ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA

LÊ VĂN VINH

PHÂN LOẠI TRÌNH TỰ METAGENOMICS TRÊN CƠ SỞ PHÂN LỚP VÀ GOM CỤM

Chuyên ngành: Khoa học Máy tính
Mã số chuyên ngành: 62.48.01.01

TÓM TẮT LUẬN ÁN TIẾN SĨ KỸ THUẬT

TP. HỒ CHÍ MINH NĂM 2016

Công trình được hoàn thành tại Trường Đại học Bách Khoa - ĐHQG-HCM

Người hướng dẫn khoa học 1: PGS. TS. Trần Văn Lăng
Người hướng dẫn khoa học 2: PGS. TS. Trần Văn Hoài

Phản biện độc lập 1:
Phản biện độc lập 2:

Phản biện 1:
Phản biện 2:
Phản biện 3:

Luận án sẽ được bảo vệ trước Hội đồng chấm luận án họp tại
vào lúc giờ ngày tháng năm

Có thể tìm hiểu luận án tại thư viện:
- Thư viện Khoa học Tổng hợp Tp. HCM
- Thư viện Trường Đại học Bách Khoa – ĐHQG-HCM

DANH MỤC CÔNG TRÌNH ĐÃ CÔNG BỐ

Tạp chí:

[1].L. V. Vinh, T. V. Lang, and T. V. Hoai, "A novel semi-supervised algorithm for the taxonomic assignment of metagenomic reads," *BMC Bioinformatics*, vol.17, no.22, ISSN: 1471-2105, 2016 (***SCIE index, IF=2.435***).

[2].L. V. Vinh, T. V. Lang, L. T. Binh, and T. V. Hoai, "A two-phase binning algorithm using *l*-mer frequency on groups of non-overlapping reads," *Algorithms for Molecular Biology*, vol. 10, no.1, ISSN: 1748-7188, 2015 (***SCIE index, IF=1.439***).

[3].L. V. Vinh, T. V. Lang, and T. V. Hoai, "A novel l-mer counting method for abundance based binning of metagenomic reads." *Journal of Computer Science and Cybernetics*, vol. 10, no.3, ISSN 1813-9663, pp. 267-277, 2014.

[4].L. V. Vinh, T. V. Lang, and T. V. Hoai, "Hiệu năng của các giải pháp gom cụm trình tự metagenomic," *Tạp chí Khoa học và Công nghệ, Viện Hàn Lâm Khoa học và Công nghệ Việt Nam*, vol. 52, no.1B, ISSN: 0866-708X, pp.28-36, 2014.

Hội nghị:

[1].L. V. Vinh, T. V. Lang, and T. V. Hoai, "MetaAB-A Novel Abundance-Based Binning Approach for Metagenomic Sequences," In *Nature of Computation and Communication*, pp. 132-141, HCM city, Vietnam: Springer International Publishing, 2014.

[2].L. V. Vinh, D. H. Nhut, T. V. Lang, and T. V. Hoai, "A combination of genomic signatures for the binning of metagenomic sequences," *Proceedings of The 2nd International Conference on Green Technology*

and Sustainable Development, HCM City Oct 30-31, ISBN 978-604-732-817-8, pp. 662-668, 2014.

[3].L. V. Vinh, T. V. Lang, and T. V. Hoai, "An abundance-based binning approach for metagenomics read using a fuzzy k-medoids methods," *Proceeding of The 7th National Conference on Fundamental and Applied IT Research - FAIR'7*, Thai Nguyen, ISBN: 978-604-913-300-8, Natural Science and Technology Publishing House, 2014.

CHƯƠNG 1

GIỚI THIỆU

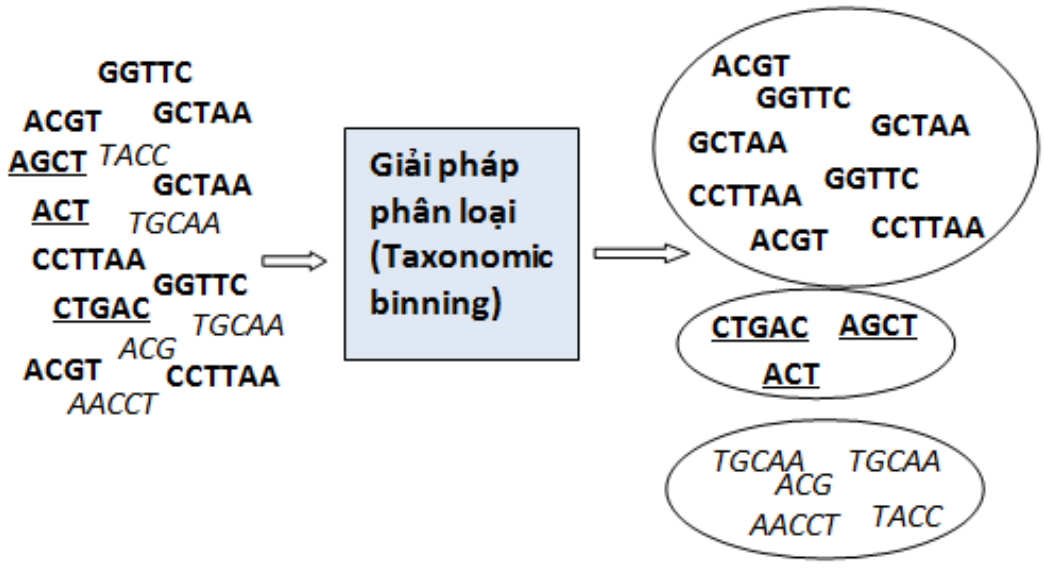
1.1. Metagenomics và bài toán phân loại trình tự

Metagenomics là lĩnh vực nghiên cứu cộng đồng vi sinh vật. Khác với phương pháp truyền thống, lĩnh vực này thực hiện phân tích trực tiếp trên mẫu thực nghiệm được thu thập từ môi trường mà không cần trải qua giai đoạn nuôi cấy và phân tách trong phòng thí nghiệm. Lĩnh vực metagenomics mang đến nhiều lợi ích trong y học, nông nghiệp, công nghệ sinh học, nghiên cứu năng lượng thay thế, hay môi trường [1].

Dữ liệu metagenomics thường không chứa trình tự của từng sinh vật riêng biệt. Chúng chứa trình tự thuộc nhiều loài khác nhau (có khi hơn 10.000 loài trong một mẫu [1]). Vì vậy, một trong những vấn đề quan trọng cần giải quyết là phân chia trình tự theo từng nhóm vi sinh vật, được gọi là bài toán phân loại trình tự metagenomics (taxonomic binning). Bài toán này được phát biểu như sau (theo Thomas và cộng sự [2]):

"*Phân loại trình tự metagenomics là quá trình sắp xếp trình tự DNA vào các nhóm bao gồm các trình tự thuộc cùng hệ gien của một cá thể hoặc hệ gien của các vi sinh vật có quan hệ gần nhau*".

Chẳng hạn, như minh họa ở hình 1.2. Tập dữ liệu bao gồm 16 trình tự DNA. Giải pháp phân loại giúp phân chia tập trình tự này vào 3 tập, mỗi tập chứa trình tự của một nhóm vi sinh vật.



Hình 1.2: Minh họa mục tiêu của bài toán phân loại trình tự metageonmic.

1.2. Vấn đề tồn tại cần giải quyết

1.2.1. Độ chính xác

Độ chính xác là một trong những khía cạnh quan trọng nhất cần được quan tâm của bài toán. Hai yếu tố chính ảnh hưởng đến chất lượng phân loại của các giải pháp hiện nay là độ dài trình tự ngắn (làm thiếu thông tin phân loại) và việc thiếu cơ sở dữ liệu tham khảo (làm giảm độ chính xác của các giải pháp phụ thuộc cơ sở dữ liệu tham khảo).

