The Saigon International University

Khóa luận tốt nghiệp

Website: www.siu.edu.vn

Thành phố Hồ Chí Minh - 2023

KHÓA LUẬN TỐT NGHIỆP

Ngành

Khoa học máy tính

Đề tài

Tích hợp dữ liệu đa phương thức trong phân tích dữ liệu tế bào đơn.

Giảng viên hướng dẫn

TS. Huỳnh Ngọc Tín

Sinh viên

Trần Lê Hải Bình

Mã sinh viên: 81011901642

The Saigon International University

Lewis Campus

Email: admission@siu.edu.vn
Website: www.siu.edu.vn

LỜI CAM ĐOAN

Em xin cam đoan những nội dung được trình bày trong khóa luận này là hoàn toàn trung thực.

Những số liệu, bảng biểu phục vụ cho việc phân tích và dẫn dắt đề tài này được thu thập từ các nguồn tài liệu khác nhau được ghi chú trong mục tài liệu tham khảo hoặc chú thích ngay bên dưới các bảng biểu.

Em cam kết hoàn toàn chịu mọi trách nhiệm thuộc về tính trung thực cũng như kết quả của đề tài nghiên cứu này.

TP. Hồ Chí Minh, tháng 6 năm 2023

Sinh viên thực hiện

Trần Lê Hải Bình

LỜI CẢM ƠN

Đầu tiên, em xin chân thành cảm ơn TS. Huỳnh Ngọc Tín, là người thầy đã tận tình hướng dẫn giúp đỡ em vượt qua những khó khăn và hoàn thành khóa luận tốt nghiệp của mình. Thầy đã đưa ra những lời khuyên bổ ích không chỉ áp dụng trong khóa luận mà còn là kim chỉ nam cho hướng phát triển sau này của em.

Đồng thời, em cũng muốn gửi lời cảm ơn sâu sắc tới tất cả thầy cô của trường Đại học Quốc tế Sài Gòn, đặc biệt là các thầy cô trong khoa Kỹ thuật và Khoa học máy tính đã tận tình giúp đỡ giảng dạy em trong thời gian học tập và rèn luyện tại trường.

TP. Hồ Chí Minh, tháng 6 năm 2023

Trần Lê Hải Bình, Sinh viên thực hiện

NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .
. . . . . .

TÓM TẮT KHÓA LUẬN

Sau quá trình tìm hiểu và thực hiện, khóa luận đã đạt được một số kết quả sau:

- Tìm hiểu tổng quan về bài toán tích hợp dữ liệu đa phương thức trong phân tích dữ liệu tế bào đơn bao gồm định nghĩa, thách thức, ứng dụng và các hướng tiếp cận cho bài toán.
- Phân tích được một số ưu và nhược điểm của của các phương pháp đề xuất trong cuộc thi *Open Problem in Single cell Analysis 2022*, và một số phương pháp ngoài cuộc thi tiếp cận dựa trên các mô hình học sâu.
- Tìm hiểu, thực nghiệm và đánh giá các phương pháp nêu trên.
- Kết quả đạt được:

  (+) Kết quả chính liên quan đến khóa luận:

  - *Submitted*, Binh Tran, Tri Pham, Tin Huynh, and Kiem Hoang, *"A Comprehensive Comparative Study in Multimodal Single-Cell Data Integration,"* 2023 International Conference on Multimedia Analysis and Pattern Recognition (MAPR).
  - TS. Huỳnh Ngọc Tín, GS. TSKH. Hoàng Văn Kiếm, Trần Lê Hải Bình, và Phạm Xuân Trí, *Cơ hội và phân tích trong phân tích dữ liệu tế bào đơn*, Hội thảo chủ đề "ĐỔI MỚI DẠY – HỌC VỚI CHATGPT VÀ TRÍ TUỆ NHÂN TẠO", Trường Đại học quốc tế Sài Gòn.

(+) Kết quả khác trong quá trình học tập nghiên cứu:

- Top 15 Hội thi Thử thách Trí tuệ Nhân tạo (AI-Challenge) TP.HCM, 2022.
- Giải nhất Cuộc thi sinh viên nghiên cứu khoa học Trường Đại học Quốc tế Sài Gòn năm 2023.

Từ khóa: Deep Learning, Single-cell analysis, Multimodal single-cell Integration.

