MỞ ĐẦU

1. Tính cấp thiết của luận án

Những thành tựu gần đây trong công nghệ giải trình tự gen thế hệ mới (Next Generation Sequencing - NGS) đã giảm đáng kể chi phí giải trình tự toàn bộ hệ gen và dẫn đến sự gia tăng nhanh chóng về số lượng DNA / RNA và chuỗi protein sẵn sàng cho các phân tích. Những dữ liệu này đại diện cho một nguồn thông tin rất hữu ích và đặt ra các vấn đề tính toán mới trong các nghiên cứu trên toàn hệ gen, điển hình là các nghiên cứu về phân bố của các biến thể di truyền trong một quần thể hay xác định các vùng gen có tác động và có ý nghĩa về mặt sinh học đối với các đặc điểm quan trọng mà ta quan tâm, … Để giải quyết những bài toán này đòi hỏi nhiều công cụ mới, đáng chú ý trong số đó là đồ thị tái tổ hợp di truyền (Ancestral Recombination Graph - ARG), một công cụ quan trọng trong nghiên cứu di truyền quần thể và các bài toán liên quan đến tìm sự đa dạng trong hệ gen.

Với một tập các chuỗi nhiễm sắc thể, đồ thị ARG đầy đủ sẽ mô tả một cách đầy đủ lịch sử di truyền, mối quan hệ của chúng với nhau và với một tổ tiên chung thông qua ba sự kiện: đột biến, tái tổ hợp và kết hợp. Trong quá trình xây dựng đồ thị ARG, sự kiện tái tổ hợp và sự kiện đột biến là 2 sự kiện cốt lõi ảnh hưởng tới đồ thị kết quả, từ đó ảnh hưởng trực tiếp tới các ứng dụng liên quan như tìm vùng gen liên quan đến bệnh, đột biến gây bệnh, đặc trưng của quần thể quan sát, … Tuy nhiên, số sự kiện tái tổ hợp và sự kiện đột biến cũng như vị trí thực sự xảy ra trong quá trình tiến hóa là không thể xác định được. Do đó, chúng ta không thể biết được ARG thực sự mà chúng ta chỉ có thể suy diễn chúng từ dữ liệu với các giả định tối ưu số sự kiện tái tổ hợp và sự kiện đột biến nhằm có được ARG với các sự kiện sát nhất với thực tế.

Tuy nhiên, các phương pháp xây dựng đồ thị ARG hiện tại vẫn gặp những hạn chế sau:

- Các phương pháp xây dựng đồ thị ARG mới chỉ giới hạn với những tập dữ liệu vừa và nhỏ hàng trăm trình tự.
- Các phương pháp xây dựng đồ thị ARG có chính xác số sự kiện tái tổ hợp ít nhất hiện thời còn tốn rất nhiều thời gian và chỉ khả thi với những tập dữ liệu rất nhỏ vài chục trình tự.

2. Mục tiêu của luận án

1) Nghiên cứu các phương pháp xây dựng đồ thị ARG hiện tại, từ đó đề xuất một thuật toán gần đúng xây dựng đồ thị ARG cho hàng nghìn trình tự, thậm chí hàng nghìn hệ gen nhằm ứng dụng hiệu quả vào các bài toán thực tế trên các tập dữ liệu lớn.

2) Đề xuất thuật toán xây dựng đồ thị ARG với hàm mục tiêu tối ưu số sự kiện tái tổ hợp trong quá trình xây dựng đồ thị ARG bằng việc kết hợp linh hoạt thuật toán đề xuất trong (1) với một số đặc trưng của dữ liệu và các kĩ thuật tối ưu được sử dụng trong các phương pháp tìm cận dưới tái tổ hợp và các phương pháp xây dựng đồ thị ARG có số sự kiện tái tổ hợp nhỏ nhất đã có.

3. Các đóng góp của luận án

Trong luận án này, dựa trên thực nghiệm, chúng tôi đề xuất một số cải tiến mới thuật toán xây dựng đồ thị ARG để giảm độ phức tạp tính toán quá trình xây dựng đồ thị và tăng khả năng xử lý được dữ liệu lớn hàng nghìn trình tự trên phạm vi toàn hệ gen người. Chúng tôi đề xuất sử dụng *đoạn đầu chung dài nhất* giữa các trình tự để xác định sự kiện tái tổ hợp. Chiến lược này giúp đảm bảo số nút trong đồ thị luôn được ổn định sau mỗi lần thực hiện bước tái tổ hợp và làm giảm đáng kể số sự kiện tái tổ hợp cũng như thời gian để xây dựng đồ thị ARG. Thực nghiệm ứng dụng trong bài toán tìm vùng gen liên quan đến bệnh sốt rét ở Châu Phi gồm 5560 trình tự trên toàn nhiễm sắc thể 11 đã nhấn mạnh thêm hiệu quả nổi trội của thuật toán đề xuất so với các thuật toán hiện tại. Luận án cũng đã đề xuất 2 thuật toán cải tiến REARG và GAMARG nhằm tối ưu thêm số sự kiện tái tổ hợp trong quá trình xây dựng đồ thị ARG. Thuật toán REARG giúp quá trình xây dựng ARG khu trú được vào các ARG có số sự kiện tái tổ hợp nhỏ nhanh hơn ARG4WG trong hữu hạn số lần chạy thuật toán đối với các tập dữ liệu vừa và lớn. Tuy nhiên, GAMARG tổng quát hơn. GAMARG có khả năng xây dựng được những ARG có chính xác hoặc gần chính xác số sự kiện tái tổ hợp nhỏ nhất.

Các kết quả của luận án đã được công bố trong 01 bài báo ở tạp chí SCI quốc tế và 02 báo cáo ở hội nghị quốc tế có phản biện.

4. Bố cục của luận án

Ngoài phần kết luận, luận án được tổ chức như sau.

Chương 1 giới thiệu khái quát về dữ liệu hệ gen người, cụ thể là cấu trúc bộ gen người, các nguyên nhân dẫn tới các biến thể di truyền ở người và các loại biến thể di truyền phổ biến. Chúng tôi cũng giới thiệu sơ lược về các loại mạng phát sinh loài, một công cụ quan trọng để biểu diễn các mối quan hệ tiến hóa trong nghiên cứu di truyền quần thể. Sau đó là phần giới thiệu về bài toán xây dựng đồ thị ARG, các giả định được sử dụng trong quá trình xây dựng đồ thị ARG. Phần cuối của chương trình bày các cách tiếp cận giải bài toán xây dựng đồ thị ARG.

Chương 2 đề xuất một thuật toán xây dựng đồ thị ARG cho dữ liệu lớn hàng nghìn mẫu độ dài toàn hệ gen. Để làm được điều đó, chúng tôi đưa ra các nhược điểm của các cách tiếp cận hiện có, đặc biệt là những hạn chế trong

thuật toán Margarita xây dựng đồ thị ARG hợp lý được đề xuất bởi Minichiello và Durbin, từ đó đưa ra thuật toán đề xuất nhằm khắc phục các nhược điểm đó. Các kết quả thực nghiệm ở phần sau của chương đã chứng tỏ hiệu quả của thuật toán đề xuất. Phần cuối của chương giới thiệu ứng dụng thuật toán đề xuất vào bài toán tìm vùng gen liên quan đến bệnh sốt rét ở Châu Phi trên tập dữ liệu lớn gồm 5560 trình tự trên toàn nhiễm sắc thể 11. Các kết quả trong phần này đã khẳng định thêm hiệu quả, khả năng ứng dụng của thuật toán đề xuất trong các bài toán thực tế trên dữ liệu lớn.

Chương 3 của luận án giới thiệu các phương pháp nhằm cực tiểu hóa số sự kiện tái tổ hợp trong quá trình xây dựng đồ thị ARG. Cụ thể, chúng tôi đề xuất hai phương pháp: (1) kết hợp một số đặc trưng của dữ liệu và (2) kết hợp các kĩ thuật tối ưu vào việc lựa chọn và thực hiện sự kiện tái tổ hợp theo thuật toán đề xuất trong chương 2. Các thực nghiệm trên các bộ dữ liệu khác nhau đã chứng tỏ hiệu quả của các phương pháp đề xuất.

Chương 1. BÀI TOÁN XÂY DỰNG ĐỒ THỊ TÁI TỔ HỢP DI TRUYỀN

1.1. Giới thiệu chung

1.1.1. Dữ liệu hệ gen người

Giới thiệu về cấu trúc bộ gen người, các nguyên nhân dẫn tới các biến thể di truyền ở người và các loại biến thể di truyền phổ biến.

Hệ gen người gồm 23 cặp nhiễm sắc thể, có khoảng 3 tỉ phân tử DNA, khoảng 20.000 đến 25.000 gen. Hầu hết các gen ở mọi người là như nhau, nhưng có khoảng 0.1% vị trí mà các nucleotit là khác nhau ở mỗi người gọi là các biến thể di truyền. Đột biến và tái tổ hợp là 2 nguyên nhân chính của biến thể di truyền. Đột biến là nguồn gốc của biến thể mới, xảy ra khi có lỗi trong quá trình sao chép DNA mà không được sửa chữa bởi các enzyme sửa chữa DNA. Trong khi tái tổ hợp di truyền là nguyên nhân chính của biến thể di truyền ở thế hệ con cái. Tái tổ hợp góp phần vào biến đổi gen bằng cách xáo trộn DNA của cha mẹ và tạo ra các tổ hợp biến thể mới. Biến thể đa hình đơn nucleotide (SNP) là loại biến thể di truyền phổ biến nhất trong hệ gen người và có vai trò đặc biệt quan trọng trong các nghiên cứu tương quan toàn bộ nhiễm sắc thể.

1.1.2. Mạng phát sinh loài

Với sự đa dạng dữ liệu sinh học hiện có ngày nay đã đặt ra những nhu cầu phát triển các mạng phát sinh loài (phylogenetic network), thay vì chỉ dùng cây phân loài như trước đây, để biểu diễn các mối quan hệ dữ liệu khác nhau.

Mạng phát sinh loài là đồ thị nào đó được sử dụng để biểu diễn các mối quan hệ tiến hóa (bằng các cạnh) giữa một tập hợp các nhãn (taxa) (là các nút lá).

Có khoảng 20 loại mạng phát sinh loài khác nhau. Mỗi mạng có vai trò khác nhau: cây phân loài mô tả mối quan hệ giữa các loài hoặc các gen; mạng phân tách mô tả sự khác nhau giữa các cây phát sinh loài; các sự kiện lai ghép hay tái tổ hợp được mô hình hóa trong các mạng lai ghép hay các mạng tái tổ hợp, … Trong đó, sự kiện tái tổ hợp là sự kiện quan trọng thu hút được nhiều sự quan tâm của các nhà nghiên cứu, đặc biệt trong di truyền quần thể. Do sự tái tổ hợp diễn ra trong tất cả các thế hệ, bộ gen mà bất kỳ cá thể nào thừa hưởng là sự pha trộn và phản ánh DNA của nhiều cá thể khác nhau qua các thế hệ tổ tiên. Sự tồn tại của bộ gen tổ hợp phong phú như vậy thúc đẩy các nghiên cứu về sự biến đổi gen trong các quần thể để khám phá mối quan hệ giữa nội dung bộ gen và các đặc điểm quan tâm có ảnh hưởng từ yếu tố di truyền. Việc phân tích và xác định được các sự kiện tái tổ hợp giúp cho quá trình xác định đa dạng di truyền, tìm hiểu các nguyên nhân dẫn đến các bệnh đa yếu tố như bệnh tiểu đường, ung thư, … và là nền tảng nghiên cứu thuốc chữa bệnh.