1.2.2. Chi phí tính toán

Chi phí tính toán là khía cạnh quan trọng khác cần được quan tâm bởi vì một dự án metagenomics thông thường cần phải phân tích một khối lượng dữ liệu rất lớn (có khi hàng trăm gigabase trình tự [3]), vốn đòi hỏi nhiều thời gian xử lý.

1.3. Mục tiêu của luận án

Mục tiêu của luận án là nhằm đề xuất giải pháp phân loại cho dữ liệu metagenomics, có khả năng xử lý tốt cho trình tự ngắn, và giải quyết hiệu quả cho trường hợp cơ sở dữ liệu tham khảo không đầy đủ.

1.4. Đóng góp của luận án

Những đóng góp chính của luận án bao gồm:

1.4.1. Về mặt khoa học

- Đề xuất mô hình thu giảm để tìm ước lượng khả năng cực đại của tham số cho mô hình thống kê về tần số xuất hiện l-mer, giúp giảm chi phí tính toán cho giải pháp phân loại dựa trên sự phong phú của hệ gien.
- Đề xuất phương pháp dự đoán số cụm trong tập dữ liệu sử dụng phương pháp lựa chọn mô hình cho vấn đề phân loại dựa trên sự phong phú của hệ gien.
- Cũng nhằm làm tăng chất lượng của giải pháp phân loại trình tự dựa trên sự phong phú của hệ gien, luận án đề xuất một phương pháp đếm l-mer với độ dài thay đổi giúp ước lượng mức độ phong phú của hệ gien hiệu quả hơn.
- Đề xuất ý tưởng chọn đại diện của tập trình tự thuộc cùng hệ gien dựa trên thông tin gối đầu trình tự. Tập này cho thấy có khả năng bảo toàn đặc trưng hợp thành và tương đồng chứa đựng trong tập dữ liệu gốc. Ý tưởng này có khả năng làm tăng chất lượng phân loại hay giảm chi phí tính toán cho các bài toán phân loại trình tự metagenomics.

1.4.2. Về mặt thực tiễn

Luận án đã đề xuất ba giải pháp phân loại trình tự metagenomics, bao gồm:

- Đề xuất giải pháp MetaAB và MetaAB-adv cho phép phân loại trình tự metagenomics dựa trên sự phong phú của hệ gien trong tập dữ liệu.
- Đề xuất giải pháp BiMeta cho phép phân loại trình tự metagenomics dựa trên đặc trưng hợp thành, không sử dụng cơ sở dữ liệu tham khảo.
- Đề xuất giải pháp SeMeta cho phép phân loại trình tự metagenomics có sử dụng cơ sở dữ liệu tham khảo.

1.5. Nội dung luận án

Cấu trúc của luận án bao gồm 7 chương. Chương 1 giới thiệu bài toán, trình bày những đóng góp và mục tiêu của luận án. Chương 2 trình bày nền tảng kiến thức cần thiết cho luận án và tình hình nghiên cứu hiện nay. Những phương pháp đóng góp cho vấn đề phân loại trình tự metagenomics dựa trên sự phong phú của hệ gien được trình bày trong chương 3. Chương 4 trình bày ý tưởng chọn tập đại diện của một tập trình tự dựa trên thông tin gối đầu sẽ được vận dụng ở hai chương tiếp theo của luận án. Chương 5 trình bày giải pháp phân loại không giám sát sử dụng đặc trưng dấu hiệu hệ gien và thông tin gối đầu giữa trình tự. Giải pháp phân loại bán giám sát SeMeta được trình bày trong chương 6 của luận án. Chương 7 là kết luận và hướng pháp triển. Phần phụ lục trình bày một số thông tin về dữ liệu được sử dụng trong các thực nghiệm được trình bày trong luận án, và một số kết quả thực nghiệm chi tiết.

CHƯƠNG 2

NỀN TẢNG KIẾN THỨC VÀ TÌNH HÌNH NGHIÊN CỨU

2.1. Nền tảng kiến thức

2.1.1. DNA và hệ gien

DNA (Deoxyribonucleic acid) là phân tử có cấu trúc ba chiều, bao gồm hai chuỗi đơn xoắn ốc, cuộn xung quanh một trục chung, tạo thành một chuỗi xoắn kép.

2.1.2. Công nghệ giải mã trình tự DNA

Giải mã trình tự DNA là quá trình xác định dãy các nucleotide trong trình tự đó. Các công nghệ giải mã được sử dụng phổ biến hiện nay như: 454 pyrose-quencing, Illumina Genome Analyzer, AB SOLiD, được gọi chung là công nghệ giải mã trình tự thế hệ tiếp theo (Next-generation sequencing). Vì mẫu DNA cần được giải mã trong thực tế thường rất dài, trong khi các máy giải mã

chỉ cho phép giải mã cho trình tự có kích thước ngắn. Vì vậy, kỹ thuật nền tảng được sử dụng cho các công nghệ này là kỹ thuật giải mã trình tự đoạn ngắn (shotgun sequencing). Kỹ thuật này thực hiện nhân bản và cắt ngẫu nhiên mẫu DNA thành những mảnh nhỏ (fragment) có độ dài phù hợp cho từng công nghệ giải mã. Máy giải mã trình tự xử lý cho từng mảnh DNA nhỏ và thông tin được lưu trữ trên máy tính được gọi là trình tự (read/sequence).

2.1.3. Đặc trưng sử dụng cho phân loại trình tự

Một giải pháp phân loại trình tự cần một phép đo mức độ giống nhau hay khoảng cách giữa các trình tự. Phép đo đó có thể được thực hiện nhờ sử dụng một số đặc trưng sau.

2.1.3.1. Tính tương đồng giữa các trình tự

Mức độ tương đồng (homology) giữa hai trình tự được tính dựa trên việc so sánh sự giống nhau tương ứng giữa các nucleotide trên hai trình tự. Hai cá thể sinh vật chứa trình tự có mức độ tương đồng cao thể hiện chúng có quan hệ sinh loài (phylogenetic relationship) gần nhau và có cùng tổ tiên. Ngược lại, mức độ tương đồng thấp thể hiện chúng có quan hệ sinh loài xa nhau [4].

2.1.3.2. Dấu hiệu hệ gien

Dấu hiệu hệ gien (genomic signature) là cấu trúc toán học đặc trưng theo loài mà có thể xây dựng từ một trình tự sinh học. Dấu hiệu hệ gien của trình tự cùng loài giống nhau nhiều hơn so với của trình tự thuộc hai loài khác nhau, và hai loài gần nhau có dấu hiệu hệ gien của trình tự giống nhau nhiều hơn so với giữa hai loài xa nhau [5]. Nhờ tính chất này mà dấu hiệu hệ gien có thể được sử dụng cho việc phân loại trình tự. Nhiều dấu hiệu hệ gien đã được nghiên cứu như: GC-content, dấu hiệu dựa trên tần số xuất hiện l-mer (đoạn trình tự ngắn có độ dài là l, thường được gọi là oligonucleotide), dấu hiệu dựa trên mô hình Markov.

2.1.3.3. Một số đặc trưng khác

Một số đặc trưng khác được rút trích ra từ sự quan sát dữ liệu metagenomics và áp dụng cho bài toán phân loại như sau:

- ***Tính duy nhất của đoạn trình tự l-mer trong tập dữ liệu***: *Hầu hết các l-mer (đoạn trình tự ngắn, có độ dài là l) không được chia sẻ bởi các hệ gien khác nhau khi l đủ lớn [6].*

- ***Sự phong phú của hệ gien trong tập dữ liệu:*** *Trong một tập trình tự metagenomics, tần số xuất hiện của l-mer thuộc cùng một hệ gien tỉ lệ thuận với sự phong phú của hệ gien đó [7].*

2.1.4. Phân lớp và gom cụm dữ liệu

2.1.4.1. Phân lớp dữ liệu

Phân lớp dữ liệu (classification) là quá trình nhằm sắp xếp các đối tượng dữ liệu vào các lớp (classes) đã biết. Các giải pháp phân lớp dữ liệu thường dựa trên hai phương pháp học chính: học có giám sát (supervised learning) và học bán giám sát (semi-supervised learning). Trong khi phương pháp học có giám sát chỉ sử dụng thông tin từ tập dữ liệu tham khảo cho việc gán nhãn dữ liệu, thì phương pháp học bán giám sát cho phép sử dụng kết hợp thông tin rút trích từ tập trình tự đang được phân tích và tập dữ liệu tham khảo. Trong luận án này, phương pháp bán giám sát gom cụm và gán nhãn (cluster-and-label) được nghiên cứu và sử dụng.

