Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.0007

ÁP DỤNG MÔ HÌNH ẨN KẾT HỢP THUẬT TOÁN BIMETA TRONG VIỆC GOM NHÓM TRÌNH TỰ METAGENOMIC

Văn Đình Vỹ Phương[1,3], Trần Văn Lăng[3], Trần Văn Hoài[1], Lê Văn Vinh[2]

[1] Khoa Khoa học và Kỹ thuật máy tính, Trường Đại học Bách khoa TPHCM
[2] Khoa Công nghệ thông tin, Trường Đại học Sư phạm Kỹ thuật TPHCM
[3] Khoa Công nghệ thông tin, Trường Đại học Lạc Hồng

phuongvdv@cse.hcmut.edu.vn, lang@lhu.edu.vn, hoai@cse.hcmut.edu.vn, vinhlv@fit.hcmute.edu.vn

TÓM TẮT*— Phân nhóm và xác định loài trong metagenomic là một trong những bài toán lớn của lĩnh vực Sinh-Tin học hiện nay. Bài báo trình bày giải pháp gom nhóm các trình tự trong metagenomic áp dụng mô hình ẩn (Latent Dirichlet Allocation) để tìm chủ đề ẩn có ý nghĩa, làm chủ đề đặc trưng cho trình tự. Từ chủ đề đặc trưng, tiến hành xác định nhóm của trình tự bằng phương pháp Kullback Leibler dựa trên sự phân bổ của chủ đề thay vì tính toán trực tiếp giữa các trình tự. Giải pháp kế thừa thuật toán BiMeta, tạo các nhóm trình tự gốc dựa vào thông tin trùng lắp trước khi áp dụng mô hình ẩn tìm chủ đề, khi đó, dữ liệu phân tích để tìm chủ đề ẩn được giảm đáng kể.*

Từ khóa*— metagenomic, gom nhóm, trình tự, LDA.*

I. GIỚI THIỆU

Sinh-Tin học (bioinformatics) là một khái niệm không còn xa lạ trong lĩnh vực nghiên cứu hiện nay. Việc giải mã trình tự DNA luôn là vấn đề tối quan trọng để hiểu rõ bản chất của sinh vật, vi sinh vật sống. Cho đến thời điểm hiện nay, giải mã trình tự có 2 cách tiếp cận; theo phương pháp truyền thống (Chain-termination methods, gọi tắt là Sanger) và phương pháp giải trình tự thế hệ mới (Next Generation Sequencing, gọi tắt là NGS [1]). Mỗi phương pháp vẫn có những ưu nhược điểm riêng.

Môi trường sống luôn là một tập thể của nhiều vi sinh vật, có sự tác động qua lại lẫn nhau cũng như sự cộng sinh trong việc tồn tại, vì thế việc tách độc lập bộ gen để nuôi cấy và tiến hành nghiên cứu là một vấn đề tốn nhiều chi phí và đôi khi không thể tách riêng biệt được. Chính vì những khó khăn này mà cơ sở dữ liệu gen cho các loài vi sinh vật đã biết vẫn còn giới hạn về số lượng. Từ thách thức trên, một hướng đi mới là các vi sinh vật trong môi trường sau khi thu thập được, không qua giai đoạn nuôi cấy mà được đưa trực tiếp vào thiết bị giải trình tự để đưa ra trình tự sinh học của nhiều vi sinh vật cộng sinh với nhau. Vấn đề đặt ra đó là làm sao xác định được trình tự của vi sinh vật cụ thể trong một tập các trình tự hỗn hợp đó.

Lĩnh vực metagenomic ra đời trong bối cảnh này; đó là sự tập hợp, sự pha trộn một lượng lớn các trình tự của rất nhiều loài vi sinh vật khác nhau. Metagenomic được lấy từ môi trường có thể chứa đến hàng triệu trình tự với sự phong phú và đa dạng khác nhau. Vì thế để tìm hiểu về các trình tự, việc đầu tiên trong nghiên cứu metagenomic là tiến hành phân tích, gom cụm các trình tự con (read, fragment) có thành phần, tính chất giống nhau theo từng nhóm. Việc phân nhóm trình tự có độ chính xác cao dẫn đến dễ dàng hơn trong việc nhận định trình tự thuộc loài vi sinh vật đã có hay mới; số loài và mức độ phong phú của chúng trong môi trường sống; từ đó bổ sung vào nguồn cơ sở dữ liệu còn ít ỏi hiện nay, làm tiền đề cho việc hiểu được các chức năng, vai trò của mỗi loài cũng như sự tác động cộng sinh của chúng.

Bài báo được trình bày thành 4 phần: phần I giới thiệu về metagenomic; phần II trình bày các giải pháp gom nhóm trình tự metagenomic đã và đang được sử dụng; phần III trình bày phương pháp đề xuất để phân nhóm trình tự và cuối cùng là phần thực nghiệm, kết luận phương pháp đã đề xuất.

II. GIẢI PHÁP GOM NHÓM TRÌNH TỰ METAGENOMIC

Hiện tại có khá nhiều phương pháp được đưa ra trong việc phân tích trình tự metagenomic. Tuy nhiên vẫn chưa có một giải pháp nào được coi là tối ưu và chính xác nhất, giải quyết trọn vẹn cho đến từng cá thể. Việc xác định, phân loài trình tự hiện nay đa phần dựa vào một số phương pháp dựa trên các đặc trưng như: tính tương đồng giữa các trình tự (homology-based), tính hợp thành (composition-based).