Trong luận án này, chúng tôi tập trung nghiên cứu về đồ thị tái tổ hợp di truyền, một loại mạng phát sinh loài mô hình hóa quan hệ di truyền giữa các trình tự hệ gen được quan sát trong một quần thể.

1.2. Xây dựng đồ thị tái tổ hợp di truyền

1.2.1. Sự kiện tái tổ hợp

Tái tổ hợp là một thành phần cơ bản trong quá trình truyền DNA từ trình tự này sang trình tự khác khi các nhiễm sắc thể được truyền từ thế hệ này sang thế hệ khác. Có 2 kiểu tái tổ hợp phổ biến là trao đổi chéo (crossing over) và chuyển đổi gen (gene conversion). Mỗi loài sinh vật có cơ chế tái tổ hợp khác nhau. Đối với loài người, trao đổi chéo là kiểu tái tổ hợp phổ biến nhất xảy ra trong quá trình giảm phân.

1.2.2. Đồ thị tái tổ hợp di truyền

Đồ thị tái tổ hợp tổ tiên đóng một vai trò quan trọng trong nghiên cứu di truyền quần thể và các bài toán liên quan đến tìm sự đa dạng trong hệ gen.

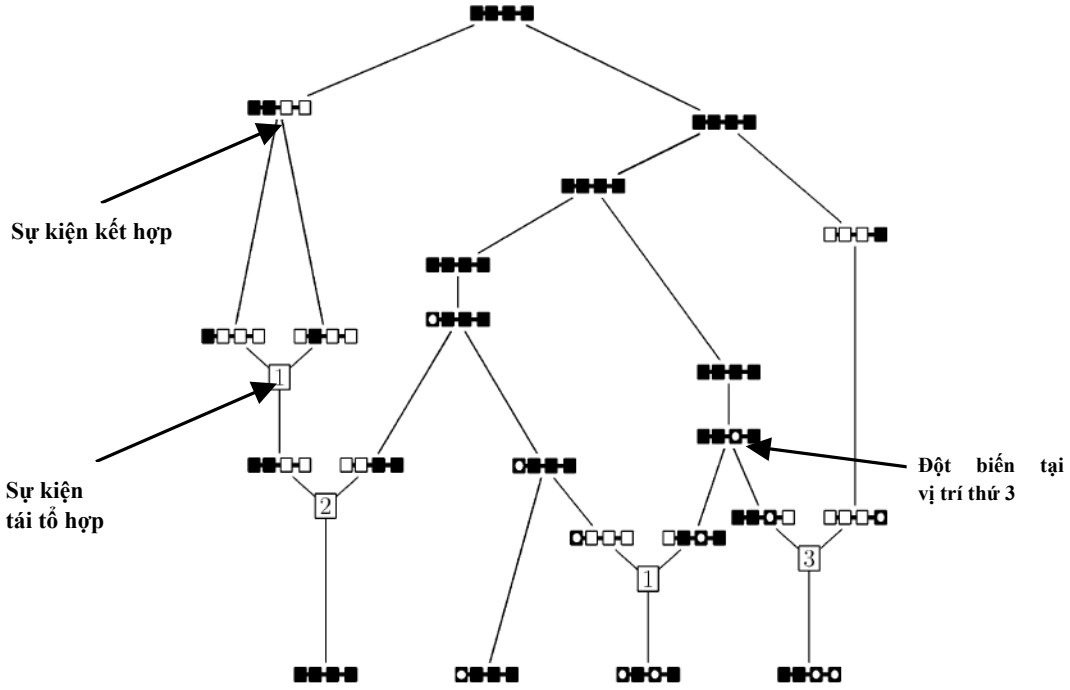
Bài toán xây dựng đồ thị ARG gắn với việc tái cấu trúc lịch sử tiến hóa của các trình tự được quan sát trong một quần thể trong đó các trình tự được tạo ra do đột biến và tái tổ hợp. Do đó, các trình tự ở đây được hiểu là các trình tự DNA đơn. Đối với quần thể người (và các loài lưỡng bội nói chung), 2 trình tự DNA trong mỗi người được coi là độc lập nhau trong quá trình xây dựng đồ thị ARG.

1.2.2.1 Mô hình các vị trí vô hạn

Trong chiều dài lịch sử tiến hóa, tại một vị trí trên tập các trình tự quan sát, sự kiện đột biến có thể xảy ra một hoặc nhiều lần (đột biến ngược hoặc đột biến lặp lại).
Quá trình xây dựng đồ thị ARG, với sự kiện tái tổ hợp là trọng tâm nghiên cứu, gắn với giả định có nhiều nhất một sự kiện đột biến xảy ra tại mỗi vị trí trong toàn bộ lịch sử tiến hóa, không cho phép đột biến ngược hoặc lặp lại. Mô hình đột biến này gọi là mô hình các vị trí vô hạn (*infinite-sites model*), mô tả sự tiến hóa của các chuỗi DNA rất dài với tỷ lệ đột biến thấp ở mỗi vị trí.

1.2.2.2 Cấu trúc đồ thị ARG

Với một tập các chuỗi nhiễm sắc thể, đồ thị ARG đầy đủ sẽ mô tả một cách đầy đủ lịch sử di truyền, mối quan hệ của chúng với nhau và với một tổ tiên chung thông qua ba sự kiện: đột biến, tái tổ hợp và kết hợp.
Có 4 thành phần cần thiết để xác định một đồ thị ARG tổng quát cho 1 tập trình tự nhị phân D cho trước: đồ thị cơ sở, các nhãn cạnh, các nhãn nút, và các trình tự quan sát.

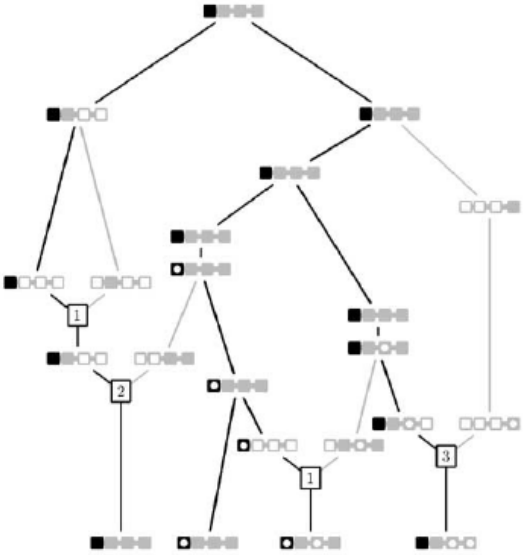


Hình 1.1: Một ví dụ đồ thị ARG với các ký hiệu: ■: trạng thái di truyền gốc, ◘: trạng thái di truyền đột biến, □: trạng thái không di truyền.

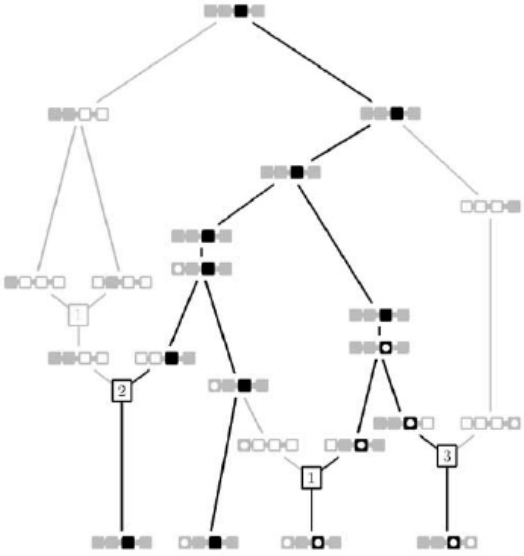
Hình 1.1 mô tả một ví dụ đồ thị tái tổ hợp tổ tiên. Đồ thị hiển thị rõ các thành phần di truyền và không di truyền trong một tập các chuỗi trình tự. Xét ngược chiều thời gian, một sự kiện kết hợp xuất hiện khi hai trình tự kết hợp với nhau thành một trình tự; một sự kiện đột biến xuất hiện khi một vị trí alen trong một trình tự bị thay đổi và một sự kiện tái tổ hợp xuất hiện khi một trình tự bị tách ra thành hai trình tự con, một trình tự mang thông tin di truyền phía trước vị trí cắt và trình tự

còn lại mang thông tin di truyền phía sau vị trí cắt. Điểm xảy ra sự kiện tái tổ hợp gọi là điểm cắt tái tổ hợp (breakpoint).

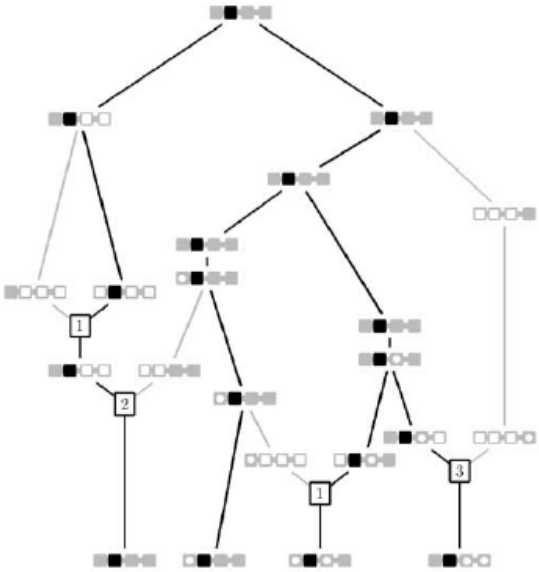
Với một đồ thị ARG đầy đủ được mô tả như trong Hình 1.1, mỗi vị trí *c* trong đồ thị sẽ có một cây thành phần (cây biên - marginal tree) T(*c*) mô tả lịch sử của các cá thể cho vị trí đó. Từ tập trình tự ban đầu, với mỗi trình tự ta lần theo các cạnh của đồ thị tái tổ hợp di truyền cho vị trí *c*; khi một sự kiện tái tổ hợp xuất hiện, ta đi theo đường bên trái nếu vị trí tái tổ hợp xảy ra sau *c* và đi theo đường bên phải trong trường hợp ngược lại. Tập tất cả các cạnh đó sẽ định nghĩa T(*c*). Hình 1.2 minh họa các cây thành phần cho đồ thị ARG trong Hình 1.1.