2.1.4.2. Gom cụm dữ liệu

Gom cụm dữ liệu là một hình thức của phương pháp học không có giám sát, nhằm phân chia các đối tượng dữ liệu vào các cụm, sao cho các đối tượng có đặc tính giống nhau thuộc cùng một cụm và các đối tượng có đặc tính khác nhau thuộc về các cụm khác nhau. Luận án này sử dụng hai phương pháp

gom cụm là *k*-means và phương pháp dựa trên mô hình (dùng thuật toán EM - Expectation Maximization).

2.1.5. Độ đo hiệu năng giải pháp phân loại

Phần này trình bày các độ đo được sử dụng đánh giá chất lượng của các giải pháp phân loại. Ba độ đo *độ chính xác* (*precision*), *độ nhạy* (*recall* hay *sensitivity*), và *F-measure* được sử dụng chung cho việc đánh giá.

2.2. Tình hình nghiên cứu

Những hướng tiếp cận chính của bài toán như sau.

2.2.1. Phương pháp có giám sát

Theo hướng tiếp cận này, trình tự DNA được phân loại dựa trên mức độ tương đồng trình tự hay mức độ giống nhau giữa dấu hiệu hệ gien của chúng với hệ gien hay trình tự của sinh vật đã biết trong cơ sở dữ liệu tham khảo. Có thể chia các giải pháp có giám sát thành ba nhóm như sau.

2.2.1.1. Phương pháp dựa trên tính tương đồng

Trình tự metagenomics được phân loại dựa trên việc so sánh để tìm ra mức độ tương đồng với trình tự trong ngân hàng gien hoặc protein. Trong các giải pháp theo hướng này, công việc so sánh tương đồng thường được thực hiện bởi các công cụ đã có sẵn như BLAST hay BLAT. Một số giải pháp thuộc nhóm này như: MEGAN, SOrt-ITEMS, và CARMA3.

2.2.1.2. Phương pháp dựa trên tính hợp thành

Phương pháp này sử dụng dấu hiệu hệ gien (genomic signature) được rút trích từ hệ gien hay trình tự tham khảo để phân loại. Một số dấu hiệu hệ gien thường được sử dụng như: GC-content, tần số xuất hiện *l*-mer. Hầu hết các giải pháp thuộc nhóm này như TACOA, TAC-ELM, AKE chỉ phù hợp cho xử lý trình tự dài. Một số nghiên cứu gần đây như MetaCV, MetaID hướng đến việc xử lý cho trình tự ngắn.

2.2.1.3. Phương pháp lai

Nhóm phương pháp lai sử dụng điểm mạnh từ sự kết hợp tính tương đồng và tính hợp thành nhằm giảm chi phí tính toán, hay cải tiến chất lượng phân loại. Một số giải pháp thuộc nhóm này như: SPHINX, MetaCluster-TA và PhymmBL.

2.2.2. Phương pháp không có giám sát

Theo hướng tiếp cận này, việc phân loại chỉ dựa trên thông tin được rút trích từ chính tập dữ liệu đang được phân tích, mà không sử dụng thông tin từ bên ngoài. Các giải pháp đã được đề xuất có thể được phân chia thành hai nhóm: giải pháp dựa trên tính hợp thành (composition feature) và giải pháp dựa trên sự phong phú của hệ gien (genome abundance-based feature).

2.2.2.1. Phương pháp dựa trên tính hợp thành

Nhóm giải pháp theo hướng tiếp cận này phân loại trình tự dựa trên dấu hiệu hệ gien được rút trích từ trình tự đang được xử lý. Một số giải pháp chỉ có khả năng phân loại tốt cho trình tự dài như: LikelyBin, Scimm, MetaCluster 2.0, MetaCluster 3.0. Một số khác có khả năng xử lý tốt hơn cho trình tự ngắn như: TOSS, MetaCluster 5.0 và MCluster.

2.2.2.2. Phương pháp dựa trên sự phong phú hệ gien

Một số giải pháp không có giám sát được đề xuất gần đây có thể phân loại trình tự ngắn sử dụng đặc trưng sự phong phú của hệ gien trong tập trình tự metagenomics. Trong số các giải pháp này, AbundanceBin phân loại dựa trên việc sử dụng giải pháp EM (expectation-maximization) nhằm ước lượng tham số của mô hình xác suất của l-mer trong trình tự.

2.2.3. Phương pháp bán giám sát

Phương pháp bán giám sát là một dạng phối hợp giữa kỹ thuật có giám sát và không giám sát nhằm đạt được chất lượng phân loại tốt hơn. Những nghiên

cứu gần đây theo hướng tiếp cận này như RAIphy, CompostBin. MetaCluster-TA cũng có thể được xếp vào nhóm phương pháp này.

CHƯƠNG 3

GIẢI PHÁP PHÂN LOẠI KHÔNG GIÁM SÁT DỰA TRÊN SỰ PHONG PHÚ CỦA HỆ GIEN

3.1. Giới thiệu

Luận án này đề xuất một phương pháp gom cụm dựa trên mô hình, được gọi là MetaAB, có khả năng phân loại trình tự một cách hiệu quả dựa trên thông tin sự phong phú của hệ gien trong tập trình tự cần phân tích. Phương pháp đề xuất sử dụng mô hình thu giảm để tìm ước lượng khả năng cực đại (MLE - maximum likelihood estimates) của tham số trong mô hình xác suất, nhằm giảm chi phí tính toán so với các giải pháp tương tự. Ngoài ra, MetaAB vận dụng một kỹ thuật lựa chọn mô hình xác suất nhằm phân loại và ước lượng số cụm dữ liệu toàn cục một cách hiệu quả. Bên cạnh đó, một phương pháp đếm tần số xuất hiện l-mer có độ dài thay đổi cũng được đề xuất trong nghiên cứu này nhằm làm tăng sự chính xác trong việc phân loại.

3.2. Phương pháp

3.2.1. Mô hình hỗn hợp của tần số xuất hiện các l-mer

Cho một tập trình tự metagenomics bao gồm n trình tự $R = \{r_1, r_2, \ldots, r_n\}$. Đặt $w_1, \ldots, w_q$ là một tập các l-mer trong tập trình tự, và $c(w_i), 1 \leq i \leq q$, là số lần xuất hiện của l-mer w_i trong tập dữ liệu. Vì mỗi l-mer được hình thành từ 4 nucleotide (A, C, G, T), ta có: $q \leq 4^l$. Như vậy, ta có một tập dữ liệu $\mathbb{X} = \{c(w_1), \ldots, c(w_q)\}$ bao gồm q quan sát của biến ngẫu nhiên $x = c(w_i), 1 \leq i \leq q$. Hàm log-likelihood tương ứng với mô hình hợp k thành phần của dữ liệu

này như sau:

$$\log \mathscr{L}(\Theta|\mathbb{X}) = \sum_{i=1}^{q} \log \left(\sum_{m=1}^{k} \alpha_m p_m(c(w_i)|\lambda_m) \right). \tag{3.1}$$