Phương pháp phân loài trình tự metagenomic theo hướng tiếp cận dựa trên tính tương đồng thực hiện so sánh trình tự cần xác định với các trình tự đã có trong cơ sở dữ liệu. Thuật toán BLAST được sử dụng phổ biến trong việc xây dựng các ứng dụng phân loài trình tự dựa theo tính tương đồng. Một số ứng dụng theo hướng này như MEGAN, CARMA thực hiện việc sắp xếp trình tự DNA trực tiếp với các gen cần so sánh. Phương pháp phân loài theo tính tương đồng có ưu điểm cho độ chính xác cao nếu đoạn trình tự cần phân tích giống hoặc gần giống với đoạn trình tự đã có trong cơ sở dữ liệu. Nhược điểm là hiện tại nguồn dữ liệu (các mẫu trình tự đã biết) ít, nên việc so sánh, tìm kiếm sự tương đồng đạt tỷ lệ thấp. Theo [2], hiện có hơn 99% trình tự gen của vi sinh vật chưa được nghiên cứu hoặc nhận diện. Dẫn đến hạn chế trong việc thực hiện so sánh với nguồn dữ liệu mẫu khi phân tích một trình tự mới nào đó.

Phương pháp tiếp cận theo tính hợp thành thực hiện việc phân loài trình tự dựa trên đặc trưng được lấy trực tiếp từ các thành phần trong trình tự metagenomic. Hiện nay, phương pháp dựa trên tính hợp thành được chia thành ba nhóm: nhóm học có giám sát (supervised learning approaches), nhóm học không giám sát (unsupervised approaches)

và nhóm học bán giám sát (semi-supervised learning approaches). Phương pháp học có giám sát có ý nghĩa gần giống với phương pháp dựa trên tính tương đồng ở điểm cần cơ sở dữ liệu tham khảo. Điều này dẫn đến hạn chế là phần lớn các vi sinh vật trong môi trường chưa được nhận diện. Để giải quyết hạn chế này, phương pháp không giám sát thực hiện việc phân loài bằng cách rút trích thông tin trực tiếp từ các trình tự cần phân loài, nghĩa là không sử dụng cơ sở dữ liệu tham khảo. Bài toán thực hiện việc phân cụm (gom cụm) các trình tự trong metagenomic có cùng một nhóm. Việc gom cụm chưa yêu cầu phải đưa ra được kết luận nhóm đó thuộc giống loài nào. Mặc dù đầu ra của phương pháp chưa đưa ra được kết quả như mong muốn của các nhà nghiên cứu sinh học. Tuy nhiên, đây là bước đi có hiệu quả trong việc phân loài trình tự có tính giống nhau trong một metagenomic mà không phải có nguồn dữ liệu tập vi sinh vật đã biết để tham chiếu.

Tình hình nghiên cứu ngoài nước về metagenomic theo phương pháp dựa trên tính hợp thành được quan tâm đáng kể. Một số nghiên cứu gần đây được đánh giá cao như: MBBC của Y. Wang và cộng sự [3] đề xuất giải pháp gom nhóm dựa trên tần suất k-mer sử dụng thuật toán Expectation Maximization. Cơ sở của phương pháp MBBC là các nhóm loài với độ phủ gen khác nhau có tần suất k-mer khác nhau; các nhóm loài có tần suất k-mer bằng hoặc gần bằng nhau thì giống nhau. Tuy nhiên, cần xác định khả năng những loài có tần suất k-mer giống nhau nhưng có thể không cùng nhóm loài và ngược lại. MetaCluster-TA của Yi Wang và cộng sự [4] nhận định việc gán nhãn phân loài các trình tự là vấn đề quan trọng trong qua trình phân tích metagenomic. Trong nghiên cứu, tác giả đưa ra khái niệm virtual contig (có chiều dài lên đến 10kb) đại diện cho mỗi nhóm. Mặc dù việc gán nhãn có kết quả khả quan hơn so với một số phương pháp khác, tuy nhiên MetaCluster-TA không phù hợp để phân tích trình tự ngắn, đồng thời phương pháp sử dụng thuật giải BLASTN, có độ phức tạp phụ thuộc vào tổng độ dài trình tự, dẫn đến thời gian thực thi tăng cao khi số lượng dữ liệu trình tự dài nhiều.

