X935	ITIH	171	6.00E-43
11	contig000433	Prokka_41698	419	Q5BIR5	Serpin B8	231	6.00E-70
12	contig000631	Prokka_52340	325	Q3T052	ITIH chain H4	58.9	3.00E-08
13	contig000726	Prokka_56621	398	Q61703	ITIH chain H2	57	2.00E-07
14	contig000981	Prokka_67673	428	Q90935	Neuroserpin	214	1.00E-62
15	contig001114	Prokka_72799	419	Q5BIR5	Serpin B8	219	4.00E-65
16	contig001390	Prokka_82416	386	A6X935	ITIH	114	4.00E-26
17	contig001690	Prokka_91580	454	Q5BIR5	Serpin B8	152	5.00E-40
18	contig001737	Prokka_93032	412	Q5BIR5	Serpin B8	214	1.00E-63
19	contig001737	Prokka_93033	223	Q99574	Neuroserpin	87.8	6.00E-19
20	contig001813	Prokka_95168	412	Q99574	Neuroserpin	207	3.00E-60
21	contig002069	Prokka_102253	280	Q8PTN8	serpin	175	7.00E-50
22	contig002236	Prokka_106102	324	A2VE29	ITIH chain H5	64.7	4.00E-10
23	contig002339	Prokka_108516	478	Q14624	ITIH chain H4	63.5	2.00E-09
24	contig002592	Prokka_114432	355	Q90935	Neuroserpin	197	3.00E-57
25	contig002838	Prokka_119867	334	Q14624	ITIH chain H4	55.5	3.00E-07
26	contig003102	Prokka_125566	401	Q8BJD1	ITIH chain H5	58.5	5.00E-08
27	contig003892	Prokka_140659	323	Q3T052	ITIH chain H4	67	7.00E-11
28	contig004584	Prokka_152589	631	Q61703	ITIH chain H2	66.2	6.00E-10
29	contig005997	Prokka_173538	430	Q8PTN8	Serpin	206	2.00E-59
30	contig006820	Prokka_184178	417	Q5BIR5	Serpin B8	237	5.00E-72
31	contig007047	Prokka_186946	497	Q61703	ITIH chain H2	122	2.00E-28
32	contig007181	Prokka_188591	146	Q96P15	Serpin B11	105	5.00E-26
33	contig007964	Prokka_197295	66	Q90935	Neuroserpin	51.6	7.00E-08
34	contig008443	Prokka_202257	443	P50453	Serpin B9	213	2.00E-62
35	contig010618	Prokka_222724	382	Q8BJD1	ITIH chain H5	64.3	8.00E-10
36	contig012483	Prokka_237228	378	Q9JK88	Serpin I2	57	1.00E-07
37	contig015758	Prokka_258680	430	Q9S7T8	Serpin-ZX	143	1.00E-36
38	contig020504	Prokka_283125	394	Q90935	Neuroserpin	73.2	8.00E-13
39	contig020772	Prokka_284248	402	Q90935	Neuroserpin	213	1.00E-62
40	contig020806	Prokka_284376	802	Q61703	ITIH chain H2	142	2.00E-33
41	contig020909	Prokka_284844	362	A6X935	ITIH	104	6.00E-23
42	contig021896	Prokka_288956	722	P56652	ITIH chain H3	170	8.00E-43
43	contig024785	Prokka_300295	303	Q29052	ITIH chain H1	55.5	3.00E-07
44	contig030105	Prokka_318139	717	Q9GLY5	ITIH chain H3	116	2.00E-25
45	contig033816	Prokka_328453	391	B4USX2	Serpin B10	220	1.00E-65
46	contig038363	Prokka_339464	376	Q9CQV3	Serpin B11	82	7.00E-16
47	contig040171	Prokka_346561	423	Q99574	Neuroserpin	211	1.00E-61
48	contig044964	Prokka_352966	457	Q5JJ64	Serpin	249	7.00E-76
49	contig060339	Prokka_377096	149	Q9UIV8	Serpin B13	66.6	4.00E-12
50	contig067320	Prokka_385523	98	Q5NBM0	Putative serpin	66.2	1.00E-12

Chú thích: ITIH: Inter-alpha-trypsin inhibitor heavy; Serpin: serine protease inhibitor.