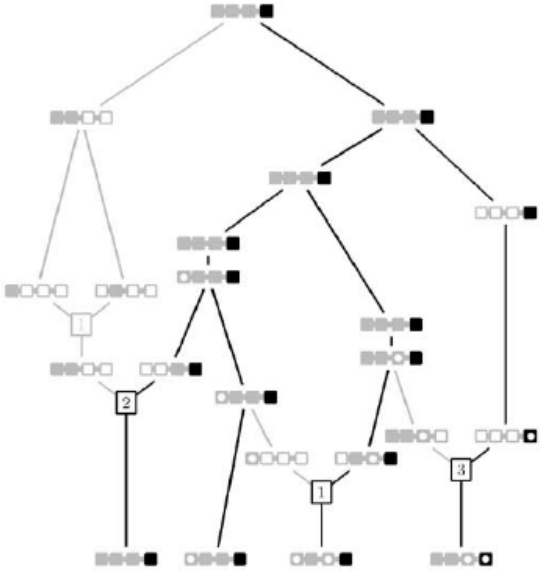
(1) Cây thành phần cho marker 1



(2) Cây thành phần cho marker 3



(3) Cây thành phần cho marker 2



(4) Cây thành phần cho marker 4

Hình 1.2: Cây thành phần của đồ thị ARG trong Hình 1.1.

Bên cạnh các thuật toán xây dựng đồ thị ARG đầy đủ, rất nhiều thuật toán, đặc biệt theo cách tiếp cận thống kê thường xây dựng đồ thị ARG không đầy đủ, tức là đồ thị ARG được biểu diễn bằng tập các cây thành phần và các sự kiện tái tổ hợp.

1.2.2. Bài toán xây dựng đồ thị ARG

Bài toán xây dựng đồ thị ARG được chứng minh là một bài toán NP-hard. Do số sự kiện tái tổ hợp và sự kiện đột biến cũng như vị trí thực sự xảy ra của chúng trong quá trình tiến hóa là không thể xác định được. Do đó, các hướng tiếp cận bài toán đều tập trung vào các giả định tối ưu số sự kiện tái tổ hợp và sự kiện đột biến. Dưới giả định các vị trí vô hạn, bài toán xây dựng đồ thị ARG được phát biểu như sau:

Cho một tập D gồm n trình tự nhị phân, mỗi trình tự có độ dài m, tìm một ARG hiển thị D với số sự kiện tái tổ hợp ít nhất.

Nhiều nghiên cứu xây dựng đồ thị ARG đã được đề xuất với các mô hình tái tổ hợp khác nhau phù hợp với quần thể quan sát và mục đích nghiên cứu khác nhau. Trong bài toán xây dựng đồ thị ARG cho các quần thể vi khuẩn, các sự kiện tái tổ hợp được xem xét và mô hình hóa là các sự kiện chuyển đổi gen. Trong nghiên cứu di truyền quần thể người, sự kiện tái tổ hợp được mô hình hóa trong quá trình xây dựng đồ thị ARG hầu hết là sự kiện trao đổi chéo. Trong nhiều thuật toán, đặc biệt là các thuật toán tổ hợp tập trung vào đặc điểm cấu trúc của đồ thị, sự kiện chuyển đổi gen có thể được biểu diễn qua 2 sự kiện trao đổi chéo liên tiếp nhau.

Trong khuôn khổ luận án này, chúng tôi tập trung vào các thuật toán tổ hợp xây dựng đồ thị ARG đầy đủ có số sự kiện tái tổ hợp ít nhất dưới giả định mô hình các vị trí vô hạn. Sự kiện tái tổ hợp trong đồ thị ARG được đề cập đến chỉ sự kiện trao đổi chéo và được sử dụng với ý nghĩa như vậy trong suốt các phần tiếp theo của luận án. Dữ liệu trình tự được xét đến trong bài toán là dữ liệu haplotype được biểu diễn ở dạng nhị phân.

Dữ liệu vào: Dữ liệu đầu vào là một tập các trình tự nhị phân độ dài m. Các trình tự có độ dài bằng nhau. Tập các trình tự được ký hiệu là D = $\{S_1, ..., S_N\}$, trong đó N là số lượng trình tự, S_x là một trình tự trong tập D, $1 \le x \le N$. S_x có độ dài m, $S_x[i]$ biểu thị giá trị của S_x tại vị trí i, $S_x[i]$ có giá trị bằng 0 hoặc 1, $1 \le i \le m$.

Bài toán: Tìm đồ thị ARG mô tả mối quan hệ của các trình tự trong tập dữ liệu vào thông qua 3 sự kiện: đột biến, kết hợp và tái tổ hợp, với giả định chỉ có nhiều nhất một đột biến xảy ra tại mỗi vị trí. Do có nhiều phương pháp khác nhau cho kết quả với độ hợp lý cũng như thời gian thực hiện khác nhau, chúng ta cần đề xuất các phương pháp cho kết quả tốt dựa trên các tiêu chí về số sự kiện tái tổ hợp ít nhất, khả thi với dữ liệu lớn hàng trăm đến hàng nghìn trình tự độ dài hệ gen, đồ thị có ứng dụng tốt trong các bài toán thực tế và có thời gian thực hiện khả thi.

Dữ liệu đầu ra: Đồ thị ARG chứa các thông tin quan hệ dưới dạng 3 sự kiện cơ bản: đột biến, kết hợp và tái tổ hợp giữa các trình tự đầu vào (nút lá) với các trình tự trung gian được sinh ra trong quá trình xây dựng đồ thị (nút cây) và với một trình tự tổ tiên chung duy nhất (nút gốc).

1.3. Các phương pháp xây dựng đồ thị ARG

Có 2 hướng nghiên cứu xây dựng đồ thị ARG: (1) Xây dựng đồ thị ARG tối thiểu (minimal ARG), tức là đồ thị có chính xác số sự kiện tái tổ hợp nhỏ nhất, và (2) xây dựng đồ thị ARG "hợp lý" (plausible ARG), tức là các thuật toán không cố gắng xây dựng ARG có chính xác số sự kiện tái tổ hợp ít nhất mà hướng đến việc xây dựng đồ thị ARG với số sự kiện tái tổ hợp được sinh ra phụ thuộc vào các phương pháp mô hình hóa sự kiện tái tổ hợp khác nhau.

1.3.1. Các phương pháp xây dựng đồ thị ARG tối thiểu

Các cách tiếp cận theo hướng nghiên cứu này hầu hết đều dựa trên các phương pháp tìm kiếm vét cạn trên đồ thị để cực tiểu hóa số sự kiện tái tổ hợp nhằm đạt tới ARG tối thiểu. Trong đó, khái niệm cặp vị trí không tương thích được sử dụng trong hầu hết các thuật toán để xác định sự kiện tái tổ hợp: Cho một tập D gồm 4 hoặc nhiều hơn 4 trình tự, một cặp vị trí bất kì gọi là không tương thích nếu tồn tại 4 trình tự trong D lần lượt chứa 4 loại giao tử (0,0), (0,1), (1,0), (1,1) cho cặp vị trí đó. Dưới giả định các vị trí vô hạn (có nhiều nhất một đột biến xảy ra tại một vị trí), cách duy nhất để có cặp vị trí không tương thích là ít nhất một sự kiện tái tổ hợp đã xảy ra trong lịch sử giữa 2 vị trí đó.

Khái niệm cặp vị trí không tương thích này là yếu tố cơ bản dẫn tới rất nhiều thuật toán tìm cận dưới tái tổ hợp và thuật toán xây dựng đồ thị ARG tối thiểu. Các phương pháp vét cạn hướng tới việc tìm ra các điểm cắt tái tổ hợp tối ưu, tức là, số sự kiện tái tổ hợp ít nhất để phá vỡ tất cả các vị trí không tương thích này.

Song và cộng sự xây dựng đồ thị ARG bằng cách duyệt qua tất cả các cây qua các vị trí. Các sự kiện tái tổ hợp cần thiết để chuyển từ tất cả các cây tại một vị trí sang tất cả các cây tại vị trí tiếp theo được tính toán. Các đồ thị ARG tối thiểu sau đó được xây dựng bằng cách lần theo các vị trí mà có số sự kiện tái tổ hợp ít nhất. Thay vì tính toán từ trái qua phải dọc theo chuỗi trình tự, Lyngsø và cộng sự sử dụng phương pháp nhánh cận, xây dựng đồ thị ARG ngược chiều thời gian, thực hiện các sự kiện đột biến, kết hợp và tái tổ hợp cho đến khi đến một tổ tiên chung tối ưu. Tìm kiếm phân nhánh được thực thi để khám phá tất cả các chuỗi sự kiện có thể, cố gắng tìm một chuỗi sự kiện với một số sự kiện tái tổ hợp cho trước. Nếu không tìm được, số sự kiện tái tổ hợp cho phép được tăng thêm một và cứ như vậy cho đến khi một đồ thị ARG được tìm thấy. Gusfield và cộng sự đề xuất thuật toán xây dựng một trường hợp đặc biệt của đồ thị ARG nếu có - đồ thị ARG với ràng buộc tất cả các chu trình tái tổ hợp không chung nút với nhau. Khi đó, đồ thị ARG là một cây có nốt sùi (galled-tree) trong đó mọi chu trình tái tổ hợp là các nốt sùi (gall) thỏa mãn không nốt sùi nào chung nút với nốt sùi nào.

Wu và cộng sự đưa bài toán xây dựng đồ thị ARG về bài toán tìm số trình tự trung gian tối thiểu cần để xây dựng ARG. Gần đây, Cámara và cộng sự đã đề xuất một kiểu đồ thị tổng hợp mới gọi là topological ARG. Tuy nhiên, các thuật toán xây dựng đồ thị ARG tối thiểu đều mới chỉ hạn chế áp dụng với các tập dữ liệu nhỏ, đến 100 trình tự ngắn, chưa khả thi với dữ liệu hệ gen người.

1.3.2. Các phương pháp xây dựng đồ thị ARG hợp lý

Các phương pháp tìm ARG tối thiểu chỉ áp dụng được cho các bộ dữ liệu nhỏ và độ phức tạp tính toán lớn. Để tương tác được với dữ liệu lớn hơn, các phương pháp xây dựng đồ thị ARG hợp lý đã được đề xuất. Theo hướng nghiên cứu này, các phương pháp xây dựng đồ thị ARG thường theo 2 cách tiếp cận chính là dựa trên kinh nghiệm và dựa trên thống kê.

Chương trình SHRUB xây dựng thuật toán tính cận trên tái tổ hợp R_{ub} và đồ thị ARG cho tập dữ liệu D sử dụng chính xác R_{ub} sự kiện tái tổ hợp bằng cách xây dựng đồ thị ARG lần lượt từ các nút lá. Các phép biến đổi kết hợp/thay thế các trình tự đầu vào được tiến hành song song tương ứng với các bước xây dựng đồ thị ARG cho đến khi đạt tới 1 nút chung duy nhất (chỉ còn lại một trình tự duy nhất qua các phép biến đổi).

Dựa trên ý tưởng từ thuật toán tìm ARG tối thiểu của Lyngso và cộng sự, Minichiello và Durbin đã đề xuất chiến lược mới để xác định sự kiện tái tổ hợp, đó là sự kiện tái tổ hợp được thực hiện trên cặp trình tự có đoạn chung dài nhất. Thuật toán chạy được với tập dữ liệu tối đa một nghìn trình tự có độ dài hàng trăm snp. Ý tưởng độ dài đoạn chung giữa 2 cá thể cũng được khai thác trong thuật toán xây dựng đồ thị ARG hợp lý của Parida và cộng sự.