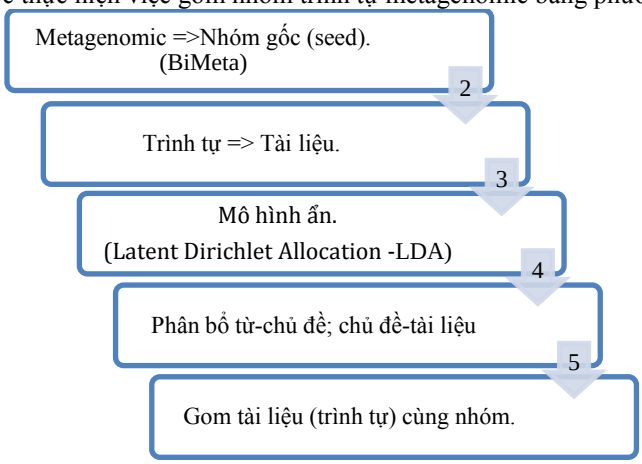
Ở trong nước, nghiên cứu về metagenomic và các hướng áp dụng cũng đang được quan tâm trong thời gian gần đây. Nghiên cứu [5, 6] của Viện Công nghệ sinh học đặt ra những vấn đề của nội tại sinh học cần giải quyết và hướng đi của việc áp dụng các ứng dụng metagenomic trong lĩnh vực sinh học. Nghiên cứu [7] áp dụng kỹ thuật metagenomic để giải quyết bài toán trong di truyền học. Lê Văn Vinh và cộng sự đưa ra những phương pháp được đánh giá khả thi trong việc giải quyết bài toán phân loài trình tự metagenomic như: [8], đánh giá hiệu năng của các phương pháp phổ biến để gom cụm trình tự metagenomic; Nghiên cứu [9] đề xuất giải pháp gom nhóm MetaAB, cho phép nhận diện và phân loài các trình tự thành các nhóm dựa trên độ phong phú bằng cách giảm kích thước mô hình (reduced-dimension model), giúp tăng tốc độ xử lý và kết hợp tiêu chuẩn Bayesion để phân loài. Kết quả thực nghiệm bằng giải pháp MetaAB cho thấy, độ chính xác trong việc xác định nhóm loài là 6/7 so với 4/7 của phương pháp AbundanceBin. Tuy nhiên thời gian xử lý vẫn chưa thể hiện rõ được sự khác biệt đối với các trình tự không bị lỗi (Error-Free Sequencing Reads); Nghiên cứu [10] đề xuất giải pháp BiMeta thực hiện phân loài dựa trên các trình tự không trùng lặp. Thuật toán BiMeta thực hiện qua hai bước: Bước 1, tiến hành gom các trình tự thành từng nhóm dựa trên thông tin trùng lặp giữa các trình tự. Điểm nổi bật của bước này là việc tạo các nhóm (gọi là nhóm gốc – seed) cùng lúc với việc phân nhóm. Một trình tự A được phân vào nhóm gốc 1 (gọi là SG_1) nếu A không liền kề với bất kỳ SG_i nào khác; Bước 2, tiến hành kết hợp các nhóm dựa trên tần suất k-mer của các tập trình tự không trùng lặp sử dụng thuật giải K-Means. Kết quả thực nghiệm so sánh giữa BiMeta, MetaCluster (5.0) và AbundanceBin thể hiện được sự cải tiến đáng kể của BiMeta về độ chính xác. Trên dữ liệu giả lập, BiMeta có độ chính xác là 8/10 mẫu trình tự so với 2/10 mẫu trình tự sử dụng MetaCluster (5.0) và hơn toàn bộ các mẫu sử dụng phương pháp AbundanceBin. Giải thuật sử dụng trong BiMeta tăng thời gian xử lý vì việc thực hiện so sánh dựa trên các nhóm gốc thay vì dựa trên toàn bộ trình tự trong tập dữ liệu cần phân nhóm. Tính chính xác của bước1 trong việc phân nhóm và xây dựng nhóm gốc phụ thuộc nhiều vào việc dự đoán đúng các trùng lặp giữa các trình tự. Đây cũng là vấn đề đặt ra bài toán cần phải giải quyết và chứng minh tính đúng đắn; Trong nghiên cứu [11, 12] đưa ra phương pháp sử dụng Fuzzy K-medoids, phương pháp đếm k-mer cho việc phân nhóm trình tự metagenomic dựa trên độ phong phú nhằm nâng cao hiệu quả việc rút trích đặc trưng độ phong phú của gen. Phương pháp thực hiện qua ba bước chính: Bước 1, thực hiện việc rút trích các k-mer; Bước 2, phân nhóm các k-mer dựa trên mức độ thành viên; Bước 3, gán các trình tự vào từng nhóm dựa trên kết quả của việc phân nhóm các k-mer. Trình tự được gán vào một nhóm nếu k-mer của nhóm đó là lớn nhất. Kết quả thực nghiệm của nhóm tác giả được so sánh với AbundanceBin. Độ chính xác có nhỉnh hơn so với AbundanceBin, tuy nhiên thời gian xử lý vẫn còn phải xem xét.

Có thể thấy rằng, việc phân loài trình tự metagenomic được các nhà nghiên cứu quan tâm và đưa ra nhiều giải pháp để thực hiện. Tuy nhiên, vẫn còn nhiều vấn đề trong việc gom cụm (bước tiền đề cho việc phân loài) và xác định thông tin nhóm loài đã biết, chưa biết, đặc trưng của loài, nhóm loài cộng hưởng với loài trong metagenomic, sự liên quan, sự độc lập của từng loài.

III. PHƯƠNG PHÁP ĐỀ XUẤT

Mô hình ẩn (Latent Dirichlet Allocation - LDA) [13, 14, 15] được sử dụng phổ biến trong việc xem xét sự tương quan, thông tin đặc trưng và tìm chủ đề ẩn của các tài liệu văn bản cần phân tích. Từ định hướng này, có thể xem xét mỗi trình tự cần xác định nhóm trong metagenomic như là một tài liệu, khi đó, áp dụng mô hình ẩn để tìm chủ đề ẩn mà trình tự trong metagenomic có thể có. Bài báo đề xuất phương pháp gom cụm trình tự bằng cách sử dụng mô hình ẩn để tìm chủ đề ẩn trong trình tự, các trình tự có cùng chủ đề thì gom thành một nhóm. Và để giản lược dữ liệu đầu vào khi xây dựng mô hình tìm chủ đề ẩn, phương pháp áp dụng việc tạo nhóm gốc trong thuật toán BiMeta, mô hình ẩn phân tích dữ liệu là tập tài liệu các nút gốc này. Sau khi có được các mô hình, chủ đề, phương pháp sử dụng phép đo Kullback Leibler [16] để gom cụm tài liệu (cũng là các trình tự) theo chủ đề tương ứng. Phép đo giữa các tài liệu bằng Kullback Leibler phù hợp hơn phương pháp SKWIC trong [17].