Mục lục

Danh sách hình vẽ

Danh sách bảng

Chương 1

TỔNG QUAN

1.1 Đặt vấn đề

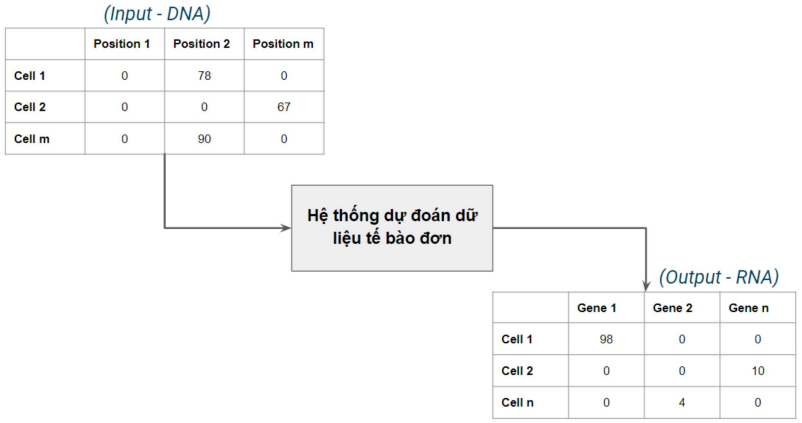
Cộng đồng khoa học trong lĩnh vực nghiên cứu tế bào đơn đã phát triển bài toán tích hợp dữ liệu trong cuộc thi *Open Problems in Single-cell Analysis - Multimodal Single-Cell Integration* tại hội nghị *NeurIPS 2022*, dựa trên hai công nghệ xét nghiệm có tên lần lượt là **Multiome** [5] và **CITEseq** [6]. Mỗi công nghệ xét nghiệm đo lường hai phương thức. Cụ thể, bộ xét nghiệm **Multiome** đo khả năng tiếp cận của chất nhiễm sắc (DNA) và mức độ biểu hiện gen (RNA), trong khi bộ xét nghiệm **CITEseq** đo mức độ biểu hiện gen (RNA) và mức độ protein bề mặt. Cuộc thi tìm kiếm các giải pháp áp dụng thế mạnh của lĩnh vực *Khoa học máy tính* nhằm giải quyết bài toán tích hợp dữ liệu đa phương thức trong phân tích dữ liệu tế bào đơn.

Trong cuộc thi *Open Problems in Single-cell Analysis - Multimodal Single-Cell Integration* tại hội nghị *NeurIPS 2022*, bài toán tích hợp dữ liệu tế bào đơn được mô tả như sau:

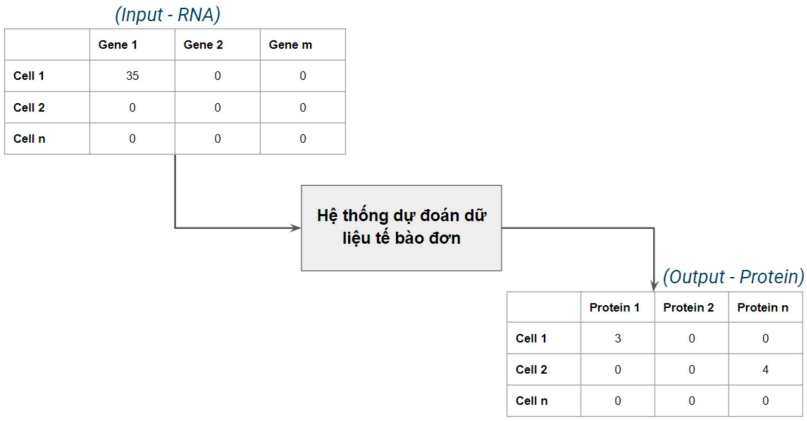
- Đối với thí nghiệm **Multiome**: Bài toán nhận đầu vào là dữ liệu đo khả năng tiếp cận của chất nhiễm sắc (DNA), yêu cầu dự đoán đầu ra là mức độ biểu hiện gen (RNA) tương ứng.

- Đối với thí nghiệm **CITEseq**: Bài toán nhận đầu vào là dữ liệu đo mức độ biểu hiện gen (RNA), yêu cầu dự đoán đầu ra là mức độ protein bề mặt tương ứng.

Hình 1.1, 1.2 mô tả một ví dụ về đầu vào và đầu ra của bài toán lần lượt trên hai thí nghiệm **Multiome** và **CITEseq**.