Trong đó, $\Theta = \{\alpha_1, \ldots, \alpha_k, \theta_1, \ldots, \theta_k\}$ là một tập các tham số của mô hình hợp này. $\alpha_1, \ldots, \alpha_k$ là các thành phần hợp và thỏa mãn điều kiện $\sum_{m=1}^{k} \alpha_m = 1, \alpha_m \geq 0$. Ngoài ra, $\theta_m, 1 \leq m \leq k$, là tập tham số của thành phần thứ m của mô hình. Trong ngữ cảnh này, với mô hình hợp Poisson, ta có: $\theta_m \equiv \lambda_m$. Giải pháp đề xuất nhằm tìm ước lượng khả năng cực đại (MLE) của tham số Θ, vốn thể hiện khả năng cao nhất mà các l-mer thuộc về các hệ gien trong tập dữ liệu.

$$\Theta^* = \arg\max_{\Theta} \log \mathscr{L}(\Theta|\mathbb{X}). \tag{3.2}$$

3.2.2. Mô hình thu giảm

Nhằm giảm chi phí tính toán của việc ước lượng tham số trong mô hình, nghiên cứu này đề xuất một mô hình thu giảm của nó. Bởi vì, hai l-mer có cùng số lần xuất hiện luôn có cùng xác suất thuộc về các thành phần trong mô hình. Vì vậy, hàm log-likelihood tương ứng với mô hình hợp k thành phần trên, được phát biểu trong biểu thức 3.1, có thể được xây dựng lại như sau:

$$\log \mathscr{L}(\Theta|\mathbb{X}) = \sum_{t=1}^{b} s_t \log \left(\sum_{m=1}^{k} \alpha_m p_m(c_t|\lambda_m) \right). \tag{3.3}$$

Với b là số nhóm l-mer mà có cùng số lần xuất hiện, s_t là số lần xuất hiện của l-mer trong nhóm t, và

$$q = \sum_{t=1}^{b} s_t. \tag{3.4}$$

Trong thực tế, một tỉ lệ lớn các l-mer xuất phát từ cùng hệ gien và thường có cùng số lần xuất hiện trong tập trình tự metagenomics (tức là $s_t \gg 1$). Vì vậy, khi sử dụng biểu thức 3.3, chi phí để tìm ước lượng khả năng cực đại của tham số Θ giảm đi đáng kể so với mô hình gốc trong 3.1.

3.2.3. Ước lượng tham số trong mô hình đề xuất

Để ước lượng khả năng cực đại của tham số trong mô hình đề xuất, nghiên cứu này sử dụng giải thuật cực đại hóa kỳ vọng (EM - Expectation Maximization [8]). Đây là một giải thuật lặp, cho phép tìm được giá trị tối ưu cục bộ của tham số trong mô hình xác suất. Mỗi vòng lặp thực thi hai bước sau (phần dưới đây thể hiện cho vòng lặp thứ $s+1$):

+ Bước kỳ vọng hóa (E-step): Tính xác suất của các l-mer mà số lần xuất hiện của chúng bằng $c_t, t \in \{1,\ldots,b\}$, thuộc về thành phần thứ m, cho trước tham số $\Theta^{(s)}$, và c_t.

$$p(z_{tm}=1|c_t,\Theta^{(s)}) = \frac{\alpha_m^{(s)} p_m(c_t|\lambda_m^{(s)})}{\sum_{v=1}^{k} \alpha_v^{(s)} p_v(c_t|\lambda_v^{(s)})} \text{(luật Bayes).} \tag{3.5}$$

+ Bước cực đại hóa (M-step): Trong bước này, các tham số được cập nhật theo biểu thức sau:

$$\Theta^{(s+1)} = \arg\max_{\Theta} Q(\Theta,\Theta^{(s)}). \tag{3.6}$$

Trong đó, hàm Q: $Q(\Theta,\Theta^{(s)}) = E[log(p(\mathbb{X},\mathbb{Z}|\Theta))|\mathbb{X},\Theta^{(s)}]$ là kỳ vọng của log-likelihood của dữ liệu đầy đủ. Với $\mathbb{Z}$ là dữ liệu cho biết thành phần nào tạo ra các l-mer. Khi các tham số trong mô hình hợp này đã được ước lượng, mỗi trình tự r_j được gán vào các thành phần (hay cụm) dựa trên xác suất các l-mer của chúng thuộc về các thành phần.

3.2.4. Ước lượng số cụm sử dụng BIC

Luận án này vận dụng phương pháp lựa chọn mô hình (model selection) BIC (Bayesian Information Criterion) nhằm tìm số thành phần của một mô hình hỗn hợp. Điều này đồng nghĩa với việc có thể ước lượng được số cụm trong tập dữ liệu. Cụ thể, giá trị BIC của mô hình m thành phần như sau:

$$\text{BIC}_m = logp(\mathbb{X}|D_m) = log\mathscr{L}(\Theta_m^*|\mathbb{X}) - \frac{d}{2}log(q). \tag{3.7}$$

Mô hình được chọn là mô hình có giá trị BIC lớn nhất.

3.2.5. Thuật toán MetaAB

Thuật toán MetaAB thực hiện các công việc như sau:

+ Tính số lần xuất hiện l-mer trong tập R.

+ Loại bỏ l-mer không tin cậy.

+ Thực thi vòng lặp thực hiện giải thuật EM với số thành phần thay đổi, và tính giá trị BIC_m.

+ Chọn mô hình có giá trị BIC lớn nhất.

3.2.6. Phương pháp đếm l-mer với độ dài thay đổi

Phương pháp đếm l-mer được trình bày trong phần này nhằm giải quyết hạn chế của phương pháp đếm l-mer có độ dài cố định, giúp cho việc tính tần số xuất hiện l-mer một cách đúng đắn, và phản ánh chính xác hơn mức độ phong phú của hệ gien chứa chúng.

3.2.6.1. Phương pháp đề xuất

Một l-mer có độ dài không cố định (với độ dài tối đa là l) được định nghĩa là một tập gồm ba phần: pre-l-mer, main-l-mer, và suf-l-mer. main-l-mer là thành phần giữa của một l-mer, và độ dài của nó được gán cố định bởi giá tri l_m. pre-l-mer và suf-l-mer là hai phần còn lại, nằm ở vị trí đầu và cuối của một l-mer. Độ dài của hai phần này không cố định và được giới hạn bởi giá trị l_p, và l_s (với $l = l_p + l_m + l_s$). Hai l-mer được so sánh như sau:

Đặt $u = \langle p(u), m(u), s(u) \rangle$ là một l-mer. $p(u), m(u)$ và $s(u)$ là pre-l-mer, main-l-mer, và suf-l-mer tương ứng của l-mer u. Chúng là các chuỗi chứa ký tự trong một tập {A, C, G, T}. Đặt $|p(u)|$, $|s(u)|$ là độ dài tương ứng của $p(u), s(u)$ ($|p(u)| \leq l_p, |s(u)| \leq l_s$). Đặt $v = \langle p(v), m(v), s(v) \rangle$ là một l-mer khác. Ký hiệu $g(s, pos, len)$ là hàm để sao chép một chuỗi con của chuỗi s từ vị trí bắt đầu

pos và trượt qua *len* ký tự. $u = v$ khi và chỉ khi:

$$\begin{cases} m(u) = m(v) \text{ và} \\ s(u) = g(s(v), 1, |s(u)|), \text{ nếu } |s(u)| \leq |s(v)| \text{ và} \\ s(v) = g(s(u), 1, |s(v)|), \text{ nếu } |s(v)| < |s(u)| \text{ và} \\ p(u) = g(p(v), |p(v)| - |p(u)| + 1, |p(u)|), \text{ nếu } |p(u)| \leq |p(v)| \text{ và} \\ p(v) = g(p(u), |p(u)| - |p(v)| + 1, |p(v)|), \text{ nếu } |p(v)| < |p(u)|. \end{cases} \tag{3.8}$$

3.3. Kết quả thực nghiệm

Trong phần này, hai phiên bản của giải pháp đề xuất được thực nghiệm. Phiên bản một tên là MetaAB sử dụng phương pháp đếm l-mer có độ dài cố định. Phiên bản hai tên là MetaAB-adv (viết tắt của MetaAB-advanced) sử dụng phương pháp đếm l-mer có độ dài thay đổi được đề xuất trong nghiên cứu này. Kết quả thực nghiệm cho thấy MetaAB và MetaAB-adv đạt chất lượng phân loại tốt hơn trong phần lớn trường hợp thực nghiệm so với AbundanceBin. MetaAB đòi hỏi chi phí tính toán thấp nhất trong số các giải pháp thực nghiệm. MetaAB-adv đạt kết quả tốt hơn so với MetaAB trong trường hợp trình tự không có lỗi giải mã.