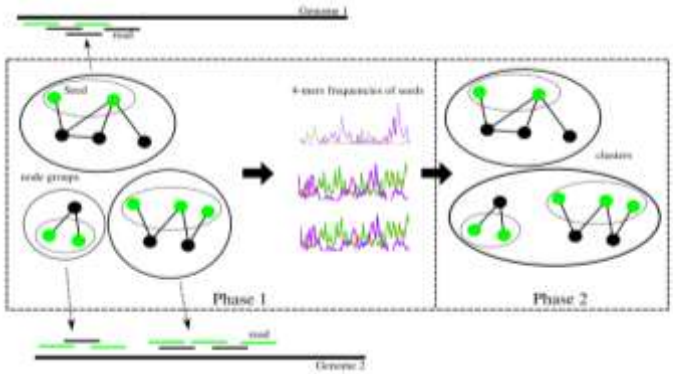
Hình *1* trình bày các bước thực hiện việc gom nhóm trình tự metagenomic bằng phương pháp đề xuất.



Hình 1. Các bước gom nhóm trình tự

A. Tạo nhóm gốc

Thuật toán BiMeta [10] được chia thành 2 bước. Bước thứ nhất (phase 1 trong **Hình *2***) gom các trình tự có sự trùng lặp thông tin lại thành nhóm (gọi là nhóm gốc - seed), bước thứ hai (phase 2 trong **Hình *2***) thực hiện gom các nhóm đã có ở bước một thành nhóm lớn hơn dựa vào rút trích đặc trưng của các nhóm. Thuật toán đã đưa ra ý tưởng thay vì phải xét đặc trưng của tất cả các trình tự, lúc này chỉ xét đặc trưng mỗi nhóm và dựa vào đó để gom nhóm. Theo [10], trùng lặp thông tin giữa hai trình tự là khi các trình tự này cùng thuộc một gen và có chung một đoạn trình tự con. Ví dụ, cho hai gen gọi lại g_1 và g_2, với $g_1 = AATTCTAG$ và $g_2 = AACGTAGTGG$. Giả sử với k-mer = 4 sẽ có một số trình tự con như sau: $read_1^{g_1} = AATT$, $read_2^{g_1} = TTCT$, $read_1^{g_2} = AACG$. $read_1^{g_1} \cap read_2^{g_1}$ (∩: trùng lặp thông tin) do trùng TT , nhưng $read_1^{g_1}, read_1^{g_2}$ lại không trùng lắp vì thuộc 2 gen khác nhau.



Hình 2. Ý tưởng của thuật toán BiMeta

B. Chuyển trình tự thành tài liệu

Như đã đề cập, mô hình ẩn thực hiện việc phân tích tập tài liệu dạng văn bản để tìm chủ đề ẩn của các tài liệu đó. Mỗi tài liệu có nội dung và số lượng từ có thể khác nhau. Vì vậy, để áp dụng mô hình ẩn cho việc phân tích trình tự metagenomic, cần chuyển đổi trình tự (là một dạng một chuỗi ký tự hợp thành từ 4 ký tự A, G, T, C) thành các từ có độ dài k-mer, ứng với mỗi từ trong tài liệu. Theo [14, 18], k=4 được đánh giá là phù hợp.

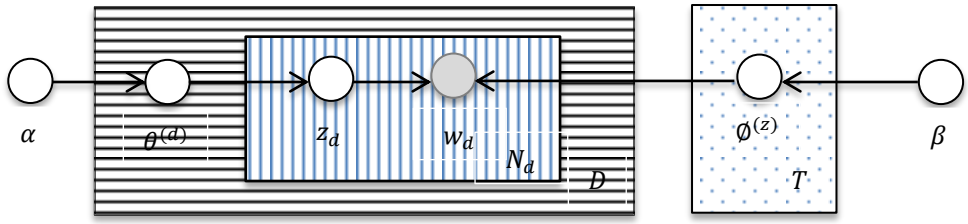
Các từ trong tài liệu sẽ được tham chiếu để xây dựng bộ từ điển (bộ từ điển là tập hợp từ có xuất hiện trong tài liệu), hỗ trợ trong việc tính toán để xây dựng mô hình trong. Số lượng ký tự để xây dựng từ là 4 (A, G, T, C), giả sử với k=5, như vậy, bộ từ điển có tổng cộng là 4^5=1024 từ. Số từ trong một tài liệu sẽ là $N_d = Length_d - k + 1$ từ. Ví dụ: trình tự AGCTCTGAGA (với k=5), sẽ được chuyển thành document là: AGCTC GCTCT CTCTG TCTGA CTGAG TGAGA ($N_d = 10 - 5 + 1 = 6$).

C. Mô hình ẩn

Có nhiều mô hình xác suất được sử dụng để phân tích chủ đề ẩn và ý nghĩa của nội dung như Blei, 2003; Griffiths and Steyvers, 2002, 2003, 2004; Hofmann, 1999, 2001. Trong đó, mô hình ẩn (Latent Dirichlet Allocation - LDA) được Blei đề xuất vào năm 2003 [13, 19] dựa trên mô hình xác suất để lựa chọn tập từ trong tài liệu có ý nghĩa đặc trưng và thể hiện được ý nghĩa hay nội dung của toàn tài liệu. Một tài liệu có thể ẩn chứa nhiều chủ đề khác nhau. Mô hình được xây dựng để tìm ra các chủ đề nổi bật của tài liệu cần phân tích. Các ký hiệu được quy ước cho việc xác định giá trị biến quan sát, biến ẩn cần suy luận trong mô hình như sau:

z_d: là một chủ đề có thể có trong tài liệu d (chứa tập các từ tạo thành chủ đề).

D: là tập tài liệu cần phân tích để tìm chủ đề ẩn. d là một tài liệu con trong tập D.
N_d: là số lượng từ có trong tài liệu d.
w_d: là tập từ trong tài liệu d.
T: là số lượng chủ đề.
$\emptyset^{(z)}$: sự phân bổ từ ứng với chủ đề z.
$\theta^{(d)}$: sự phân bổ của chủ đề ứng với tài liệu d.