Một cách tiếp cận khác gần đây là lấy mẫu (sampling) các ARG từ xác suất hậu nghiệm của các mô hình xấp xỉ quá trình kết hợp và tái tổ hợp (coalescent-with-recombination – CwR). Các thuật toán này cố gắng tích hợp quá trình kết hợp và tái tổ hợp vào các mô hình học máy để xây dựng tập hợp các cây phả hệ.

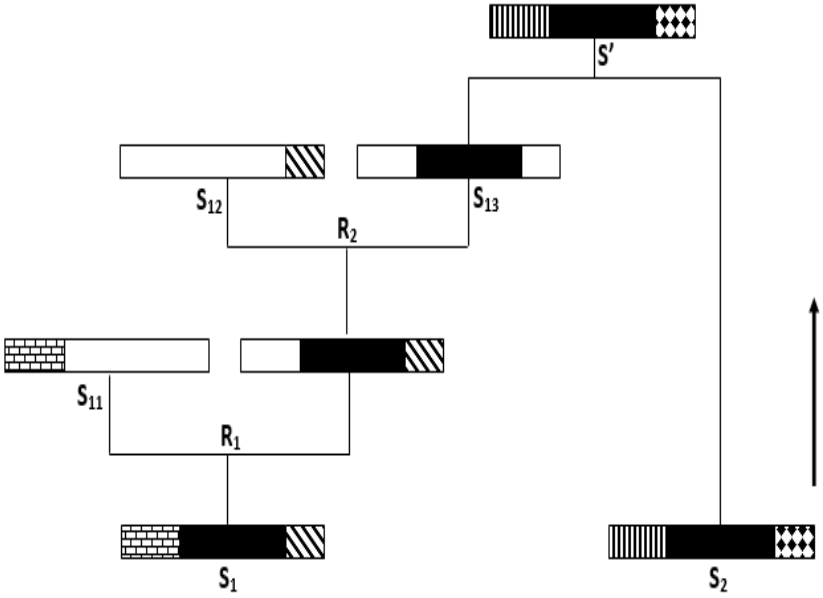
Các phương pháp theo cách tiếp cận thống kê là một hướng tiếp cận được nhiều nhà nghiên cứu phát triển gần đây. Tuy nhiên, các phương pháp này không suy luận được các ARG đầy đủ mà chỉ là tập các cây biên với tập các sự kiện tái tổ hợp tương ứng. Các phương pháp này thường được dùng trong việc mô phỏng dữ liệu. Hơn nữa, cách tiếp cận này rất phức tạp, đòi hỏi chi phí tính toán lớn nên vẫn chưa có được những ứng dụng thực tế trên những tập dữ liệu lớn.

Chương 2. THUẬT TOÁN ARG4WG XÂY DỰNG ĐỒ THỊ TÁI TỔ HỢP DI TRUYỀN CHO DỮ LIỆU LỚN

2.1. Giới thiệu

Qua khảo sát các phương pháp tìm ARG hợp lý, chúng tôi nhận thấy cách tiếp cận dựa trên kinh nghiệm của Minichiello và Durbin được cài đặt trong chương trình Margarita khả thi với tập dữ liệu một nghìn trình tự có độ dài hàng trăm SNP và đã có những ứng dụng vào một số bài toán thực tế. Tuy nhiên, thuật toán bị giới hạn với dữ liệu lớn do chiến lược thực hiện sự kiện tái tổ hợp trong quá trình xây dựng đồ thị ARG.

Để thực hiện bước tái tổ hợp, Margarita tìm một cặp trình tự có đoạn giống nhau liên tục dài nhất (longest shared tract) và thực hiện tái tổ hợp tại hai đầu của đoạn chung đó (xem Hình 2.1). Do đó, nếu đoạn chung được tìm thấy nằm bên trong trình tự, Margarita sẽ phải thực hiện 2 sự kiện tái tổ hợp, sinh ra 3 trình tự con từ 1 trình tự để có được trình tự chỉ chứa đoạn chung để thực hiện kết hợp với trình tự còn lại. Chiến lược này gây ra sự bùng nổ về số nút trong đồ thị khi số lượng sự kiện tái tổ hợp tăng.



Hình 2.1: Vấn đề trong việc thực hiện sự kiện tái tổ hợp của Margarita. Hai trình tự S_1 và S_2 với dải chung dài nhất giữa hai trình tự được biểu diễn bằng màu đen. Thuật toán thực hiện 1 cặp tái tổ hợp R_1 và R_2 trên trình tự S_1 để sinh ra 3 trình tự con S_{11}, S_{12} và S_{13}. Sau đó, S_{13} sẽ được kết hợp với S_2. Vì vậy, khi đoạn chung được tìm thấy bên trong trình tự, thuật toán phải thực hiện 2 sự kiện tái tổ hợp trên một trình tự và một cặp trình tự ban đầu sẽ biến thành 3 trình tự.

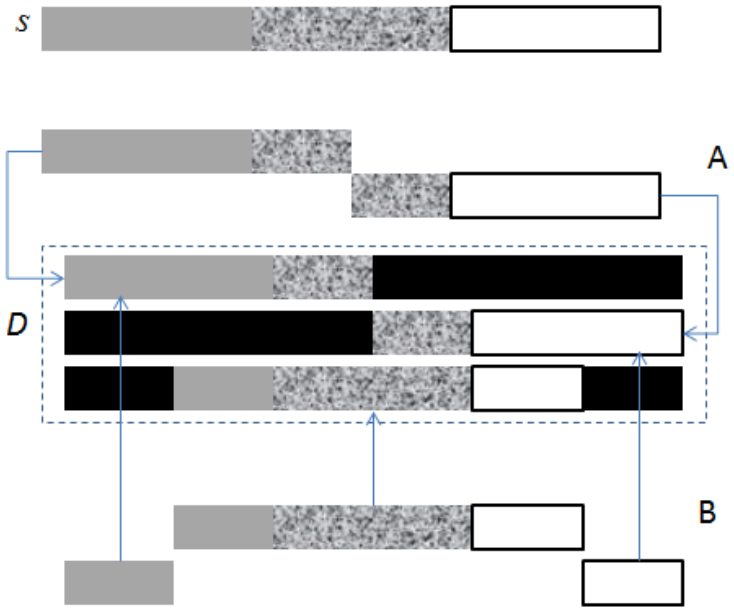
Luận án đề xuất thuật toán ARG4WG xây dựng đồ thị ARG hợp lý cho dữ liệu lớn hàng nghìn mẫu độ dài toàn nhiễm sắc thể. Cùng cách tiếp cận như Margarita, tuy nhiên, chúng tôi thực hiện tái tổ hợp theo chiến lược tìm đoạn đầu chung dài nhất. Các chứng minh, thực nghiệm và ứng dụng trên các bộ dữ liệu khác nhau đã chứng minh hiệu quả của thuật toán đề xuất.

2.2. Chiến lược tìm đoạn đầu chung dài nhất

Cho trước một tập trình tự D và một trình tự s, ta sẽ chứng minh rằng việc lấy lặp lại đoạn chung dài nhất tại một đầu của s mà có thể kết hợp với một trình tự trong D cho chúng ta số sự kiện tái tổ hợp ít nhất. Ta có thể lấy phía bên trái hoặc phía bên phải. Từ đó chỉ ra rằng chiến lược lấy đoạn chung dài nhất trong trình tự không phải luôn luôn cho ta số sự kiện tái tổ hợp ít nhất.

Mệnh đề 1: Cho một tập các trình tự trong D, và 1 trình tự s có cùng độ dài m. Số cực tiểu sự kiện tái tổ hợp, $f(s, D)$, để tách s thành các trình tự con mà có thể kết hợp với các trình tự trong D có thể đạt được bằng cách lặp lại việc lấy các đoạn dài nhất từ phía trái của s.

Chúng ta có thể có được cực tiểu số sự kiện tái tổ hợp bằng cách lặp lại việc lấy ra các đoạn chung dài nhất từ phía bên trái của s. Tương tự với trường hợp lấy từ phía bên phải. Và điều này là không đúng nếu chúng ta không chọn các đoạn chung dài nhất từ hai phía của s. Hình 2.1 mô tả giải pháp tối ưu mà chỉ cần một sự kiện tái tổ hợp (xem Kịch bản **A**). Tuy nhiên, nếu ta chọn đoạn chung dài nhất không phải từ 2 phía của s (ở đây là chọn đoạn chung dài nhất trong s) thì ta có thể phải cần đến 2 sự kiện tái tổ hợp (Kịch bản **B**).



Hình 2.2: Phân tách s bằng cách chọn các đoạn chung dài nhất trong s để kết hợp với các trình tự trong D có thể không dẫn tới số cực tiểu sự kiện tái tổ hợp.

Từ đó, chúng tôi định nghĩa *đoạn đầu chung dài nhất* (longest shared end) là đoạn chứa thông tin di truyền giống nhau liên tục dài nhất tính từ 2 đầu của các trình tự.

2.3. Thuật toán ARG4WG

ARG4WG được xây dựng ngược chiều thời gian, xây dựng 1 ARG từ một tập các trình tự (haplotypes) cho tới khi đạt tới một tổ tiên chung. ARG4WG gồm 3 bước chính: Bước kết hợp, bước đột biến và bước tái tổ hợp.

Đầu tiên, thuật toán tìm các trình tự đồng nhất để thực hiện kết hợp. Bước này giúp giảm số lượng trình tự cho đến khi tới một tổ tiên chung duy nhất. Trong bước đột biến, thuật toán tìm các vị trí mà ở đó chỉ có một trình tự có giá trị khác với tất cả các trình tự còn lại. Kết quả của bước này có thể sinh ra các trình tự đồng nhất để thực hiện bước kết hợp. Khi không thực hiện được sự kiện kết hợp hay đột biến, thuật toán sẽ chuyển sang bước tái tổ hợp.

Để xác định điểm cắt tái tổ hợp, thuật toán sẽ chọn một cặp trình tự (S_1, S_2) có đoạn chung dài nhất từ 2 đầu. Giả sử S1 chứa ít vật liệu di truyền trong phần chung hơn S_2, thuật toán thực hiện một sự kiện tái tổ hợp bằng việc tách S_1 thành 2 trình tự con mới. Trình tự con chứa đoạn chung sẽ được kết hợp với S_2 ngay sau đó (xem Hình 2.3).

Đặt "*" là trạng thái không di truyền trong các nút trong của đồ thị ARG.

ĐẦU VÀO: Tập dữ liệu $D = \{S_1, ..., S_N\}$ trình tự (haplotype), S_x có m marker, $S_x[i]$ có giá trị bằng 0 hoặc 1, $1 \le x \le N$, $1 \le i \le m$.