Hình 1.1: Minh họa bài toán tích hợp dữ liệu trong thí nghiệm Multiome



Hình 1.2: Minh họa bài toán tích hợp dữ liệu trong thí nghiệm CITEseq

Hiện nay, kỹ thuật giải trình các thành phần sinh học và các thông tin di truyền không chỉ được thực hiện trên các mẫu mô, nhưng còn có thể thực hiện

được trên từng tế bào đơn lẻ, điều này đã tạo ra những bước ngoặt mới trong lĩnh vực khoa học sự sống [7]. Mỗi một tế bào đơn lẻ chính là đơn vị cấu trúc và chức năng cơ bản nhất của sự sống [8]. Dù mỗi cơ thể sống của người trưởng thành chứa hơn 37,2 nghìn tỷ tế bào [9], nhưng trong khía cạnh bệnh học, ví dụ như ung thư, sự tăng sinh bất thường của một tế bào cũng có thể dẫn đến sự suy tàn của toàn bộ cơ thể sinh vật sống. Do đó, sự phát triển các kỹ thuật giải trình tế bào đơn trong những năm gần đây đã thay đổi mô hình nghiên cứu, đặc biệt trong lĩnh vực di truyền học. Thay vì tập trung nghiên cứu trên toàn bộ mẫu mô, các nhà nghiên cứu y sinh hướng đến nghiên cứu chi tiết và toàn diện trên từng tế bào [10].

Với khả năng có thể phân tích hàng nghìn hoặc thậm chí hàng triệu tế bào chỉ trong một thí nghiệm duy nhất, các công nghệ giải trình tế bào đơn đã tạo nên một cuộc cách mạng dữ liệu trong y sinh và đặt ra các vấn đề khoa học dữ liệu hấp dẫn. Tuy nhiên, phân tích và khai thác dữ liệu tế bào đơn phải đối mặt nhiều thách thức, không chỉ từ góc nhìn của khoa học dữ liệu nhưng còn vì sự đa dạng sinh học của khoa học sự sống. Một số thách thức lớn đang được cộng đồng nghiên cứu quan tâm có thể kể đến như dữ liệu có số chiều quá lớn và độ thưa thớt cao, nhu cầu tích hợp dữ liệu giữa các loại phép đo tế bào đơn khác nhau (ví dụ: RNA, DNA, protein và quá trình methyl hóa) và trên các mẫu, có thể là từ các thời điểm khác nhau, các nhóm điều trị, hoặc thậm chí cả sinh vật khác nhau, hay nhu cầu đánh giá hiệu quả của các phương pháp đo lường một cách có hệ thống. Các thách thức này được thúc đẩy từ những câu hỏi nghiên cứu trong lĩnh vực y sinh, do đó nỗ lực giải quyết các thách thức này chính là mở ra một hướng nghiên cứu giao thoa mới, và nhiều hiểu biết sâu rộng hơn trong các ngành khoa học sự sống [11].

Từ năm 2017, các phương pháp *"joint method"*, lược dịch là *"phương pháp tích hợp"*, được giới thiệu, đã có thể thực hiện cùng lúc từ hai phương thức đo đạt trở lên [12, 13, 14, 15]. Điều này đã tạo nên một tiền đề quan trọng để giải

quyết nhu cầu tích hợp dữ liệu giữa các phương thức đo đạt khác nhau đã nhắc đến ở trên. Cụ thể hơn, những phương pháp này đã cung cấp các tập dữ liệu tiêu chuẩn - các tập dữ liệu dù từ nhiều phương thức do đạt khác nhau nhưng đồng thời được thực hiện trên cùng một quần thể tế bào tại một thời điểm xác định. Do đó, đảm bảo quá trình tích hợp nhiều tập dữ liệu của các phương thức khác nhau có thể đưa ra những nhận định và đánh giá có tính khoa học, khách quan, đồng thời khắc phục được những hạn chế khi phân tích đơn lẻ các tập dữ liệu [16].

Qua hai cuộc thi được tổ chức trong hội nghị *NeurIPS 2021 & 2022*, cộng đồng nghiên cứu đã thúc đẩy sự giao thoa nghiên cứu giữa hai lĩnh vực *Khoa học máy tính* và *Phân tích tế bào đơn*. Một số giải pháp tham gia tại hai cuộc thi đã góp phần đưa ra giải pháp hiệu quả giải quyết bài toán. Các phương pháp trong khuôn khổ cuộc thi không chỉ dừng lại ở việc thiết kế các mô hình huấn luyện cho kết quả cao, nhưng khai thác các yếu tố tiền dữ liệu và lựa chọn đặc trưng hợp lý, nhằm phát triển những phương pháp phù hợp với dữ liệu trong lĩnh vực này. Vì đây là một nghiên cứu giao thoa mới, các phương pháp được đề xuất ở cuộc thi được thực nghiệm với nhiều mô hình, nhiều giải pháp tiền xử lý dữ liệu khác nhau. Nhìn chung các phương pháp này xây dựng dựa trên các mô hình học sâu đơn giản, kết hợp các phương pháp tiền xử lý dữ liệu dựa trên thống kê. Tuy nhiên, tính đến thời điểm thực hiện đề tài này, qua khảo sát, chưa có những phân tích chính thức nhằm đánh giá tính hiệu quả của các phương pháp nêu trên. Do đó, trong đề tài lần này tôi tập trung đánh giá lại một số phương pháp có kết quả cao tại cuộc thi, tập trung sâu vào các phương pháp sử dụng kiến trúc *học máy và học sâu*. Tôi sẽ tiến hành phân tích, so sánh và đánh giá khách quan bốn giải pháp cho kết quả cao nhất tại cuộc thi năm 2022, nhằm cung cấp những tiền đề cho các nghiên cứu tiếp theo cho các phân tích này. Bên cạnh đó, tôi cũng khảo sát một số phương pháp mới đang được đề xuất để giải quyết bài toán, nhưng chưa được áp dụng trong cuộc thi nhằm cung cấp một cái nhìn đa dạng và khách