Hình 3. Mô hình ẩn

Mô hình được xác định sự phân bổ của từ được tính như sau:

$$P(w_i) = \sum_{j=1}^{T} \emptyset^{(j)} \theta^{(d)} \quad (1)$$

Trong đó; $\emptyset^{(j)} = P(w_i|z_i = j)$. $\theta^{(d)} = P(z)$. Với $P(w_i|z_i = j)$ thể hiện xác suất của từ w_i trong chủ đề j. $P(z)$ thể hiện xác suất phân bổ của chủ đề z trong tài liệu đang được phân tích.

Biến α là mật độ xác suất phân bổ trực tiếp được định nghĩa bằng công thức (2) với α_i thể hiện sự tiền quan sát số lần xuất hiện của chủ đề j trong tài liệu, trước khi quan sát cụ thể 1 từ nào đó trong tài liệu. Để đơn giản hóa vấn đề, giả thiết các α_i có giá trị bằng nhau, để chỉ xét 1 giá trị α duy nhất.

$$Dir(\alpha_1, \dots, \alpha_T) = \frac{\Gamma(\sum_i \alpha_j)}{\prod_j \Gamma(\alpha_j)} \prod_{j=1}^{T} p_j^{\alpha_j - 1} \quad (2)$$

Ngoài ra, một biến khác khác là β thể hiện số lần xuất hiện của từ có trong chủ đề, trước khi từ được quan sát cụ thể. Theo Blei, Griffiths và Steyvers, $\alpha = 50 / T$ và $\beta = 0.01$ là giá trị phù hợp cho việc xác định mô hình.

Mô hình ẩn (**Hình *3***) thể hiện các biến quan sát (T, D, N, w) và các biến ẩn $(\emptyset, \theta, z)$ là ẩn số cần tìm kiếm giá trị. Mũi tên chỉ sự phụ thuộc điều kiện giữa các biến. Hình chữ nhật thể hiện quá trình lặp để xác định mẫu ứng với số lần tương ứng. Ví dụ, hình chữ nhật gạch dọc (chỉ chứa z và w), thể hiện N_d bước lặp đối với tài liệu d, hình chữ nhật gạch ngang (chứa θ) xác định phân bổ chủ đề trên mỗi tài liệu d so với tổng số tài liệu là D, Hình chữ nhật chấm nhỏ (chứa ∅) thể hiện phân bổ của các từ trong chủ đề, cho đến khi T chủ đề được tạo ra.

D. Phân bổ từ-chủ đề, phân bổ chủ đề-tài liệu

Hofmann sử dụng thuật toán Expectation-Maximization để tính $\emptyset, \theta$. Tuy nhiên, thuật toán này gặp vấn đề cực đại cục bộ. Vì thế phương pháp Gibb Sampling đã được thực hiện để tính toán $\emptyset, \theta$ dựa vào chủ đề (không tính $\emptyset, \theta$ trực tiếp), với mỗi $z_t, t \epsilon [1..T]$ thể hiện 1 chủ đề với N_{wd} từ ($N_{wd} < N_d$). Thuật toán Gibbs Sampling Markov Chain Monte Carlo, phù hợp cho việc rút trích chủ đề từ một tập dữ liệu lớn.

Các tài liệu được thể hiện bằng một tập các từ với chỉ số w_i và chỉ số tài liệu là d_i. Gibbs Sampling xem xét mỗi từ trong tập dữ liệu được chọn và tính toán sự phân bổ của từ-chủ đề so với tất cả các từ còn lại. Khi đó, sự phân bổ của từ thứ i trong chủ đề j được tính toán theo công thức (3).

$$P(z_i = j|z_{-i}, w_i) \propto \emptyset_i^j \theta_j^d \quad (3)$$

Và phân bổ từ-chủ đề (∅), phân bổ chủ đề-tài liệu (θ) được tính theo công thức (4) và (5). Trong đó C^{WT} là ma trận số lượng từ w gán trong chủ đề j và C^{DT} là ma trận số lượng chủ đề j trong tài liệu d.

$$\emptyset_i^j = \frac{c_{ij}^{WT} + \beta}{\sum_{k=1}^{W} c_{kj}^{WT} + W\beta} \quad (4)$$

$$\theta_j^d = \frac{c_{dj}^{DT} + \alpha}{\sum_{k=1}^{T} c_{dk}^{DT} + T\alpha} \quad (5)$$

E. Gom tài liệu cùng nhóm

Các chủ đề được rút trích từ tập tài liệu có thể được suy dẫn để trả lời các câu hỏi về sự liên quan giữa các tài liệu, ý chính của tài liệu. Tương ứng với phân tích trình tự trong metagenomic trả lời cho câu hỏi tìm đặc trưng của các trình tự và trình tự nào có liên quan với nhau (hay cùng nhóm với nhau). Hai tài liệu hay trình tự được coi là tương đồng (cùng nhóm) nếu có chung chủ đề. Để tính toán sự tương đồng, thay vì xem xét nội dung chi tiết của tài liệu, ta tính toán sự tương đồng bằng sự phân bổ θ của tài liệu d1 (θ_{d1}) và d2 (θ_{d2}), bài báo sử dụng phương pháp Kullback Leibler [16] được xem là phù hợp hơn so với phép đo bằng K-Mean. Công thức được tính như sau:

$$D(\theta_{d1}, \theta_{d2}) = \sum_{j=1}^{T} \theta_{d1_j} log_2 \frac{\theta_{d1_j}}{\theta_{d2_j}} \quad (6)$$