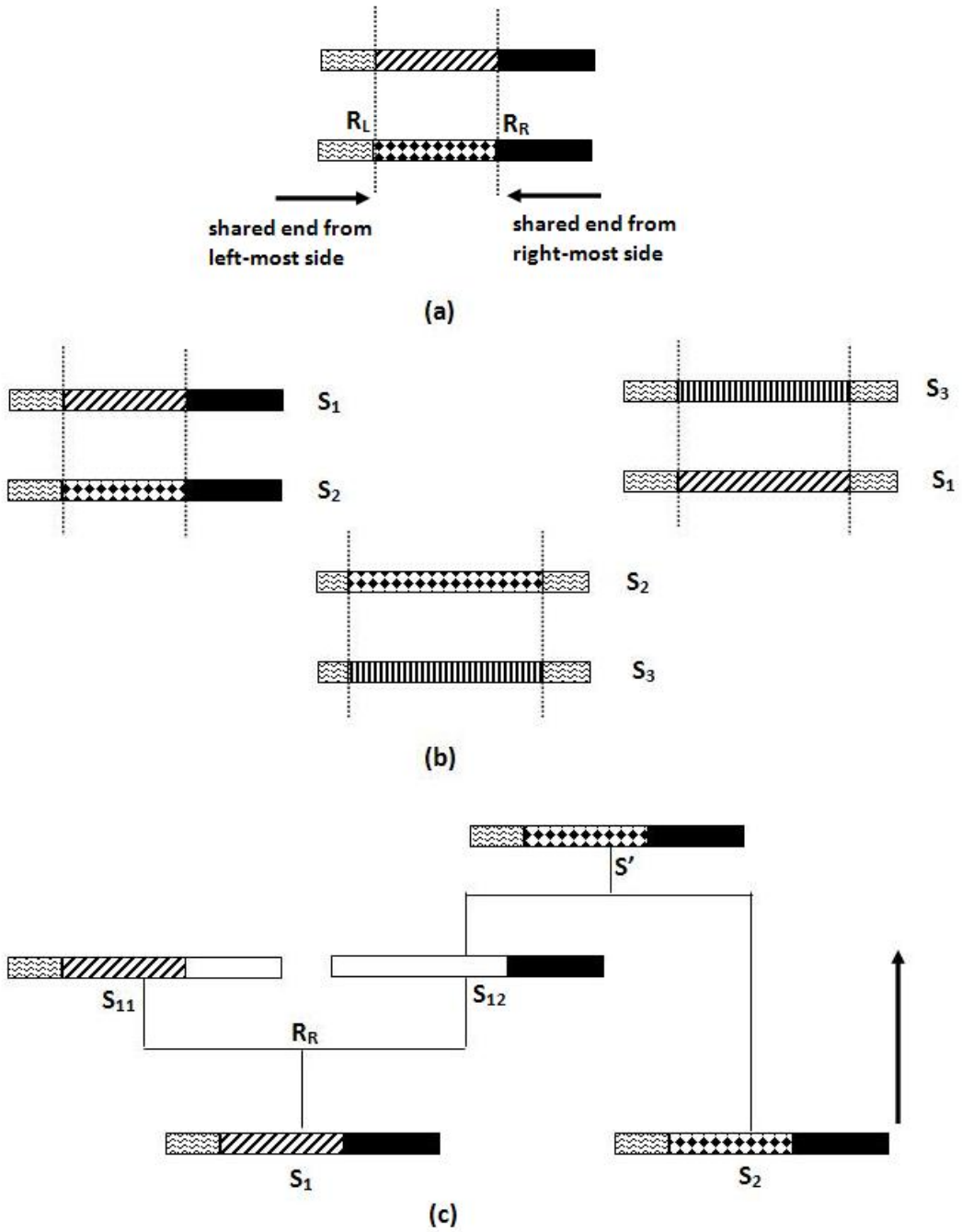
ĐẦU RA: một đồ thị ARG mô tả các mối quan hệ (các sự kiện kết hợp, đột biến, tái tổ hợp) giữa các nút (các trình tự) trong đồ thị đến một tổ tiên chung duy nhất S_{SCA}.

Một trình tự S_1 được coi là dài hơn trình tự S_2 ($L(S_1) > L(S_2)$) nếu S_1 chứa nhiều vật liệu di truyền hơn S_2. Ta cũng định nghĩa ($L(S_1) > L(S_2)$)[a,b] nếu S_1 dài hơn S_2 trong khoảng [a,b].

Một toán tử bù, ¬, được định nghĩa để nếu $S[i] = 0$ thì $\neg S[i] = 1$ và ngược lại, và * là phần bù của chính nó.

Với một cặp (S_1, S_2), đặt (S_1, S_2){d} là đoạn đầu chung của chúng. Cụ thể, (S_1, S_2){d=*left*} là phần chung của đầu bên trái của (S_1, S_2); (S_1, S_2){d=*right*} là phần chung của đầu bên phải của (S_1, S_2).

Chúng tôi định nghĩa $S_1[i]$ khớp với $S_2[i]$ nếu hoặc cả 2 trình tự có cùng trạng thái hoặc trạng thái của ít nhất một trong 2 trình tự là *.



Hình 2.3: Sự kiện tái tổ hợp được biểu thị trong thuật toán ARG4WG. (a) Xét 2 trình tự S_1 và S_2, các đoạn chung ở 2 đầu của 2 trình tự từ phía bên trái (hình lượn sóng) và từ phía bên phải (màu đen) được xác định. (b) Với 1 tập 3 trình tự S_1, S_2 và S_3, các đoạn chung ở 2 đầu của mỗi cặp được tính toán (hình lượn sóng) và đoạn đầu chung dài nhất được xác định được mô tả bằng màu đen. (c) Một sự kiện tái tổ hợp được thực hiện trên trình tự S_1 để sinh ra 2 trình tự con S_{11} và S_{12}. S_{12} chứa đoạn đầu chung dài nhất sau đó sẽ được kết hợp với S_2. Do đó, chỉ cần thực hiện một sự kiện tái tổ hợp và số trình tự không bị tăng lên trong quá trình xây dựng đồ thị ARG.

Chúng tôi định nghĩa cặp có đoạn đầu chung cực đại $(S_1,S_2)\{d,l\}$ với độ dài đoạn chung l $(0 < l \leq m)$ của S_1 và S_2 nếu nó thỏa mãn các điều kiện sau:

1. Nếu d = *left* thì $S_1[i]$ khớp với $S_2[i]$ với mọi $1 \leq i \leq l$ và hoặc $l = m$ hoặc $S_1[l+1]$ không giống $S_2[l+1]$.
2. Nếu d = *right* thì $S_1[i]$ khớp với $S_2[i]$ với mọi $m\text{-}l < i \leq m$ và hoặc $l = m$ hoặc $S_1[m\text{-}l]$ không giống $S_2[m\text{-}l]$.
3. Vùng giống nhau phải có ít nhất một vị trí i mà $S_1[i] = S_2[i] \neq *$.

Điều kiện đầu tiên và thứ 2 xác định rằng 2 trình tự là đồng nhất trên một đoạn đầu chung dài nhất của chúng. Điều kiện thứ 3 nhấn mạnh rằng đoạn đầu chung giữa 2 trình tự có chung ít nhất một vị trí mang vật liệu di truyền. Điều này làm giảm số các nhánh dư thừa trong quá trình xây dựng đồ thị ARG.

Xét cặp trình tự (S_1, S_2) (không đồng nhất) có các đoạn đầu chung cực đại từ phía bên trái và từ phía bên phải tương ứng là l_L và l_R. Nếu đoạn đầu chung cực đại từ phía bên phải chứa nhiều phần vật liệu di truyền chung hơn đoạn đầu chung cực đại từ phía bên trái thì l_R được xác định là đoạn đầu chung dài nhất của cặp trình tự (S_1, S_2).

Thuật toán bắt đầu từ thời gian $t = 1$. Tập các trình tự tại thời gian t được ký hiệu là D_t ($D_1 = D$). Với mỗi D_t chúng tôi xây dựng 3 danh sách ứng cử viên cho các sự kiện kết hợp, đột biến và tái tổ hợp như sau:

- Danh sách kết hợp **C**: Với một cặp có đoạn đầu chung dài nhất $(S_1,S_2)\{d,l\}$, nếu $l = m$, ta cho cặp này vào danh sách kết hợp.
- Danh sách đột biến **M**: Với một vị trí i $(1 \leq i \leq m)$, nếu tồn tại duy nhất một trình tự S_1, và với mọi trình tự S_2 trong $D_t \setminus \{S_1\}$ ta có $S_2[i] = \neg S_1[i]$ thì $S_1[i]$ được cho vào danh sách đột biến.
- Danh sách tái tổ hợp **R**: Với một cặp có đoạn đầu chung dài nhất $(S_1,S_2)\{d,l\}$, nếu $0 < l < m$, $(S_1,S_2)\{d,l\}$ được cho vào danh sách tái tổ hợp.

Khi một trong ba sự kiện xuất hiện, tập trình tự tiếp theo D_{t+1} được tạo ra từ tập trình tự hiện thời D_t và 3 danh sách ứng cử viên được cập nhật.

BEGIN

$t = 1$; $D_t = D$;

Gán danh sách kết hợp **C** = {tất cả các cặp $(S_x,S_y)\{d,l\}$ $(1 \leq x, y \leq N)$ có $l = m$};

Gán danh sách đột biến **M** = {tất cả các trình tự chứa các vị trí đột biến đơn};

Gán danh sách tái tổ hợp **R** = {tất cả các cặp $(S_x,S_y)\{d,l\}$ $(1 \leq x, y \leq N)$ có $0 < l < m$ };

while chưa đạt tới một tổ tiên chung duy nhất **do**

if (danh sách kết hợp **C** không rỗng) **then**

Lấy ngẫu nhiên một cặp trình tự có đoạn đầu chung (S_1,S_2);

Thực hiện kết hợp như sau:

Gán $S' = S_1$ nếu $L(S_1) > L(S_2)$; ngược lại $S' = S_2$;

$D_{t+1} = (D_t \setminus \{S_1,S_2\}) \cup \{S'\}$;

Cập nhật 3 danh sách **C, M, R**;

else

if (danh sách đột biến **M** không rỗng) **then**

Lấy ngẫu nhiên một trình tự S với một đột biến tại vị trí i;

Thực hiện sự kiện đột biến như sau:

$D_{t+1} = (D_t \setminus \{S\}) \cup \{S'\}$ với $S'[i] = \neg S[i]$ và $S'[j] = S[j]$ với mọi $j \neq i$ và $1 \le i,j \le m$

Cập nhật 3 danh sách **C, M, R**;

else

Lấy cặp trình tự có đoạn đầu chung dài nhất $(S_1,S_2)\{d,l\}$ từ danh sách tái tổ hợp;

Thực hiện tái tổ hợp như sau:

if $d = left$ **then** // *đoạn đầu chung dài nhất của* (S_1, S_2) *là từ đầu phía bên trái*

Gán $S_R = S_1$ nếu $(L(S_1) < L(S_2))[1,l]$; ngược lại $S_R = S_2$;

$S_{R1}[i] = S_R[i]$ với mọi $1 \le i \le l$; $S_{R1}[j] = *$ với mọi $l < j \le m$

$S_{R2}[i] = *$ với mọi $1 \le i \le l$; $S_{R2}[j] = S_R[j]$ với mọi $l < j \le m$

else // $d = right$

Gán $S_R = S_1$ nếu $(L(S_1) < L(S_2))[m\text{-}l\text{+}1,m]$; ngược lại $S_R = S_2$;

$S_{R1}[i] = *$ với mọi $1 \le i \le m\text{-}l$; $S_{R1}[j] = S_R[j]$ với mọi $m\text{-}l < j \le m$

$S_{R2}[i] = S_R[i]$ với mọi $1 \le i \le m\text{-}l$; $S_{R2}[j] = *$ với mọi $m\text{-}l < j \le m$

endif

$D_{t+1} = (D_t \setminus \{S_R\}) \cup \{S_{R1},S_{R2}\}$;

Cập nhật 3 danh sách **C, M, R**;

endif

endif

endwhile

END;

Thuật toán 2.1: Thuật toán ARG4WG xây dựng một đồ thị ARG từ một tập trình tự D cho trước.