CHƯƠNG 4

CHỌN ĐẠI DIỆN CỦA MỘT TẬP TRÌNH TỰ DỰA TRÊN TÍNH CHẤT GỐI ĐẦU

4.1. Giới thiệu

Luận án này đề xuất ý tưởng chọn đại diện cho một tập trình tự DNA dựa trên tính gối đầu giữa các trình tự. Việc lựa chọn tập đại diện là nhằm giảm chi phí tính toán, đồng thời giảm nhiễu trong dữ liệu do độ phủ của tập dữ liệu không đồng nhất để đạt được hiệu quả phân loại trình tự tốt hơn.

4.2. Định nghĩa bài toán

4.2.1. Một số ký hiệu và khái niệm

- Cho hai trình tự DNA r và s. Nếu r và s được lấy mẫu từ cùng hệ gien, ta ký hiệu là $r \bowtie s$.
- Ta ký hiệu *r gối đầu s* là $r \sqcap s$. Ta cũng ký hiệu *r không gối đầu s* là $r \not\sqcap s$

4.2.2. Tính chất của tập đại diện

Cho một tập trình tự G, sao cho $\forall r, s \in G, r \bowtie s$. Ta định nghĩa một tập đại diện của G, được ký hiệu là $S(G)$, là tập được xây dựng sao cho thỏa tính chất sau:

i) $S(G) \subseteq G$

ii) $\forall r, s \in S(G), r \not\sqcap s$

4.2.3. Định nghĩa bài toán tìm tập đại diện

Cho một đồ thị không có trọng số $D = (V, E)$. Trong đó, V là một tập gồm $|V|$ đỉnh thể hiện cho các trình tự trong tập G, và E là một tập các cạnh. Mỗi cạnh $(r, s), r, s \in V$, thể hiện mối quan hệ $r \sqcap s$.

Một điều có thể thấy rằng, tập đại diện $S(D)$ của D tương đương với một *tập độc lập (independent set)* hay *tập ổn định (stable set)* của một đồ thị mà trong đó không có đỉnh nào kề nhau. Bài toán tìm tập đại diện của một tập trình tự là bài toán tìm tập độc lập lớn nhất (maximum independent set) của một đồ thị, được định nghĩa như sau:
Đặt $x_r = 1$ nếu $r \in S(D)$. Ngược lại, $x_r = 0$ nếu $r \notin S(D)$. Mục tiêu của bài toán là tìm tập $S(D) \subset D$ nhằm:

$$\text{maximize } f(x) = \sum_{r=1}^{|V|} x_r, \tag{4.1}$$

sao cho thỏa mãn các ràng buộc sau: $x_r + x_s \leq 1, \forall (r, s) \in E$, và $x_r \in \{0, 1\}, \forall r \in V$.

4.3. Sự bảo toàn đặc trưng của nhóm trình tự

Một vấn đề chính cần được quan tâm là khả năng bảo toàn đặc trưng của đại diện của tập dữ liệu. Cụ thể, hai đặc trưng chính được sử dụng trong nghiên cứu này là tính tương đồng và tính hợp thành dựa trên tần số xuất hiện l-mer. Đặc trưng của tập dữ liệu sẽ được bảo toàn trong tập đại diện nếu tập đại diện có khả năng phủ hết các vị trí trên hệ gien gốc mà tập dữ liệu đó phủ. Nhờ sử dụng tính chất không gối đầu (non-overlapping), các trình tự trong tập đại diện có xu hướng phủ hầu hết các vị trí trên hệ gien gốc, qua đó có thể bảo toàn phần lớn đặc trưng của tập dữ liệu ban đầu. Một số thực nghiệm đã được thực hiện trong luận để kiểm chứng cho khả năng bảo toàn đặc trưng này.

CHƯƠNG 5

GIẢI PHÁP PHÂN LOẠI KHÔNG GIÁM SÁT SỬ DỤNG DẤU HIỆU HỆ GIEN

5.1. Giới thiệu

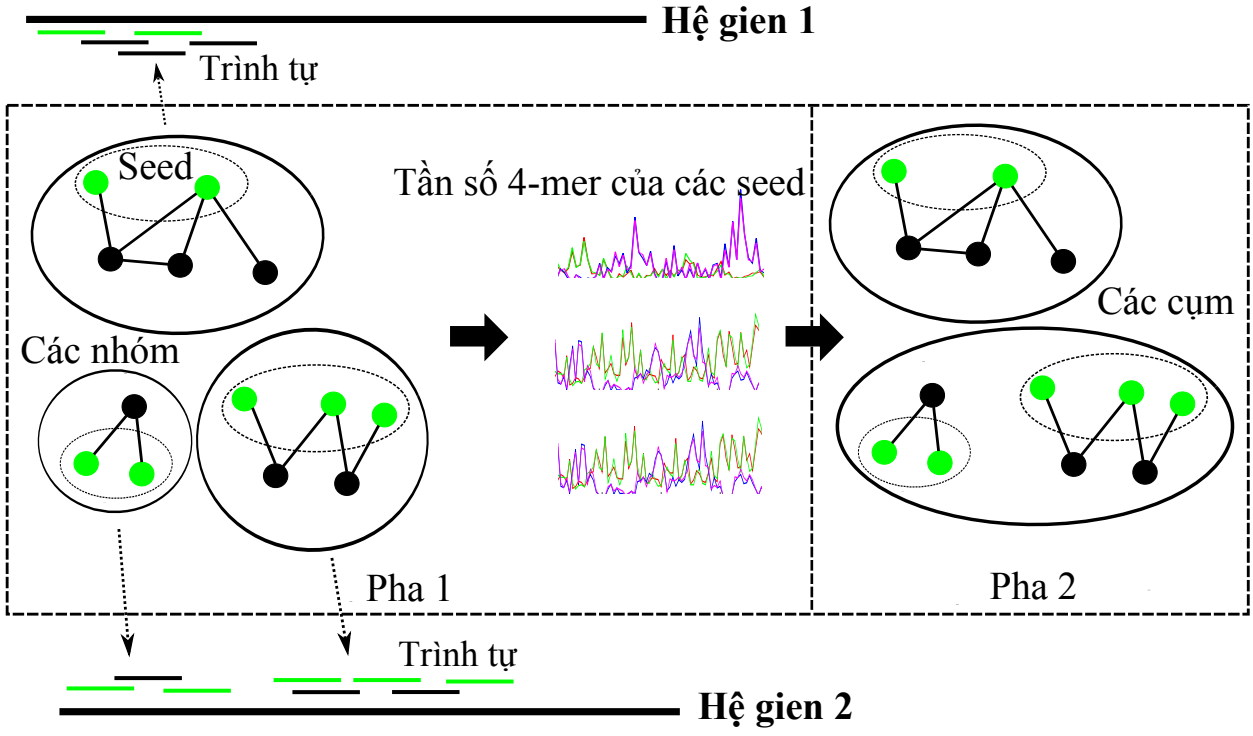
Nghiên cứu này đề xuất một giải pháp phân loại trình tự metagenomics, được gọi là BiMeta, sử dụng kỹ thuật gom cụm, và dựa trên dấu hiệu hệ gien của nhóm trình tự không gối đầu (non-overlapping reads). Giải pháp BiMeta vận dụng ý tưởng sử dụng tập đại diện của tập trình tự thuộc cùng hệ gien được trình bày ở chương 4 nhằm làm tăng chất lượng phân loại, cũng như giảm chi phí tính toán.