quan hơn.

1.2 Phạm vi và mục tiêu

1.2.1 Mục tiêu

Để hoàn thành công việc, tôi đã đề ra mục tiêu cụ thể sau:

- Tìm hiểu tổng quan và các hướng tiếp cận cho bài toán tích hợp dữ liệu đa phương thức trong phân tích dữ liệu tế bào đơn.
- Khảo sát và phân tích một số phương pháp học sâu trong và ngoài cuộc thi *Open Problems in Single-cell Analysis - Multimodal Single-Cell Integration* tại hội nghị *NeurIPS 2022*
- Thực nghiệm các phương pháp nói trên nhằm đưa ra các phân tích, so sánh và đánh giá.

1.2.2 Phạm vi

Trong khuôn khổ giới hạn, khóa luận tập trung hoàn thành các công việc sau:

- Tìm hiểu tổng quan về định nghĩa, tính cấp thiết, thách thức và các hướng tiếp cận của bài toán tích hợp dữ liệu đa phương thước trong phân tích dữ liệu tế bào đơn.
- Tìm hiểu ưu, nhược điểm của một số phương pháp học sâu tại cuộc thi *Open Problems in Single-cell Analysis - Multimodal Single-Cell Integration* tại hội nghị *NeurIPS 2022*. Từ đó đúc kết các bài học và kinh nghiệm tạo cơ sở cho các nghiên cứu về sau trong lĩnh vực này.

1.3 Đóng góp của khóa luận

Sau quá trình tìm hiểu và thực hiện khóa luận, tôi đã có một số đóng góp sau đây:

- Tài liệu tổng quan về bài toán tích hợp dữ liệu đa phương thước trong phân tích dữ liệu tế bào đơn, bao gồm định nghĩa, những thách thức và ứng dụng của bài toán.
- Phân tích những ưu điểm và hạn chế của các giải pháp theo hướng tiếp cận sử dụng mô hình học máy và học sâu tại cuộc thi *Open Problems in Single-cell Analysis - Multimodal Single-Cell Integration* tại hội nghị *NeurIPS 2022*, đồng thời cũng khảo sát qua một số phương pháp khác ngoài cuộc thi theo cùng hướng tiếp cận này, cung cấp các tiền đề cho những nghiên cứu về sau.
- Bản thảo bài báo *A Comprehensive Comparative Study in Multimodal Single-Cell Data Integration* tại Hội nghị MAPR 2023 (The International Conference on Multimedia Analysis and Pattern Recognition (MAPR) supported by VAPR, *IEEE Xplore*®)

1.4 Cấu trúc khóa luận

Cấu trúc của khóa luận gồm 6 chương như sau:

- **Chương 1:** Giới thiệu tổng quan về đề tài của khóa luận.
- **Chương 2:** Trình bày tổng quát về định nghĩa, thách thức và động lực của bài toán. Đồng thời tìm hiểu các hướng tiếp cận có thể giải quyết bài toán tích hợp dữ liệu đa phương thước trong phân tích dữ liệu tế bào đơn. Tập trung khảo sát các giải pháp tiếp cận theo hướng học máy và học sâu.

- **Chương 3:** Trình bày các phân tích, đánh giá với những giải pháp cho kết quả cao nhất tại cuộc thi *Open Problems in Single-cell Analysis - Multimodal Single-Cell Integration* tại hội nghị *NeurIPS 2022* và một số giải pháp mới được đề xuất, chưa được áp dụng trong cuộc thi.
- **Chương 4:** Trình bày các kết quả thực nghiệm và những phân tích đánh giá của các phương pháp trên tập dữ liệu trong cuộc thi *Open Problems in Single-cell Analysis - Multimodal Single-Cell Integration* tại hội nghị *NeurIPS 2022*.
- **Chương 5:** Đưa ra kết luận và hướng phát triển tiềm năng của đề tài.