Như vậy, một sự kiện kết hợp làm giảm số trình tự đi một. Sự kiện đột biến xuất hiện chỉ tại một vị trí đơn. Một sự kiện tái tổ hợp chỉ thay thế một trình tự bằng một trình tự mới có ít vật liệu di truyền hơn con của nó. Chính vì vậy, ARG4WG luôn đạt tới được một tổ tiên chung duy nhất.

Những lựa chọn ngẫu nhiên trong các bước của thuật toán dẫn tới việc sinh ra các đồ thị ARG khác nhau cho mỗi lần chạy. Thuật toán ARG4WG đã đơn giản hóa cách thực hiện sự kiện tái tổ hợp nhưng nó vẫn suy luận ra những đồ thị ARG hợp lý. So với Margarita, chiến lược đoạn đầu chung dài nhất này không những

cho ta số sự kiện tái tổ hợp ít hơn mà còn làm giảm thời gian tìm đoạn chung dài nhất và số lượng nút trong quá trình xây dựng đồ thị.

2.4. Kết quả

Các thực nghiệm trên các bộ dữ liệu khác nhau đã cho thấy hiệu quả của thuật toán đề xuất. Mặc dù có hình thái cây kém hơn một chút so với Margarita khi so sánh trên dữ liệu mô phỏng nhưng ARG4WG nhanh hơn hàng nghìn lần so với Margarita. Các kết quả thực nghiệm cũng cho thấy số sự kiện tái tổ hợp của Margarita nhiều hơn trung bình là 1.4 lần so với ARG4WG. Đặc biệt, ARG4WG có thể sinh ra 1 ARG với thời gian ~4.5 giờ trong 1 lần chạy sử dụng 1 máy tính 16-thread cho dữ liệu 4246 haplotype (2123 mẫu gen người) trên toàn nhiễm sắc thể 1 (174,234 SNPs – nhiễm sắc thể dài nhất trong bộ gen người) từ dự án 1kGP. Kết quả này nói lên rằng thuật toán ARG4WG đề xuất có thể chạy được với dữ liệu lớn hàng nghìn trình tự trên toàn hệ gen.

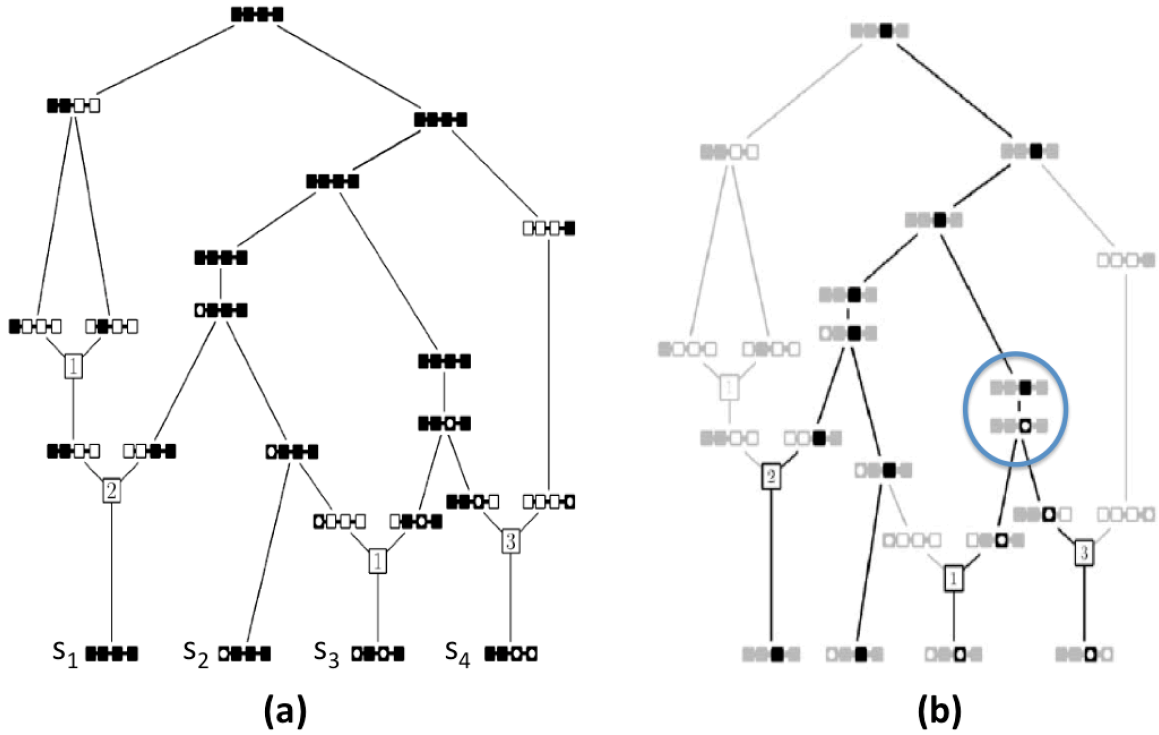
2.5. Ứng dụng ARG4WG trong nghiên cứu tương quan toàn hệ gen

Trong nghiên cứu tương quan người bệnh-người không bệnh (case-control association study), các tần số alen tại các vị trí quan tâm được so sánh trong các quần thể gồm các cá thể bị bệnh và các cá thể không bị bệnh. Tần số trong người bị bệnh mà cao hơn là minh chứng cho alen đó liên quan đến nguy cơ gây bệnh tăng lên. Bằng việc phân tích sự phân biệt của các alen SNP giữa các quần thể người bệnh và người không bệnh này ta có thể xác định được các vị trí có liên quan một cách thống kê tới bệnh.

2.5.1 Cách tiếp cận sử dụng đồ thị ARG vào bài toán tìm ánh xạ tương quan

Di chuyển dọc theo nhiễm sắc thể, các hình thái của các cây biên liên tiếp nhau dịch chuyển theo tác động của các sự kiện tái tổ hợp mang tính lịch sử. Các sự kiện tái tổ hợp định nghĩa vùng nhiễm sắc thể mà mỗi cây biên mở rộng. Với một vị trí cho trước, cây biên có thể được trích xuất từ ARG bằng cách lần vết phả hệ của vị trí đó ngược chiều thời gian từ các nút lá. Khi một tái tổ hợp xuất hiện, phả hệ sẽ theo đường của thế hệ cha mẹ phía bên trái nếu điểm cắt tái tổ hợp nằm phía phải của vị trí được xét, và ngược lại thì đi theo thế hệ cha mẹ phía bên phải.

Nếu có một đột biến nguy cơ gây bệnh tại một vị trí cụ thể trên nhiễm sắc thể (giả sử đột biến chỉ xuất hiện nhiều nhất một lần tại mỗi vị trí trong suốt lịch sử tiến hóa), nó sẽ xảy ra trên một số nhánh bên trong cây biên tại vị trí đó. Vì vậy, một cách để tìm các tương quan đến bệnh là kiểm tra các cây biên để tìm những cây có nhánh phân biệt rõ nhất giữa người bệnh và người không bệnh, tức là các nhánh mà ở đấy nhiều người bệnh và chỉ số rất ít hoặc không có người không bệnh thuộc nhánh đó. Cụm các người bệnh tập trung vào một nhánh như vậy sẽ gợi ý rằng có một đột biến gây bệnh xuất hiện trên nhánh đó (Hình 2.4).



Hình 2.4: (a) Đồ thị ARG cho tập 4 trình tự, trong đó trình tự s_1, s_2 là từ 2 cá thể khỏe mạnh, trình từ s_3, s_4 là từ 2 cá thể bị bệnh. (b) Đột biến 3 (vùng khoanh tròn) trên cây biên tại vị trí 3 của đồ thị ARG trong (a) cho ra sự phân biệt rõ nhất giữa các trình tự bệnh và trình tự không bệnh.

Cách làm ánh xạ tương quan sử dụng đồ thị ARG được tóm tắt như sau:

1. Với một tập D các haplotype từ các cá thể người bệnh và người không bệnh, xây dựng đồ thị ARG $\mathcal{G}$ cho D sử dụng thuật toán xây dựng đồ thị ARG.
2. Tìm tập các cây biên $\mathcal{T}$ của $\mathcal{G}$ tại các vị trí của D.
3. Các cạnh e trong cây T $\in \mathcal{T}$ được tính điểm độ tốt (theo cách nào đó) trong việc phân biệt các lá gán nhãn bệnh với các lá gán nhãn không bệnh. Sau đó, cạnh e với điểm cao nhất được thiết lập cho T.
4. Đặt T là cây biên trong $\mathcal{T}$ có độ tương quan lớn nhất, tức là chứa cạnh e với độ phân biệt lớn nhất giữa nhãn bệnh và nhãn không bệnh trong tất cả các cạnh trong tất cả các cây biên trong $\mathcal{T}$. Nếu T đủ tốt (đạt một ngưỡng cho trước), kết luận là đột biến gây bệnh có khả năng xảy ra quanh vị trí cây biên T và đột biến xuất hiện có thể xảy ra trong thời gian được biểu thị bởi cạnh e được tìm thấy trong T.

2.5.2 Ứng dụng ARG4WG vào bài toán tìm vùng gen liên quan đến bệnh sốt rét ở Châu Phi

ARG4WG được sử dụng để xây dựng các đồ thị ARG từ tập dữ liệu Gambia chứa 5560 haplotypes (tương ứng với 2780 mẫu cá thể, trong đó có 1533 mẫu khỏe mạnh và 1247 mẫu bị bệnh sốt rét) trên toàn nhiễm sắc thể 11 (Band et al. 2013). Chương trình Margarita xây dựng thuật toán tìm ánh xạ tương quan dựa trên đồ thị ARG hợp lý được sử dụng để thử nghiệm các đồ thị ARG kết quả từ thuật toán ARG4WG vào bài toán tìm ánh xạ tương quan toàn hệ gen.