5.2. Phương pháp

5.2.1. Nền tảng của phương pháp đề xuất

Giải pháp đề xuất gồm hai pha như sau (hình 5.1): Đặt R là một tập n trình tự metagenomics. Trong pha 1, trình tự được gom vào các nhóm $G_i, i \in \{1, \ldots, p\}$ và $p \leq n$, dựa trên thông tin gối đầu trình tự. Nói một cách khác, hai

trình tự $r, s \in R$ có thể được gom vào cùng nhóm nếu chúng được kết luận là $r \sqcap s$. Điều này có nghĩa là các trình tự $r, s \in R$ ở trong cùng nhóm được xem như thuộc cùng hệ gien ($r \bowtie s$). Để trộn các nhóm này vào các cụm mà có thể thể hiện hệ gien của các sinh vật có quan hệ sinh loài gần nhau, phương pháp đề xuất tính vectơ tần số l-mer $\mathbf{f}$ cho mỗi nhóm G_i. Luận án sử dụng tập đại diện của mỗi nhóm G_i thay vì G_i nhằm giảm thiểu sự mất cân bằng trong độ phủ của tập trình tự, cũng như giảm chi phí rút trích thông tin từ các nhóm. Tập đại diện này chỉ bao gồm các trình tự không gối đầu (như được trình bày ở chương 4), và được gọi là một *seed* của G_i. Trong pha 2, phương pháp đề xuất nhằm mục tiêu trộn các nhóm $G_i, i \in \{1, \ldots, p\}$, vào k cụm ($k \leq p$) sử dụng vectơ $\mathbf{f}$ của đại diện của các nhóm.



Hình 5.1: Quá trình phân loại của BiMeta.

Xác định các trình tự gối đầu và không gối đầu

Cho trước $m, q \in \mathbb{N}$, nếu r và s chia sẻ ít nhất m q-mer, chúng được xem như là gối đầu nhau. Ngược lại, chúng không gối đầu nhau.

5.2.2. Thuật toán BiMeta

5.2.2.1. Pha 1 - Gom nhóm các đỉnh và xây dựng seed

Pha này thực hiện các công việc:

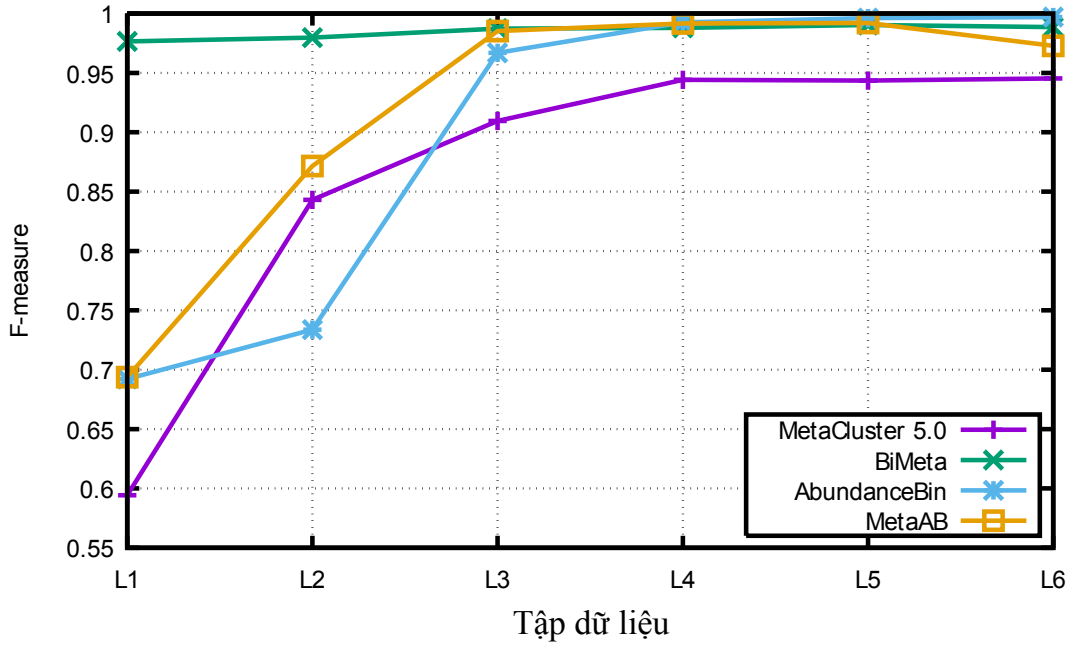
+ Xây dựng các nhóm và xây dựng seed của chúng sử dụng một thuật toán tham lam.

+ Tính vectơ tần số của đại diện của các nhóm.

5.2.2.2. Pha 2 - Trộn các nhóm

Trong pha này, giải thuật gom cụm *k*-means được sử dụng để trộn các nhóm vốn được tạo trong pha 1, thành các cụm.

5.3. Kết quả thực nghiệm



Hình 5.3: Hiệu năng của MetaCluster 5.0, BiMeta, AbundanceBin và MetaAB trên các tập dữ liệu từ L1 đến L6.

BiMeta được so sánh với các giải pháp MetaCluster 5.0, AbundanceBin, MetaCluster 2.0 và MetaAB. Kết quả thực nghiệm cho thấy BiMeta đạt chất lượng phân loại tốt hơn so với các giải pháp còn lại trong hầu hết trường hợp thực nghiệm (chẳng hạn như kết quả ở hình 5.3), và tốn ít chi phí trên các tập

dữ liệu được khảo sát. BiMeta cũng cho thấy có thể phân tích tốt cho trình tự có độ dài khác nhau, và trên các bộ dữ liệu có các mức độ phong phú khác nhau. Ngoài ra, BiMeta cũng đạt giá trị F-measure cao hơn so với MetaCluster 2.0 khi xử lý trên bộ dữ liệu thực AMD (Acid Mine Drainage).

CHƯƠNG 6

GIẢI PHÁP PHÂN LOẠI BÁN GIÁM SÁT SỬ DỤNG ĐẶC TRƯNG KẾT HỢP

6.1. Giới thiệu

Chương này trình bày một giải pháp phân loại trình tự metagenomics mới, sử dụng phương pháp phân lớp bán giám sát, được gọi là SeMeta. Ý tưởng tìm tập đại diện của tập trình tự cũng được áp dụng nhằm giúp giải pháp này đạt được tốc độ xử lý nhanh, trong khi vẫn bảo toàn chất lượng phân loại như trường hợp sử dụng toàn bộ tập trình tự.

6.2. Phương pháp

6.2.1. Nền tảng của phương pháp đề xuất

Cho một tập R gồm n trình tự metagenomics. Bước đầu tiên của giải pháp đề xuất là nhằm phân chia n trình tự vào k tập $C_1, C_2, \ldots, C_k, k \leq n$. Ở bước thứ hai, mỗi cụm $C_i, 1 \leq i \leq k$, được gán nhãn dựa trên việc so sánh tương đồng giữa trình tự trong cụm với trình tự tham khảo. Một trong những ý tưởng được áp dụng trong nghiên cứu này là việc sử dụng tập đại diện của cụm như được trình bày ở chương 4. Thay vì tìm kiếm tương đồng cho tất cả trình tự trong các cụm $C_i, 1 \leq i \leq k$, giải pháp này chỉ thực hiện trên đại diện $S(C_i)$ của chúng.