Bảng 5. Gen có hoạt tính sinh học mới (so với ở Việt Nam)

STT	Tên gen	Nu length	Siminarity (%)	Đặc điểm
1	Predicted_gene_346561	1398	55,98	Neuroserpin; AltName: Full=Peptidase inhibitor 12; Short=PI-12; AltName: Full=Serpin I1; Flags: Precursor
2	Predicted_gene_91473	1893	49,75	Inter-alpha-trypsin inhibitor heavy chain H2; Short=ITI heavy chain H2; Short=ITI-HC2; Short=Inter-alpha-inhibitor heavy chain 2



Hình 6. Kết quả so sánh trình tự a xít amin sàng lọc trong metagenomics với protein trên NCBI

Chất ức chế Serine protease là một họ quan trọng và lớn nhất của chất ức chế protease. Chúng hoạt động như một điều biến (modulator) và tham gia vào rất nhiều quá trình phân giải protein quan trọng, liên kết hóa trị với protein đích và bất hoạt chúng. Vì vậy, chất ức chế protein thuộc nhóm này luôn được các nhà khoa học trên thế giới quan tâm nghiên cứu và tìm kiếm chất mới. Ngoài việc phân lập được các chất ức chế protease bằng phương pháp truyền thống, thú vị thay, Jiang et al. (2011) dựa trên trình tự của thư viện metagenomic từ vi sinh vật biển không nuôi cấy đã sàng lọc được gen ức chế serine protease mới (serpin) gọi là Spi1C. Gen có ORF 642 bp, mã hóa cho polypeptide 214 amino acid với khối lượng phân tử dự đoán 28,7 kDa. Protein Spi1C có hoạt tính ức chế một loạt các serine proteases như α-chymotrypsin và trypsin. Như vậy có thể thấy sàng lọc gen có hoạt tính sinh học mới trong đó có hoạt tính ức chế protease từ vi sinh vật liên kết với hải miên bằng phương pháp metagenomics là hướng đi mới rất tiềm năng (Jiang et al., 2011).

KẾT LUẬN

Trong nghiên cứu này, từ 6 mẫu DNA của vi sinh vật liên kết với hải miên thu thập tại biển Quảng Trị (Việt Nam), chúng tôi đã lựa chọn được 1 mẫu DNA đạt hàm lượng và độ tinh sạch cao là DNA QT2 để giải trình tự metagenomics (nồng độ DNA: 202,5 ng/µl, độ tinh sạch A260/A280 đạt 1,80). Sau khi giải trình tự shortgun metagenome toàn bộ của mẫu QT2 đã nhận được 44.117.722 reads, từ đó sắp xếp được 120.236 contigs. Tổng số khung đọc mở dự đoán (predict ORFs theo Prodigal) là 386.416 và đã chú giải chức năng gen theo 7 cơ sở dữ liệu khác nhau (NR, COG, CAZy, Swiss-Prot, GO, KEGG, Pfam). Dựa trên kết quả chú giải gen, đã sàng lọc được 50 gen hoàn chỉnh

mã hóa protein ức chế protease. Trong đó, 28 gen (trên 50%) mã hóa cho các protein thuộc serpin (ức chế serine protease), còn lại 22 gen mã hóa cho các protein thuộc nhóm ức chế Inter-alpha-trypsin.

Lời cảm ơn: Nghiên cứu này được thực hiện bằng nguồn kinh phí của ĐTĐLCN.17/14.

TÀI LIỆU THAM KHẢO

*A*shburner M., Ball C., Blake J., 2006. Gen ontology: tool for the unification of biology. The gen ontology consortium database resources of the national center for biotechnology information. *Nucleic acids research*, 34.

Bairoch A., Apweiler R., 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic acids research*, 28(1): 45–48.

