ĐẠI HỌC QUỐC GIA HÀ NỘI

ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA TOÁN - CƠ – TIN HỌC

**TÓM TẮT LUẬN VĂN**

Học viên: **Tạ Văn Nhân.**

Giáo viên hướng dẫn: **PGS. TS. Nguyễn Thị Hồng Minh.**

Tên đề tài: **Áp dụng phương pháp gióng hàng trình tự cho bài toán dự đoán biến thể gen.**

Giải trình tự DNA đang ngày càng trở nên nhanh chóng và kinh tế. Tuy nhiên, để ráp các trình tự thu được dựa trên bộ gen tham chiếu và tìm kiếm các biến thể, chúng ta cần có những hệ thống đủ mạnh để xử lý và phân tích dữ liệu. Phương pháp dóng hàng trình tự là một giải pháp hữu hiệu cho vấn đề này. Đã có nhiều kết quả nghiên cứu liên quan tới phương pháp cũng như phát triển công cụ dóng hàng trình tự được công bố. Tuy nhiên vấn đề về thời gian thực hiện, mức độ chính xác và phạm vi áp dụng của các kĩ thuật dóng hàng vẫn còn là những chủ đề cần được phát triển.

Mục đích của luận văn là nghiên cứu sâu kĩ thuật dóng hàng trình tự, đề xuất cải tiến để tăng hiệu quả về thời gian của thuật toán cũng như khả năng triển khai trên các hạ tầng tính toán phổ dụng. Đồng thời áp dụng thuật toán để khám phá trong phạm vi rộng hơn các biến thể gen so với một số nghiên cứu trước đây, khám phá mức độ ảnh hưởng của các biến thể đến chức năng của Protein.

Cụ thể trong nghiên cứu này, chúng tôi phát triển thuật toán dóng hàng dựa trên chuyển dạng Burrows-Wheeler và thuật toán Smith-Waterman. Trong đó, các mã giả được viết chi tiết để có thể triển khai bằng các ngôn ngữ lập trình khác nhau. Chúng tôi sử dụng ngôn ngữ Go với kỹ thuật song song và đồng thời để triển khai thuật toán dóng hàng trình tự dựa trên chuyển dạng Burrows-Wheeler, các chương trình được triển khai có thể chạy trên các hệ thống tính toán hiệu năng cao nhiều bộ xử lí, và cũng có thể chạy trên các máy tính cá nhân với khả năng tận dụng tất cả các logic proccessor của bộ xử lí. Kết quả thực nghiệm thuật toán bằng chương trình của chúng tôi được so sánh với kết quả nhận được từ công cụ BWA-MEM nhằm kiểm nghiệm tính chính xác của thuật toán mà chúng tôi phát triển. Đồng thời, việc thử nghiệm này cũng giúp hiểu rõ hơn về các tham số cho phù hợp với dữ liệu để sử dụng thuận lợi các công cụ dóng hàng trên các hệ thống đã có.

Trong chương 1, luận văn giới thiệu một số kiến thức cơ sở về sinh học phân tử, tin sinh học, các công nghệ giải trình tự. Từ những kiến thức cơ sở đó, những nghiên cứu sâu về phương pháp dóng hàng trình tự và những đề xuất cải tiến được trình bày chi tiết trong chương 2, bao gồm cả phần phương pháp và phần thực nghiệm. Cuối cùng, trong chương 3 trình bày những kết quả áp dụng các phương pháp và công cụ dóng hàng để tìm biến thể gen của bệnh tâm thần phân liệt (Schizophrenia), một hội chứng rối loạn tâm thần nghiêm trọng có liên quan đến nhiều gen với yếu tố di truyền cao. Dữ liệu được tiền xử lý và khớp với bộ gen tham chiếu sử dụng thuật toán dựa trên chuyển dạng Burrows-Wheeler. Sau đó, thuật toán dóng hàng Smith-Waterman sắp xếp lại các Haplotype ở một số vùng hoạt động giúp kết quả dóng hàng ban đầu chính xác hơn. Các quá trình được triển khai trên nền tảng Galaxy và máy chủ Linux 64CPUs. Kết quả những biến thể tìm được trên các gen được so sánh với một số kết quả nghiên cứu của một số nhà khoa học và tổ chức đã được công bố.