Trong bước gán nhãn cho cụm, một kỹ thuật lọc hai mức (two-level filtering) được đề xuất nhằm loại bỏ những BLAST hit (tên hệ gien tham khảo được trả về bởi công cụ so sánh tương đồng BLAST). Mức một (mức trình tự) lọc

những BLAST hit có giá trị bit-score thấp cho từng trình tự bằng việc sử dụng hai ngưỡng min-score (loại bỏ những hit có bit-score thấp) và top-percent (lựa chọn và giữ lại những hit có bit-score cao hơn phần còn lại). Mức hai (mức cụm) tiếp tục loại bỏ những hit không tin cậy nhờ thông tin tương đồng kết hợp của các trình tự trong từng cụm.

6.2.2. Thuật toán SeMeta

Hình 6.3 thể hiện quá trình thực hiện của phương pháp này, bao gồm hai bước chính: Gom cụm (Clustering), và Gán nhãn sinh học (Taxonomic Assignment).

6.2.2.1. Bước 1: Gom cụm

Trong bước này, trình tự được phân loại vào các cụm chứa sinh vật có mối quan hệ sinh loài gần nhau, sử dụng phiên bản cải tiến của giải pháp BiMeta được đề xuất ở chương 5. Điểm khác biệt của SeMeta và BiMeta trong bước gom cụm này là: (1) SeMeta loại bỏ những nhóm có kích thước nhỏ nhằm nâng cao độ chính xác; (2) SeMeta có khả năng phát hiện tự động số cụm dữ liệu.

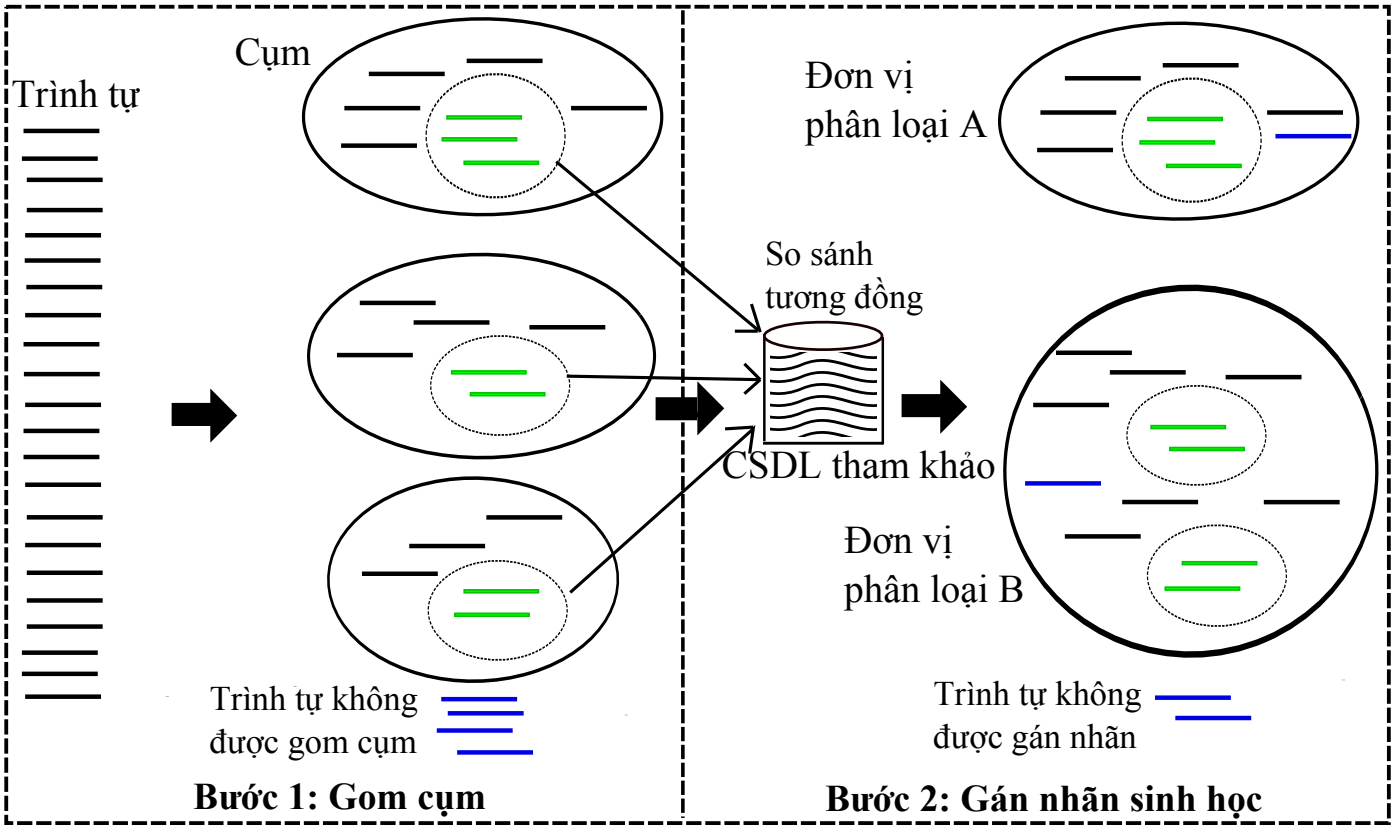
Xây dựng đại diện của cụm

Sau khi trình tự được chia vào k cụm $C_1, \ldots, C_k$, đại diện của các cụm được xây dựng dựa trên thông tin gối đầu giữa các trình tự. Nhằm cố gắng gán nhãn cho những trình tự đã bị loại bỏ khỏi quá trình gom cụm ở bước 1, SeMeta xem các trình tự này như những cụm và đưa vào bước gán nhãn sinh học cho các cụm.

6.2.2.2. Bước 2: Gán nhãn sinh học

Bước này bao gồm ba công việc chính:

- *Công việc 1 - Tìm kiếm tương đồng*: Thực hiện so sánh tương đồng của trình tự trong đại diện của các cụm với cơ sở dữ liệu tham khảo.
- *Công việc 2 - Gán nhãn cho cụm*: SeMeta thực thi một kỹ thuật lọc ở hai mức sau:



Hình 6.3: Quá trình thực hiện của SeMeta.

+ *Mức trình tự*: Sử dụng hai tham số min-score s_{min} và top-percent p_{top}.
+ *Mức cụm*: Mức này sử dụng ngưỡng max-occur o_{max} để loại bỏ thêm những hit không tin cậy.

Cuối cùng, giải thuật LCA (Lowest Common Ancestor) được sử dụng để tìm đơn vị phân loại chung thấp nhất của những hit còn lại sau giai đoạn lọc.

- *Công việc 3 - Hậu xử lý*: Giai đoạn này thực hiện trộn các cụm mà được gán cùng đơn vị phân loại vào cùng một cụm, và xác định những trình tự không được gán nhãn.

6.3. Kết quả thực nghiệm

SeMeta được so sánh với hai giải pháp dựa trên tính tương đồng thường được sử dụng hiện nay là MEGAN và SOrtITEMS. Thực nghiệm này đánh giá ở cả hai khía cạnh sau: (1) Khả năng gán nhãn vào một nhóm (clade) trên cây sinh loài; (2) Khả năng gán nhãn chính xác vào một vị trí trên cây sinh loài. Hai kịch bản cơ sở dữ liệu được tạo ra là: (1) Loài đã biết (vi sinh vật trong dữ liệu cần phân tích có trong cơ sở dữ liệu tham khảo); (2) Loài chưa biết (vi sinh vật trong cơ sở dữ liệu cần phân tích không có trong cơ sở dữ liệu tham khảo).