Các kết quả thực nghiệm cho kết quả vùng có tín hiệu mạnh liên quan đến bệnh là từ 4.43Mb tới 6.28Mb trên nhiễm sắc thể 11 với p-values $\leq 10^{-7}$. Kết quả này đồng ý với phân tích của nhóm tác giả Garvin Band (Band et al. 2013) là vùng HBB (4.5Mb-5.5Mb) có nhiều khả năng liên quan đến bệnh sốt rét. Trong nghiên cứu của họ, họ đã chỉ ra giá trị P-value thấp nhất trong vùng này là 5.7×10^{-13} sử dụng phương pháp phân tích SNPTEST meta-analysis. Tuy nhiên, do hạn chế của phương pháp kiểm thử hoán vị trong công cụ Margarita nên ta không thể thực hiện phân tích sâu đến $>10^7$. Sự thống nhất trong kết quả của các thực nghiệm cũng chỉ ra rằng ARG4WG rất ổn định trong việc xây dựng ARG với số lượng SNP khác nhau.

Các kết quả nghiên cứu của chương này đã được công bố trong một bài báo trên tạp chí quốc tế *IEEE/ACM Transactions on Computational Biology and Bioinformatics* năm 2017 (công trình khoa học số 1).

Chương 3. PHƯƠNG PHÁP TỐI ƯU HÓA SỐ SỰ KIỆN TÁI TỔ HỢP TRONG QUÁ TRÌNH XÂY DỰNG ĐỒ THỊ ARG

Thuật toán ARG4WG do chúng tôi đề xuất được giới thiệu trong chương 2 có thể xây dựng được đồ thị ARG cho dữ liệu lớn hàng nghìn mẫu trên toàn hệ gen. Tuy nhiên, thuật toán không được thiết kế nhằm tối ưu hóa số sự kiện tái tổ hợp.

Trong chương này, chúng tôi trình bày 2 phương pháp: (1) kết hợp các đặc trưng của dữ liệu và (2) kết hợp các kĩ thuật tối ưu vào thuật toán ARG4WG nhằm tối ưu hóa số sự kiện tái tổ hợp trong quá trình xây dựng đồ thị ARG.

3.1. Một số định nghĩa và khái niệm sử dụng trong các thuật toán

Dưới giả định các vị trí vô hạn, ta gọi 2 vị trí i và j là *không tương thích* nếu chúng chứa tất cả 4 loại giao tử 00, 01, 10, 11. Sẽ có ít nhất một sự kiện tái tổ hợp giữa 2 vị trí không tương thích i và j.

Đặt $D = \{S_1, S_2, \ldots, S_N\}$ là tập N trình tự nhị phân có độ dài m, $S_x[i]$ có giá trị bằng 0 hoặc 1, $1 \leq x \leq N$, $1 \leq i \leq m$; gọi * là trạng thái không di truyền, tức là không

mang thông tin di truyền từ dữ liệu quan sát. Chúng tôi sử dụng một số định nghĩa giống như trong thuật toán ARG4WG như sau:

- Xét 1 vị trí i, $S_x[i]$ khớp với $S_y[i]$ nếu $S_x[i] = S_y[i]$ hoặc $S_x[i] = *$ hoặc $S_y[i] = *$.
- $(S_x,S_y)\{d,l\}$ là một cặp trình tự S_x và S_y có đoạn đầu chung với độ dài tối đa l từ phía bên trái ($d = left$) hoặc từ phía bên phải ($d = right$).
- $(S_x,S_y)\{d,l\}$ tồn tại nếu và chỉ nếu có ít nhất một vị trí i trong phần chung thỏa mãn $S_x[i] = S_y[i] \neq *$.
- Cặp (S_x,S_y) gọi là cặp có đoạn đầu chung dài nhất nếu cặp đó chứa phần vật liệu di truyền chung dài nhất trong đoạn đầu chung.

Với 1 cặp có đoạn đầu chung $(S_x,S_y)\{d,l\}$, theo chiến lược đoạn đầu chung dài nhất thì điểm cắt tái tổ hợp được xác định giữa:

- l và $l + 1$ khi $d = left$ và $S_x[i]$ khớp với $S_y[i]$ với mọi $1 \leq i \leq l$ và $S_x[l+1] \neq S_y[l+1]$.
- l -1 và l khi $d = right$ và $S_x[i]$ khớp với $S_y[i]$ với mọi $l \leq i \leq m$ và $S_x[l-1] \neq S_y[l-1]$.

Cũng giống với ARG4WG, các thuật toán đề xuất dưới đây đều hoạt động ngược thời gian và có cùng giả định có nhiều nhất một đột biến xảy ra tại mỗi vị trí trong suốt quá trình xây dựng đồ thị ARG.

3.2. Hạn chế của thuật toán ARG4WG trong bài toán xây dựng đồ thị ARG tối thiểu

Phần này chỉ ra hạn chế của chiến lược đoạn đầu chung dài nhất trong việc xây dựng đồ thị ARG tối thiểu.

Chiến lược đoạn đầu chung dài nhất giúp cho ARG4WG chạy được với dữ liệu lớn gồm hàng nghìn trình tự độ dài toàn hệ gen. Tuy nhiên, nhiều khi cách chọn điểm cắt tái tổ hợp theo chiến lược này không giúp phá vỡ bất kì 1 cặp vị trí không tương thích nào, dẫn đến thuật toán không xây dựng được ARG tối thiểu.

3.3. Thuật toán REARG

3.3.1 Động cơ nghiên cứu

Xuất phát từ các quan sát trong quá trình làm thực nghiệm, chúng tôi nhận thấy rằng việc lựa chọn cặp trình tự có độ dài đoạn đầu chung dài nhất cho việc thực hiện tái tổ hợp trong thuật toán ARG4WG thường là không duy nhất. Nói cách khác, ARG4WG thường phải chọn ngẫu nhiên 1 cặp trình tự cho việc thực hiện tái tổ hợp từ rất nhiều cặp có cùng độ dài đoạn đầu chung dài nhất.

Các phân tích thực nghiệm của chúng tôi cho thấy, bên cạnh tiêu chí độ dài đoạn đầu chung dài nhất, các yếu tố khác như độ tương đồng của cặp trình tự được chọn hay độ dài của trình tự được chọn để thực hiện tái tổ hợp cũng có ảnh hưởng đáng kể đến số sự kiện tái tổ hợp. Do đó, việc kết hợp các yếu tố này trong việc lựa chọn cặp trình tự thích hợp nhất cho việc tái tổ hợp có thể giúp định hướng quá

trình xây dựng đồ thị ARG tới đồ thị ARG với số sự kiện tái tổ hợp tối ưu hơn khi chạy với dữ liệu lớn với số lần chạy giới hạn.

3.3.2 Thuật toán REARG

Chúng tôi định nghĩa:

Độ tương đồng của 2 trình tự S_1 và S_2:

$$Sim(S_1, S_2) = \sum_{1}^{m} Sim(S_1[i], S_2[i])$$

Với

$$Sim(S_1[i], S_2[i]) = \begin{cases} 1 & if \ \ S_1[i] = S_2[i] \neq * \\ 0 & otherwise \end{cases}$$

Độ dài của trình tự S:

$$Len(S) = \sum_{1}^{m} Len(S[i])$$

Với

$$Len(S[i]) = \begin{cases} 1 & if \ \ S[i] \neq * \\ 0 & if \ \ S[i] = * \end{cases}$$

Trong thuật toán REARG, các thủ tục cho bước kết hợp và đột biến giống như trong thuật toán ARG4WG. Chúng tôi sử dụng thêm một số tiêu chuẩn khác để lựa chọn ứng cử viên tốt nhất cho bước tái tổ hợp. Dưới đây, chúng tôi mô tả 3 phiên bản khác nhau của thuật toán REARG: REARG_SIM, REARG_LEN và REARG_COM.

Bước tái tổ hợp của thuật toán REARG_SIM

- Bước 1: Tính độ dài của các đoạn đầu chung cho tất cả các cặp trình tự. Các cặp trình tự có đoạn đầu chung dài nhất được chọn là các cặp ứng cử viên cho việc tái tổ hợp.
- Bước 2: Tính độ tương đồng của tất cả các cặp ứng cử viên. Chọn cặp ứng cử viên có độ tương đồng cao nhất để thực hiện tái tổ hợp. Trong trường hợp có nhiều ứng cử viên có cùng độ tương đồng cao nhất, một cặp trong số đó sẽ được chọn ngẫu nhiên để thực hiện tái tổ hợp.

Bước tái tổ hợp của thuật toán REARG_LEN

- Bước 1: Tính độ dài của các đoạn đầu chung cho tất cả các cặp trình tự. Các cặp trình tự có đoạn đầu chung dài nhất được chọn là các cặp ứng cử viên cho việc tái tổ hợp.
- Bước 2: Tính độ dài của trình tự ngắn hơn của tất cả các cặp ứng cử viên. Chọn ứng cử viên có độ dài trình tự dài nhất để thực hiện tái tổ hợp. Trong

trường hợp có nhiều ứng cử viên có cùng độ dài trình tự dài nhất, một trong số đó sẽ được chọn ngẫu nhiên để thực hiện tái tổ hợp.

Bước tái tổ hợp của thuật toán REARG_COM

- Bước 1: Tính độ dài của các đoạn đầu chung cho tất cả các cặp trình tự. Các cặp trình tự có đoạn đầu chung dài nhất được chọn là các cặp ứng cử viên cho việc tái tổ hợp.
- Bước 2: Tính độ tương đồng của tất cả các cặp ứng cử viên và tính độ dài của trình tự ngắn hơn của các cặp ứng cử viên.
- Bước 3: Chọn ngẫu nhiên một cặp ứng cử viên có độ tương đồng cao nhất hoặc một ứng cử viên có độ dài trình tự dài nhất để thực hiện tái tổ hợp.

3.4. Thuật toán GAMARG

3.4.1 Động cơ nghiên cứu

Do chiến lược đoạn đầu chung dài nhất không dẫn đến số sự kiện tái tổ hợp cực tiểu, nên ý tưởng đặt ra là kết hợp ARG4WG với các tiêu chí tối ưu khác để giảm số sự kiện tái tổ hợp. Đáng chú ý, kiểm thử 4 giao tử (four-gamete test) là ý tưởng then chốt dẫn đến nhiều thuật toán khác nhau trong bài toán tìm cận dưới số sự kiện tái tổ hợp và trong bài toán xây dựng đồ thị ARG có chính xác số sự kiện tái tổ hợp nhỏ nhất. Do đó, chúng tôi đề xuất thuật toán GAMARG kết hợp giữa ràng buộc trong kiểm thử 4 giao tử với chiến lược đoạn đầu chung dài nhất trong ARG4WG để tối ưu hóa số sự kiện tái tổ hợp trong quá trình xây dựng đồ thị ARG. Các kết quả thực nghiệm trên các tập dữ liệu khác nhau cho thấy GAMARG có thể chạy được với hàng nghìn trình tự với hàng chục nghìn snp và có thể đạt đến ARG với số sự kiện tái tổ hợp nhỏ nhất.