$D(\theta_{d1}, \theta_{d2}) = 0$ khi với tất cả các giá trị j, $\theta_{d1_j} = \theta_{d2_j}$. Do $D(\theta_{d1}, \theta_{d2})$ là số không âm, khi đó, sự khác biệt sẽ là:

$$KL(\theta_{d1}, \theta_{d2}) = \frac{1}{2}[D(\theta_{d1}, \theta_{d2}) + D(\theta_{d2}, \theta_{d1})] \quad (7)$$

IV. THỰC NGHIỆM VÀ KẾT LUẬN

A. Thực nghiệm

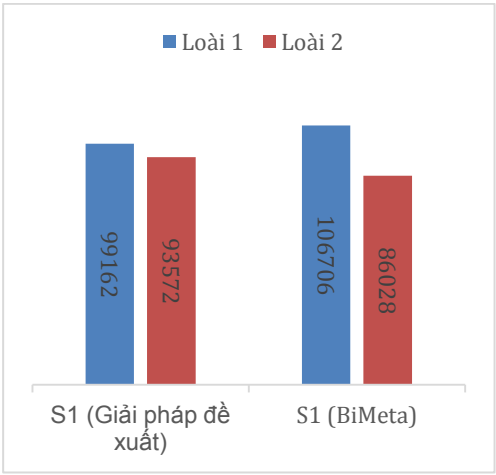
Dữ liệu sử dụng để thực nghiệm được kế thừa từ dữ liệu thực nghiệm trong nghiên cứu [10], là dữ liệu được phát sinh giả lập bằng ứng dụng MetaSim [20]. Dữ liệu được chia thành 2 loại, loại trình tự ngắn và loại trình tự dài. **Bảng *1*** và **Bảng *2*** thể hiện thông tin các mẫu thực nghiệm. Trong đó, tên mẫu để phân biệt các mẫu thực nghiệm, với ký hiệu S thể hiện cho metagenomic chứa các trình tự ngắn (mỗi trình tự ngắn có chiều dài ~100bp), ký hiệu R thể hiện cho metagenomic chứa các trình tự dài (mỗi trình tự dài > 700bp). Số loài, ứng với số nhóm loài có trong mỗi mẫu (giả thiết là biết trước số loài trong mẫu thực nghiệm). Tỉ lệ, cho biết tỉ lệ của từng loài có trong mẫu (ví dụ 1:1 nghĩa là số trình tự của các loài trong mẫu là bằng nhau). Số trình tự, cho biết số lượng trình tự có trong mẫu (cần nhận diện trình tự thuộc nhóm nào).

Bảng 1. Dữ liệu trình tự ngắn

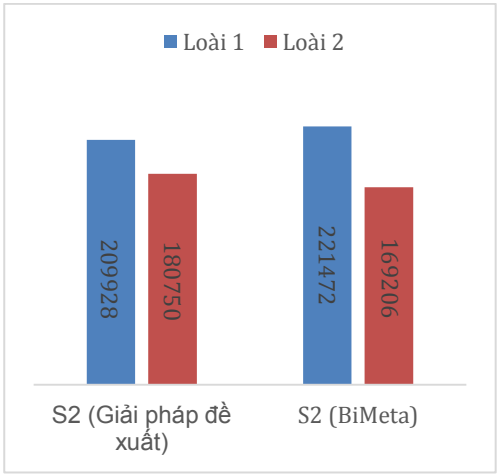
Tên mẫu	Số loài	Tỉ lệ	Số trình tự
S1	2	1 :1	192734
S2	2	1 :1	390678
S3	3	3 :2 :1	1426776

Bảng 2. Dữ liệu trình tự dài

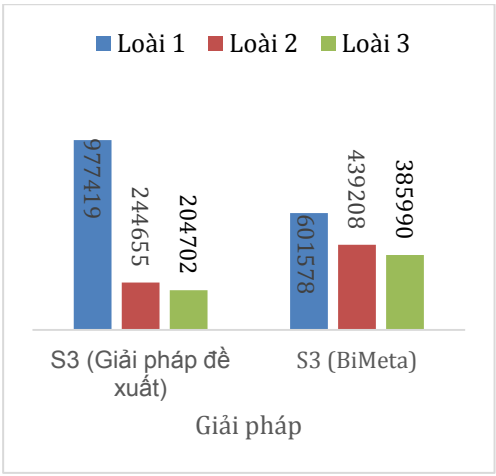
Tên mẫu	Số loài	Tỉ lệ	Số trình tự
R1	2	1:1	82960
R2	2	1:1	77293
R7	3	1:1 :8	290473
R9	6	1:1:1:1:2:14	285065



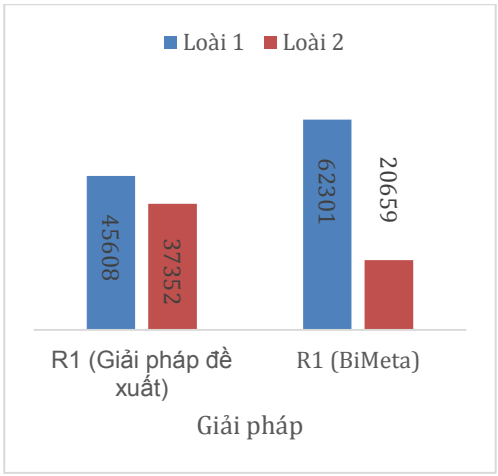
Hình 4. Phân nhóm trình tự ngắn S1 (tỉ lệ 1:1)



Hình 5. Phân nhóm trình tự ngắn S2 (tỉ lệ 1:1)