Kết quả thực nghiệm cho thấy, SeMeta đạt chất lượng phân loại tốt hơn hai giải pháp còn lại trong phần lớn trường hợp thực nghiệm, đặc biệt là khi xét ở bậc phân loại thấp (mức loài, mức chi), và ở kịch bản cơ sở dữ liệu loài chưa biết. Điểm nổi bật của SeMeta là cần chi phí tính toán ít hơn nhiều (chẳng hạn, ít hơn 5.6 lần cho bộ dữ liệu *ds*2) so với MEGAN và SOrtITEMS. Ngoài ra, SeMeta có khả năng xử lý tốt cho hai bộ dữ liệu thực AMD (Acid Mine Drainage) và HGM (Human Gut Metagenome).

CHƯƠNG 7

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

7.1. Kết luận

Lĩnh vực nghiên cứu metagenomics mở ra cơ hội lớn giúp con người hiểu hơn về cộng đồng vi sinh vật, và có thể mang đến nhiều lợi ích thiết thực cho cuộc sống. Mặc dù vậy, vấn đề phân tích dữ liệu metagenomics có nhiều thách thức lớn bởi sự phức tạp và đa dạng sinh học của môi trường vi sinh vật. Ba yếu tố chính làm cho việc phân tích trình tự trở nên khó khăn, bao gồm: phần lớn vi sinh vật chưa được khám phá; sự hạn chế của công nghệ giải mã trình tự, dẫn đến trình tự được tạo ra có kích thước ngắn; và dữ liệu cần phân tích lớn. Điều này đòi hỏi những công cụ phân tích dữ liệu hiệu quả góp phần thành công cho các dự án metagenomics.

Luận án này đã đề xuất các giải pháp phân loại trình tự metagenomics trên cơ sở sử dụng các kỹ thuật phân lớp và gom cụm, kết hợp với việc khám phá tính chất của dữ liệu để hướng đến giải quyết những thách thức hiện nay của bài toán. Trong đó, cả hai khía cạnh chất lượng phân loại và chi phí tính toán đều được quan tâm. Cụ thể, giải pháp phân loại không giám sát dựa trên sự phong phú của hệ gien - MetaAB - sử dụng mô hình thu giảm vốn đòi hỏi ít chi phí tính toán mà không ảnh hưởng đến chất lượng phân loại. Ngoài ra, việc sử dụng kỹ thuật lặp nhằm phát hiện số cụm trong tập dữ liệu dựa trên kỹ thuật lựa chọn mô hình thống kê và một phương pháp đếm l-mer có độ dài thay đổi giúp làm tăng chất lượng phân loại của giải pháp đề xuất. Giải pháp BiMeta cũng là giải pháp phân loại không giám sát nhưng sử dụng đặc trưng tần số xuất hiện l-mer, và thông tin gối đầu giữa các trình tự. Việc áp dụng ý tưởng sử dụng tập đại diện của tập trình tự giúp BiMeta có khả năng gom cụm với độ chính xác cao mà không đòi hỏi chi phí và tài nguyên tính toán lớn. SeMeta là giải pháp phân loại cho phép gán nhãn trình tự dựa trên kỹ thuật bán giám sát. Kỹ thuật này cho phép sử dụng kết hợp tính hợp thành và tính tương đồng của trình tự nhằm làm tăng chất lượng phân loại. Ý tưởng thực hiện so sánh tương đồng cho đại diện của cụm thay vì tất cả trình tự trong cụm trong bước gãn nhãn giúp giảm đáng kể chi phí tính toán so với các giải pháp dựa trên tính tương đồng khác nhưng vẫn giữ được chất lượng phân loại tốt.

Kết quả thực nghiệm cho thấy sự hiệu quả của các giải pháp đề xuất ở cả hai khía cạnh chất lượng phân loại và chi phí tính toán so với giải pháp cùng loại trên dữ liệu giả lập và dữ liệu thực. Trong đó, xử lý cho trình tự ngắn là thế mạnh của các giải pháp đề xuất trong luận án này. Các giải pháp đề xuất còn cho phép thực thi trên cả hai kiểu dữ liệu trình tự dạng single-end và paired-end, và hứa hẹn là những công cụ hữu ích phục vụ cho các dự án metagenomics nhằm khám phá cộng đồng vi sinh vật. Mã nguồn của các giải pháp và dữ liệu thực nghiệm trong luận án có thể được tải về từ trang web `http://it.hcmute.edu.vn/bioinfo/metapro/index.html`.

7.2. Hướng phát triển

Trong tương lai, một số khía cạnh có thể được khai thác và cải tiến nhằm nâng cao hiệu quả phân loại của các giải pháp đề xuất. Kết quả thực nghiệm cho thấy rằng khi số lượng của loài trong tập dữ liệu hay kích thước tập dữ liệu càng lớn, chất lượng phân loại của các giải pháp đề xuất giảm đi, đồng thời chi phí tính toán tăng lên một cách đáng kể. Vì vậy, việc nghiên cứu và vận dụng đặc trưng phân loại phù hợp cần tiếp tục được nghiên cứu cho trường hợp dữ liệu lớn. Bên cạnh đó, một số độ đo khoảng cách khác có thể được nghiên cứu thay thế cho độ đo Euclide được sử dụng trong hai giải pháp BiMeta và SeMeta nhằm làm tăng chất lượng phân loại. Ngoài ra, công nghệ tính toán hiệu năng cao có thể được áp dụng giúp giảm thời gian tính toán cũng như nâng cao chất lượng nghiệm của bài toán.

Đối với vấn đề gán nhãn trình tự, khả năng gán nhãn trình tự vào một vị trí thực tế trên cây sinh loài của giải pháp SeMeta mặc dù tốt hơn so với các giải pháp được thực nghiệm trong luận án này, nhưng vẫn còn thấp bởi sự nghiêm ngặt của độ đo này. Một trong những hướng tiềm năng là quan tâm đến mức độ tương đồng khác nhau (được thể hiện bởi BLAST bit-scores) của các BLAST hit tin cậy. Đồng thời, thông tin này có thể kết hợp với việc khảo sát và ước lượng ngưỡng giá trị thể hiện mức độ tương đồng của các trình tự theo từng bậc phân loại để đạt được khả năng dự đoán tốt hơn.

Ngoài ra, luận án này chưa phân tích mức độ ảnh hưởng của lỗi giải mã trình tự đối với hiệu năng của các giải pháp phân loại. Mặc dù vậy, thực nghiệm ở chương 3 cho thấy có sự khác biệt về kết quả phân loại giữa trường hợp trình tự có lỗi giải mã và không có lỗi giải mã. Vì vậy, vấn đề này cần được nghiên cứu trong tương lai. Qua đó, phương pháp sửa lỗi trình tự cũng có thể được áp

dụng nhằm làm tăng chất lượng phân loại của các giải pháp.

TÀI LIỆU THAM KHẢO

[1] J. C. Wooley, A. Godzik, and I. Friedberg, "A primer on metagenomics," *PLoS Comput Biol*, vol. 6, no. 2, p. e1000667, 2010.

[2] T. Thomas, J. Gilbert, and F. Meyer, "Metagenomics-a guide from sampling to data analysis," *Microb Inform Exp*, vol. 2, no. 3, pp. 1–12, 2012.

[3] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, *et al.*, "A human gut microbial gene catalogue established by metagenomic sequencing," *nature*, vol. 464, no. 7285, pp. 59–65, 2010.

[4] J. G. Black, *Microbiology: Principles and Explorations (Chapter 9)*. US: Wiley, 8th ed., January 2012.

[5] J. Bohlin, "Genomic signatures in microbes - properties and applications," *The Scientific World Journal*, vol. 11, 2011.

[6] Y. Wang, H. C. Leung, S. M. Yiu, and F. Y. Chin, "Metacluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample," *Bioinformatics*, vol. 28, pp. i356 – i362, September 2012.

[7] Y. W. Wu and Y. Ye, "A novel abundance-based algorithm for binning metagenomic sequences using l-tuples," *Journal of Computational Biology*, vol. 18, no. 3, pp. 523 – 534, 2011.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society.SeriesB (Methodological)*, vol. 39, no. 1, pp. 1 – 38, 1977.