3.4.2 Thuật toán GAMARG

Các phương pháp vét cạn hướng tới việc tìm ra các điểm cắt tái tổ hợp tối ưu, tức là, số sự kiện tái tổ hợp ít nhất để phá vỡ tất cả các cặp vị trí không tương thích. Tuy nhiên, việc quét tất cả các khả năng có thể để đưa ra phương án tối ưu là không khả thi với các tập dữ liệu vừa và lớn. Do đó, chúng tôi đưa ra một số quan sát trong quá trình xây dựng ARG sử dụng kiểm thử 4 giao tử, từ đó dẫn đến một số mở rộng được đề xuất khi áp dụng kiểm thử 4 giao tử vào thuật toán.

Đặt $FreqGamete_{i,j} = \{freq00_{i,j}, freq01_{i,j}, freq10_{i,j}, freq11_{i,j}\}$ là tần số của các loại giao tử 00, 01, 10, 11 xuất hiện giữa vị trí i và vị trí j.

Đặt δ là kích thước của cửa sổ trượt mà chúng tôi sẽ quét để tìm tất cả các cặp vị trí không tương thích trong vùng đó. Cụ thể, chúng tôi sẽ quét qua tất cả các vị trí. Với mỗi vị trí i ($0 \leq i < m$), chúng tôi sẽ quét để tìm tất cả các cặp vị trí không tương thích trong phạm vi $[i, i+\delta]$.

Đặt $S_x(i,j)$ là một trình tự có loại giao tử có tần số bằng 1 giữa cặp vị trí không tương thích i và j ($0 \le i < m, j - i \le \delta$). Tức là, $S_x(i,j)$ thỏa mãn điều kiện sau:

$$\begin{cases} freq00_{i,j} > 0 \; and \; freq01_{i,j} > 0 \; and \; freq10_{i,j} > 0 \; and \; freq11_{i,j} > 0 \\ freq00_{i,j} = 1 \; or \; freq01_{i,j} = 1 \; or \; freq10_{i,j} = 1 \; or \; freq11_{i,j} = 1 \end{cases}$$

Khi đó, nếu ta thực hiện tái tổ hợp trên trình tự S_x giữa vị trí i và j ta sẽ phá vỡ được ít nhất là cặp vị trí không tương thích (i,j).

Xuất phát từ quan sát đó, chúng tôi đơn giản hóa chiến lược kiểm tra 4 giao tử bằng cách chỉ xem xét các cặp vị trí không tương thích có ít nhất một loại giao tử có tần số bằng 1. Giả định này đảm bảo rằng thuật toán sẽ luôn phá vỡ ít nhất một cặp vị trí không tương thích khi thực hiện một tái tổ hợp giữa một cặp vị trí không tương thích i và j.

Thuật toán GAMARG bắt đầu từ thời điểm $t = 1$. Tập các trình tự tại thời điểm t được kí hiệu là D_t (D_1=D). Với mỗi D_t, các danh sách cho các sự kiện kết hợp, đột biến và tái tổ hợp được xây dựng như sau:

- Danh sách kết hợp **C**: Đối với một cặp trình tự S_x và S_y có đoạn đầu chung $(S_x,S_y)\{d,l\}$, nếu $l = m$ thì $(S_x,S_y)\{d,l\}$ được thêm vào danh sách kết hợp.
- Danh sách đột biến **M**: Với một vị trí i ($1 \le i \le m$), nếu $S_x[i] = 1$ và $\forall S_y \in D_t \setminus \{S_x\}: S_y[i] \neq 1$ hoặc $S_x[i] = 0$ và $\forall S_y \in D_t \setminus \{S_x\}: S_y[i] \neq 0$, thì $S_x[i]$ được thêm vào danh sách đột biến.
- Danh sách giao tử **G**: Đối với một cặp vị trí không tương thích (i,j) ($0 \le i < m, j - i \le \delta$), nếu tồn tại một trình tự S_x chứa loại giao tử có tần số bằng 1 thì $S_x(i,j)$ được thêm vào danh sách giao tử.
- Danh sách đoạn đầu chung **S**: Với một cặp trình tự S_x và S_y có đoạn đầu chung $(S_x,S_y)\{d,l\}$, nếu $0 < l < m$ thì $(S_x,S_y)\{d,l\}$ được thêm vào danh sách đoạn đầu chung.

Khi một trong 3 sự kiện xảy ra, tập trình tự tiếp theo D_{t+1} được tạo ra từ tập trình tự D_t hiện thời và 4 danh sách ứng cử viên được cập nhật.

Thuật toán GAMARG

Đầu vào: Một tập N trình tự nhị phân độ dài m.

Đầu ra: Một đồ thị ARG chứa các sự kiện kết hợp, đột biến, tái tổ hợp giữa các trình tự, các trình tự trung gian được sinh ra và trình tự tổ tiên chung duy nhất được tìm thấy.

- **Bước 1:** Nếu danh sách kết hợp **C** không rỗng, thực hiện tất cả các kết hợp có thể.
- **Bước 2:** Nếu danh sách đột biến **M** không rỗng, thực hiện tất cả các đột biến có thể

sau đó chuyển sang Bước 1. Nếu không có đột biến nào, chuyển sang Bước 3.

- **Bước 3:** Nếu danh sách giao tử **G** không rỗng, thực hiện một tái tổ hợp sau đó chuyển sang Bước 1.
- **Bước 4:** Nếu danh sách đoạn đầu chung **S** không rỗng, thực hiện một tái tổ hợp theo sau là một sự kiện kết hợp. Chuyển đến Bước 1.
- **Bước 5:** Lặp lại Bước 1, Bước 2, Bước 3, và Bước 4 cho tới khi đạt đến một tổ tiên chung duy nhất.

Trong danh sách Giao tử **G**, nếu một trình tự ứng cử viên $S_x(i, j)$ có khoảng cách ngắn nhất từ vị trí i đến vị trí j, tức là, $(j - i)$ có giá trị nhỏ nhất thì S_x có thứ tự ưu tiên hàng đầu để thực hiện tái tổ hợp. Các ứng cử viên trong các bước kết hợp, đột biến và tái tổ hợp sẽ được lấy ngẫu nhiên khi chúng cùng đạt các tiêu chuẩn đặt ra.

3.5 Kết quả

Các thực nghiệm trên các bộ dữ liệu với các kích thước khác nhau cho thấy REARG có thể giúp tìm ra các ARG có số sự kiện tái tổ hợp ít hơn so với ARG4WG với những tập dữ liệu vừa và lớn. Tuy nhiên, cả thuật toán ARG4WG và REARG đều không phù hợp với các tập dữ liệu nhỏ. Thuật toán GAMARG tổng quát hơn khi có kết quả tốt nhất trong tất cả các thực nghiệm. GAMARG có khả năng xây dựng được những ARG có chính xác hoặc gần chính xác số sự kiện tái tổ hợp nhỏ nhất. Ngoài ra, các thực nghiệm cũng cho thấy thuật toán Margarita không ổn định khi chạy với các tập dữ liệu kích thước trung bình trích xuất từ dữ liệu 1kGP.

Các kết quả nghiên cứu của chương này đã được công bố trên một báo báo tại hội thảo quốc tế KSE năm 2017 (công trình khoa học số 2) và một báo cáo đã được chấp nhận tại hội thảo quốc tế ICBBB năm 2019 (công trình khoa học số 3).

Kết luận

Xác định nguồn gốc di truyền của bệnh bằng việc xác định các gen và alen nhạy cảm với bệnh là mục tiêu then chốt của nghiên cứu di truyền học con người. Đồ thị tái tổ hợp di truyền đóng một vai trò quan trọng trong nghiên cứu di truyền quần thể, đa dạng hệ gen và đa hình di truyền SNP. Tuy nhiên, bài toán xây dựng đồ thị ARG là một bài toán NP-khó và đòi hỏi tính toán khối lượng lớn nên ứng dụng vào thực tế còn hạn chế.

Thông qua việc nghiên cứu các phương pháp xây dựng đồ thị ARG, tập trung theo hướng tiếp cận xây dựng đồ thị ARG có ít số sự kiện tái tổ hợp nhất và thuật

toán Margarita, chúng tôi đã đề xuất thuật toán ARG4WG xây dựng đồ thị ARG hợp lý cho dữ liệu lớn hàng nghìn mẫu trên toàn hệ gen.

Bằng cách tiếp cận vấn đề theo cách của Margarita, cải tiến sử dụng đoạn đầu chung dài nhất cho bước tính toán sự kiện tái tổ hợp, thuật toán ARG4WG đề xuất đã cho ra các đồ thị ARG có ít sự kiện tái tổ hợp hơn Margarita. Đồng thời, chiến lược này không những giúp đảm bảo số nút trong đồ thị luôn được ổn định sau mỗi lần thực hiện bước tái tổ hợp mà còn làm giảm đáng kể thời gian tìm kiếm các đoạn chung dài nhất trong quá trình xây dựng đồ thị ARG. Kết quả thực nghiệm cho thấy thuật toán ARG4WG nhanh hơn hàng trăm đến hàng nghìn lần thuật toán Margarita. Đặc biệt, ARG4WG có thể chạy được với hàng nghìn mẫu trên toàn nhiễm sắc thể trong 1 lần chạy trong khoảng thời gian hợp lý thông qua xử lý đa luồng.

Chúng tôi cũng đã thực hiện một ứng dụng thuật toán đề xuất vào một bài toán thực tế xác định tương quan toàn bộ nhiễm sắc thể trên tập dữ liệu lớn. Cụ thể, chúng tôi đã thử nghiệm ứng dụng ARG4WG trong bài toán tìm vùng gen liên quan đến bệnh sốt rét ở Châu Phi trên 5560 trình tự độ dài toàn nhiễm sắc thể 11. Kết quả vùng tín hiệu bệnh sốt rét tìm được trùng với các kết quả phân tích đã có. Các kết quả này đã cho thấy khả năng ứng dụng của thuật toán ARG4WG vào các bài toán thực tế trên dữ liệu lớn.

Luận án cũng đã đề xuất 2 thuật toán cải tiến REARG và GAMARG nhằm tối ưu thêm số sự kiện tái tổ hợp trong quá trình xây dựng đồ thị ARG. Thuật toán REARG giúp quá trình xây dựng ARG khu trú được vào các ARG có số sự kiện tái tổ hợp nhỏ nhanh hơn ARG4WG trong hữu hạn số lần chạy thuật toán đối với các tập dữ liệu vừa và lớn. Tuy nhiên, GAMARG tổng quát hơn. GAMARG có khả năng xây dựng được những ARG có chính xác hoặc gần chính xác số sự kiện tái tổ hợp nhỏ nhất.

Trong thời gian tới, việc xác định tham số δ trong GAMARG cần được thực hiện một cách hệ thống hơn. Ý tưởng sử dụng các thuật toán trong bài toán tìm các khối haplotype (haplotype blocks) có thể được áp dụng. Bên cạnh đó, chúng tôi sẽ tiếp tục nghiên cứu và triển khai các ứng dụng thuật toán ARG4WG, GAMARG vào các bài toán thực tế khác như bài toán tìm đa hình di truyền đơn nucleotide, xử lý dữ liệu bị khuyết, …