THÔNG TIN LUẬN ÁN

Tên luận án: **PHÂN LOẠI TRÌNH TỰ METAGENOMICS TRÊN CƠ SỞ PHÂN LỚP VÀ GOM CỤM.**

Chuyên ngành: **Khoa học máy tính.**

Mã ngành: **62.48.01.01.**

Họ tên NCS: **Lê Văn Vinh.**

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

2. PGS. TS. Trần Văn Hoài

Cơ sở đào tạo: **Trường Đại học Bách Khoa, Đại học Quốc gia Tp. Hồ Chí Minh.**

1. Mục tiêu của luận án

Luận án này nhằm nghiên cứu đề xuất giải pháp phân loại cho dữ liệu metagenomic, 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.

2. Những đóng góp chính của luận án

- **Đóng góp 1:** Luận án đề xuất các phương pháp nhằm nâng cao chất lượng của vấn đề phân loại trình tự metagenomics dựa trên sự phong phú của hệ gien. Ba đóng góp chính của luận án trong vấn đề này là: (1) Sử dụng mô hình thu giảm vốn đòi hỏi ít chi phí tính toán để tìm ước lượng khả năng cực đại của tham số cho mô hình xác suất; (2) Vận dụng phương pháp lựa chọn mô hình nhằm phát hiện số cụm trong tập dữ liệu, giúp làm tăng chất lượng phân loại. (3) Đề xuất một phương pháp đếm *l*-mer với độ dài thay đổi, giúp làm tăng chất lượng của giải pháp phân loại dựa trên sự phong phú của hệ gien khi dữ liệu không có lỗi giải mã.
- **Đóng góp 2:** Luận án đề xuất ý tưởng xây dựng tập đại diện của một tập trình tự cùng hệ gien dựa trên thông tin gối đầu trình tự. Tập đại diện này cho thấy nó vẫn chứa đựng đặc trưng tương đồng và hợp thành. Do đó, nó có khả năng bảo toàn

đặc trưng của tập dữ liệu ban đầu và có thể được vận dụng nhằm giúp giảm chi phí tính toán mà vẫn giữ được chất lượng của giải pháp phân loại.

- **Đóng góp 3:** Một giải pháp không giám sát được đề xuất cho phân loại trình tự sử dụng đặc trưng tần số xuất hiện *l*-mer, và thông tin gối đầu giữa các trình tự. Giải pháp đề xuất này, được gọi là BiMeta, sử dụng ý tưởng tìm tập đại diện của tập trình tự thuộc cùng hệ gien nhằm mục đích vừa giảm chi phí tính toán, vừa đạt được chất lượng phân loại tốt. BiMeta bao gồm hai pha chính. Trong pha đầu, trình tự được gom thành từng nhóm dựa trên thông tin gối đầu giữa chúng. Pha hai trộn các nhóm vào các cụm dựa trên đặc trưng phân bố tần số xuất hiện *l*-mer được rút trích từ tập đại diện của các nhóm này.
- **Đóng góp 4:** Một giải pháp bán giám sát, được gọi là SeMeta, được đề xuất nhằm phân loại có gán nhãn cho trình tự. SeMeta cũng sử dụng ý tưởng tìm tập đại diện của tập dữ liệu, nhưng giải pháp này vận dụng khả năng bảo toàn tính tương đồng của chúng để phân loại trình tự. Giải pháp này bao gồm hai bước chính. Sau bước gom cụm sử dụng phương pháp cải tiến của BiMeta, nó thực hiện gán nhãn từng cụm vào từng đơn vị phân loại phù hợp dựa trên sự tương đồng giữa trình tự trong đại diện của các cụm với cơ sở dữ liệu tham khảo. Bên cạnh đó, một kỹ thuật lọc những thông tin nhiễu (BLAST hit không tin cậy) từ quá trình so sánh tương đồng cũng được áp dụng giúp làm tăng chất lượng phân loại của giải pháp.

Tập thể hướng dẫn	Nghiên cứu sinh
PGS. TS. Trần Văn Lăng	Lê Văn Vinh
PGS. TS. Trần Văn Hoài	