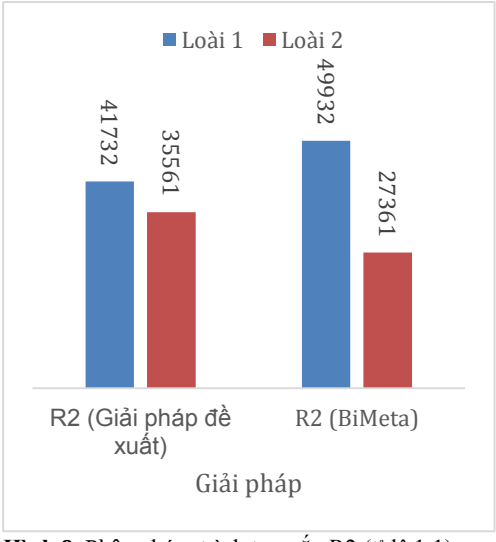


Hình 6. Phân nhóm trình tự ngắn S3 (tỉ lệ 3:2:1)

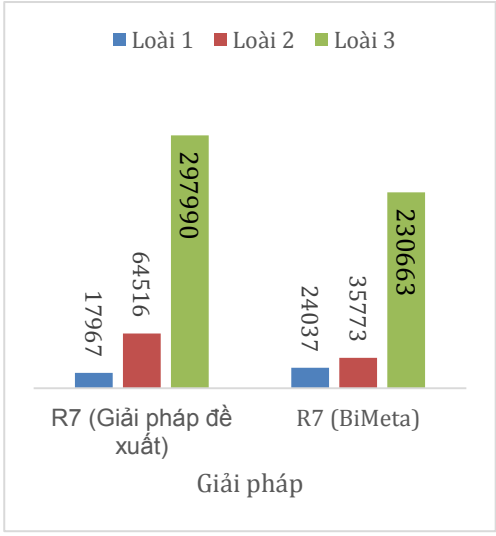


Hình 7. Phân nhóm trình tự dài R1 (tỉ lệ 1:1)

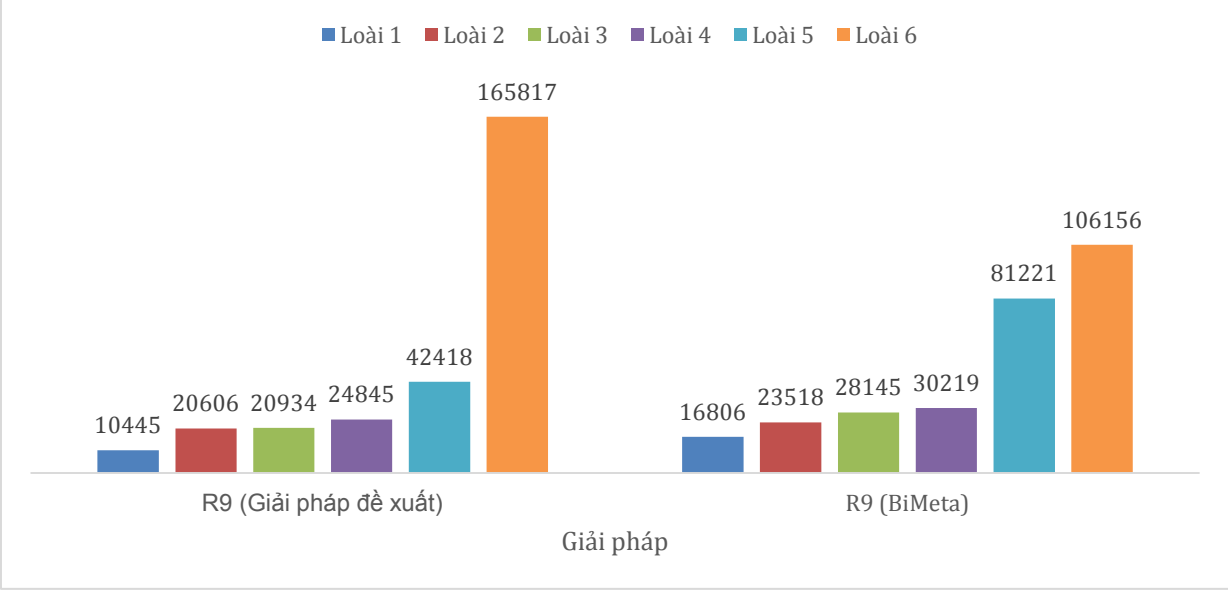
Kết quả thực nghiệm gom nhóm trình tự trong các mẫu dữ liệu được so sánh giữa phương pháp đề xuất với phương pháp BiMeta cho thấy, trình tự phân nhóm sử dụng dữ liệu mẫu bằng phương pháp đề xuất có sự cải thiện so với phương pháp BiMeta. Kết quả thực nghiệm đối với mẫu S1, S2, R1, R2 thể hiện ở **Hình 4**, **Hình 5**, **Hình 7**, **Hình 8**, **Hình *10*** cho thấy trình tự được phân theo số nhóm được cải thiện để đạt được ngưỡng mong muốn là cân bằng theo đúng tỉ lệ đã cho. Kết quả thực nghiệm đối với mẫu R6, R9 thể hiện ở **Hình *6***, **Hình *9*** cho kết quả chưa đạt như mong đợi, tuy nhiên cũng có thể xác định đây là một phương pháp phù hợp để gom cụm trình tự metagenomic.