Bankevich A., Nurk S., Antipov D., Gurevich A. A., Dvorkin M., Kulikov A. S., Lesin V. M., Nikolenko S. I., Pham S., Prjibelski A. D., et al., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology*, 19(5): 455–477.

Bolger A. M, Lohse M., Usadel B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15): 2114–2120.

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 research*, 37(1): 233–238.

Carr R and Borenstein E., 2014. Comparative Analysis of Functional Metagenomic Annotation and the Mappability of Short Reads. *PLoS One*, 9(8): e105776.

García-Alcalde F., Okonechnikov K., Carbonell J., Cruz L. M., Götz S., Tarazona S., Dopazo J., Meyer T. F., Conesa A., 2012. Qualimap: evaluating next-genration sequencing alignment data. *Bioinformatics*, 28(20): 2678–2679.

Gurgui C., Piel J., 2010. Metagenomic approaches to identify and isolate bioactive natural products from microbiota of marine sponges. *Methods Mol. Biol.*, 668: 247–264.

He R., Bochu W., Wakimoto T., Wang M., Zhu L and Abe I., 2013. Cyclodipeptides from Metagenomic Library of a Japanese Marine Sponge. *J. Braz. Chem. Soc.*, 24(12): 1926–1932.

He R., Wakimoto T., Egami Y., Kenmoku H., Ito T., Asakawa Y., Abe I., 2012. Heterologously expressed b-hydroxyl fatty acids from a metagenomic library of a marine sponge. *Bioorganic & Medicinal Chemistry Letters*, 22: 7322–7325.

Horn H., Slaby B. M., Jahn M. T., Bayer K., Moitinho-Silva L., Förster F., et al., 2016. An enrichment of CRISPR and other defense-related features in marine sponge-associated microbial metagenomes. *Front Microbiol.*, 7: 1751.

Hyatt D., Chen G. L., LoCascio P. F., Land M. L., Larimer F. W., Hauser L. J., 2010. Prodigal: prokaryotic gen recognition and translation initiation site identification. *BMC bioinformatics*, 11(1):1.

Jiang C. J., Hao Z. Y., Zeng R., Shen P. H., Li J. F. and Wu B., 2011. Characterization of a Novel Serine Protease Inhibitor Gene from a Marine Metagenome. *Mar. Drugs*, 9: 1487–1501.

Kanehisa M., Goto S., Sato Y., Furumichi M., Tanabe M., 2011. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*: gkr988.

Karuppiah V., Li Z., 2017. Marine Sponge Metagenomics. *Springer Handbook of Marine Biotechnology*: 457–473.

Langmead B., Salzberg S. L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4): 357–359.

Li W., Godzik A., 2006. Cd-hit: a fast program for clustering and comparing large sets of

protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

O¨ztu¨rk., Jaeger L. D., Smidt H & Sipkema D., 2013. Culture-dependent and independent approaches for identifying novel halogenases encoded by *Crambe crambe* (marine sponge) microbiota. *Sci. Reports*, 3: 2780.

Pruitt K. D., Tatusova T., Maglott D. R., 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 35(suppl 1): 61–65.

Tatusov R. L., Natale D. A., Garkavtsev I. V., Tatusova T. A., Shankavaram U. T., Rao B. S., Kiryutin B., Galperin M. Y., Fedorova N. D., Koonin E. V., 2001. The COG database: new developments in phylogentic classification of proteins from complete genomes. *Nucleic acids research*, 29(1): 22–28.

Thomas T., Silva L. M., Lurgi M., Björk J. R., Easson C., García C. A., et al., 2016. Diversity, structure and convergent evolution of the global sponge microbiome. *Nat. Commun*, 7: 11870.

Yin Y., Mao X., Yang J., Chen X., Mao F., Xu Y., 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic acids research*, 40(1): 445–451.

Zhu W., Lomsadze A., Borodovsky M., 2010. Ab initio gen identification in metagenomic sequences. *Nucleic acids research*, 38(12): 132–132.