Hình 8. Phân nhóm trình tự ngắn R2 (tỉ lệ 1:1)



Hình 9. Phân nhóm trình tự dài R7 (tỉ lệ 1:1:8)



Hình 10. Phân nhóm trình tự dài R9 (tỉ lệ 1:1:1:1:2:14)

B. Kết luận

Bài báo đề xuất việc sử dụng mô hình ẩn để tìm kiếm sự tương quan giữa các trình tự trong metagenomic thông qua chủ đề ẩn. Các trình tự có cùng chủ đề được xem là thuộc một nhóm. Phương pháp sử dụng phép đo Kullback Leibler để đo khoảng cách giữa các chủ đề (thay vì đo khoảng cách trực tiếp giữa các trình tự). Ngoài ra, để giản lược số lượng trình tự cần phân tích tìm mô hình ẩn, tiền dữ liệu được xử lý bằng thuật toán BiMeta trong việc gom nhóm trình tự gốc. Kết quả thực nghiệm cho thấy phương pháp đề xuất có tỉ lệ xác định trình tự và gom nhóm có sự cải thiện hơn so với thuật toán BiMeta.

TÀI LIỆU THAM KHẢO

[1] Michael L. Metzker et al, "Sequencing technologies – the next generation," *Nature Rev. Genet*, vol. 11, pp. 31-46, 2010.

[2] Teeling H, Hanno Glöckner, Frank Oliver, "Curent opportunities and challenges in microbial metagenome analysis – A bioinformatic perpective," *Briefings In Bioinformatics*, vol. 13, no. 6, pp. 728-742, 2012.

[3] Y. Wang, H. Hu, X. Li, "MBBC: An efficient approach for metagenomic binning based on clustering," *BMC Bioinformatics*, vol. 16, no. 1, pp. 1-11, 2015.

[4] Y. Wang, H. Chi Ming Leung, S. Ming, Yiu et al, "MetaCluster-TA: Taxonomic annotation for metagenomic data based on assembly-assisted binning," *BMC Genomics*, vol. 15, no. 1, pp. 1-9, 2014.

[5] Thi Huyen Do et al, "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," *Journal of Bioscience and Bioengineering*, vol. 118, no. 6, p. 665–671, 2014.

[6] Nguyễn Minh Giang, Đỗ Thị Huyền, Trương Nam Hải, "Sử dụng công cụ tin sinh trong nghiên cứu metagenomics-hướng nghiên cứu và ứng dụng mới trong sinh học," *Tạp chí Khoa học Trường Đại học Sư phạm TP.HCM,* vol. 2, no. 67, pp. 167-177, 2015.

[7] N. T. Thảo, "Nghiên cứu gene mã hoá Enzyme tham gia thuỷ phân Cellulose từ khu hệ vi khuẩn trong ruột mối bằng kỹ thuật metagenomics," Đại học Quốc gia Hà Nội, Hà Nội, 2015.

[8] Lê Văn Vinh, Trần Văn Lăng, Trần Văn Hoài, "Hiệu năng của các giải pháp gom cụm trình tự," *Tạp chí Khoa học và Công nghệ 52 (1B),* vol. 52, pp. 28-36, 2014.

[9] Van Vinh Le, Lang Van Tran, Hoai Van Tran, "MetaAB - A novel abundance-based binning approach for metagenomic sequences," in *Nature of Computation and Communication*, Springer International Publishing, 2015, pp. 132-141.

[10] Vinh LV, Lang TV, Binh LT, Hoai TV, "A two-phase binning algorithm using l-mer frequency on groups of non-overlapping reads," *Algorithms for Molecular Biology: AMB*, vol. 10, no. 2, 2015.

[11] Le Van Vinh, Tran Van Lang, Tran Van Hoai, "An abundance-based binning of metagenomic reads using Fuzzy K-medoids method," *Kỷ yếu Hội nghị Quốc gia lần thứ VII về Nghiên cứu cơ bản và ứng dụng công nghệ thông tin (FAIR),* pp. 25-30, 2014.

[12] Le Van Vinh, Tran Van Lang, Tran Van Hoai, "A novel l-mer counting method abundance-based binning of metagenomic reads," *Journal of Computer Science and Cybernetics*, vol. 30, no. 3, pp. 267-277, 2014.

[13] D. Blei, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1020, 2003.

[14] Chor B, Horn D, Goldman N, Levy Y, Massingham T, "Genomic DNA k-mer spectra: models and modalities," *Genome biology*, vol. 10, no. (10):R108, 2009.

[15] Thomas L. Griffiths, Mark Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 5228-5235, 2004.

[16] K. S, Information Theory and Statistics, John Wiley & Sons, 1959.

[17] Zhang R, Cheng Z, Guan J, Zhou S, "Exploiting topic modeling to boost metagenomic reads binning," *BMC Bioinformatics*, vol. 16, no. (Suppl 5):S2, 2015.

[18] Zhou F, Olman V, Xu Y, "Barcodes for genomes and applications," *BMC Bioinformatics*, vol. 9, no. 546, 2008.

[19] David M. Blei, Andrew Y. Ng, Michael I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, no. 4, pp. 993-1022, 2012.

[20] Richter DC, Ott F, Auch AF, Schmid R, Huson DH, "Metasim – a sequencing simulator for genomics and metagenomics," *PloS ONE*, vol. 3, no. (10):e3373, 2008.

LATENT DIRICHLET ALLOCATION AS A PROBABILISTIC TOPIC MODEL IN COMBINATION WITH BIMETA ALGORITHM FOR METAGENOMIC BINNING

Van Dinh Vy Phuong, Tran Van Lang, Tran Van Hoai, Le Van Vinh

ABSTRACT*—Binning and taxanomical classification are two challenging problems in bioinformatics. The paper proposes a method using Laten Dirichlet Allocation to find hidden topics embedded as characteristic within genomic sequences. From these topics, the method classifies the group of sequences by using Kullback Leibler to calculate the similarity based on the distribution of topics instead of calculating directly from sequences. The proposed method is combined with BiMeta algorithm to create seed froup based on overlap information before using Latent Dirichlet Allocation to reduce the size of data to create model.*

Keywords*— Metagenomic; binning; reads; LDA.*