ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

——————o0o——————

BÁO CÁO BÀI TẬP LỚN CUỐI KỲ

GENOTYPING AND HAPLOTYPING

Giảng viên: **TS. Lê Sỹ Vinh**

Lớp môn học: **2324II_INT3407_1 - Tin sinh học**

Sinh viên thực hiện: **Phạm Thanh Sơn** **21020027**

Nguyễn Anh Tú **21020030**

Nguyễn Thành Đạt **21020064**

HÀ NỘI, 2024

Mục lục

1 Phân công công việc

Thành viên	Mã sinh viên	Đóng góp
21020027	Phạm Thanh Sơn	33%
21020030	Nguyễn Anh Tú	33%
21020064	Nguyễn Thành Đạt	33%

2 Lời giới thiệu

Di truyền học là một nền tảng quan trọng của sinh học phân tử và là một công cụ thiết yếu trong việc giải mã sự phức tạp của các quá trình sinh học. Điều này đã dẫn đến những khoản đầu tư đáng kể và sự phát triển công nghệ trong lĩnh vực di truyền học cho các ứng dụng học thuật và công nghiệp. Sự tiến bộ của di truyền học đã dẫn đến những phát hiện nghiên cứu quan trọng và các ứng dụng lâm sàng, cải thiện sự hiểu biết của chúng ta về các bệnh và tính trạng. Điều này cũng hỗ trợ việc điều trị và phát triển thuốc, mở đường cho y học cá nhân hóa dựa trên di truyền học.

Báo cáo này trình bày một cách tổng quan về hai kỹ thuật phân tích genotyping và haplotyping, từ quá trình hình thành phát triển cho đến một số phương pháp phổ biến được sử dụng, từ đó, cho ta thấy được tính ứng dụng thực tiễn của genotyping và haplotyping không chỉ trong lĩnh vực sinh học y tế mà còn trong các ngành công nghiệp khác.

3 Genotyping

Kiểu gen của một cá nhân được xác định bởi sự kết hợp của các alen thừa hưởng từ cả hai cha mẹ. Kiểu gen có thể được trình bày dưới dạng carriage hoặc dosage. Carriage là sự tồn tại của một alen nhất định bất kể số lượng bản sao (0 hoặc 1), trong khi dosage chỉ rõ số lượng bản sao của alen (0, 1

hoặc 2). Genotyping là quá trình xác định trình tự DNA tại một vị trí cụ thể trong bộ gen, tức là xác định alen.

3.1 Quá trình phát triển của các kỹ thuật Genotyping

Trước đây, kiểu gen được suy luận dựa trên kiểu hình được biểu hiện của nó, điều này sau đó đã được phát triển thành xét nghiệm chẩn đoán Đánh mô (tissue typing), do sự xuất hiện của việc cấy ghép nội tạng đòi hỏi việc xác định các kháng nguyên HLA của cá nhân nhằm đảm bảo tính tương thích với vật chủ

Các phương pháp giải trình tự DNA được tiên phong sử dụng vào những năm 1970, cho phép phát triển các công cụ có thể xác định kiểu gen. Các biến thể trong trình tự DNA được nhận diện bởi các enzyme giới hạn của vi khuẩn gây ra sự phân cắt DNA tại các vị trí khác nhau, dẫn đến sự khác biệt về chiều dài các đoạn DNA. Ban đầu, các đa hình chiều dài đoạn giới hạn (Restriction Fragment Length Polymorphisms) này đã được sử dụng để định kiểu gen trong những năm 1970-1980.

Những kỹ thuật ban đầu này đòi hỏi nhiều công sức và thời gian, với các quy trình phức tạp thường kéo dài trong vài ngày, bao gồm điện di gel và hybridization với các đầu dò đánh dấu phóng xạ để hiển thị kết quả. Sau đó, với sự xuất hiện của phản ứng chuỗi polymerase (PCR) vào năm 1985 (Mullis và cộng sự, 1986) đã cho phép khuếch đại vô hạn các bản sao của một đoạn DNA cụ thể, cách mạng hóa lĩnh vực di truyền học và y học bằng cách cho phép so sánh DNA, chẩn đoán các rối loạn di truyền và phát hiện virus trong tế bào người.

Độ phân giải của định kiểu gen (Genotyping resolution) đã tăng lên đáng kể với sự phát triển của công nghệ vi mạch DNA (DNA microarray). Vi mạch được sử dụng để phân tích đồng thời một số lượng lớn các biến thể.

Các oligonucleotide với các trình tự DNA cụ thể, được gọi là đầu dò, gắn vào DNA mục tiêu để phát hiện các biến thể trình tự. Các vi mạch truyền thống thể rắn bao gồm các đầu dò được đặt cố định trên một chip, với mỗi vị trí đại diện cho một trình tự cụ thể. Gần đây hơn, các vi mạch hạt cho phép một số lượng cực lớn các đầu dò mà DNA được lai trên các hạt polystyrene mã hóa vi mô.

Giải trình tự toàn bộ bộ gen (WGS) lâm sàng được giới thiệu vào năm 2014. WGS xác định toàn bộ trình tự DNA của một bộ gen trong một thí nghiệm duy nhất, bao gồm cả DNA ty thể và, đối với thực vật, DNA lục lạp. Đây là phương pháp định kiểu gen có độ phân giải cao nhất. WGS xác định các biến thể nhân quả từ các nghiên cứu liên kết để giúp dự đoán khả năng mắc bệnh và phản ứng thuốc. Phiên bản đơn giản hóa và ít tốn kém hơn là giải trình tự toàn bộ exon (WES), trong đó chỉ 1% đến 2% bộ gen, đại diện cho các vùng mã hóa được biểu hiện thành protein, được giải trình tự. Việc giải trình tự toàn bộ bộ gen gần đây trở nên dễ tiếp cận hơn nhờ sự phát triển nhanh chóng và ứng dụng rộng rãi của công nghệ giải trình tự thế hệ mới (NGS), bao gồm cải tiến trong phân tích song song hàng loạt, thông lượng cao và giảm chi phí.

3.2 Các yếu tố ảnh hưởng đến việc lựa chọn kỹ thuật

Việc lựa chọn các kỹ thuật genotyping phụ thuộc vào các thông số như số lượng (tức là số lượng các dấu hiệu di truyền) và loại thông tin di truyền cần thiết (tức là độ chính xác), kích thước của nhóm đối tượng nghiên cứu (tức là số lượng cá nhân), khả năng tính toán sẵn có (ví dụ: tiền xử lý, hậu xử lý) và các hạn chế về tài chính.

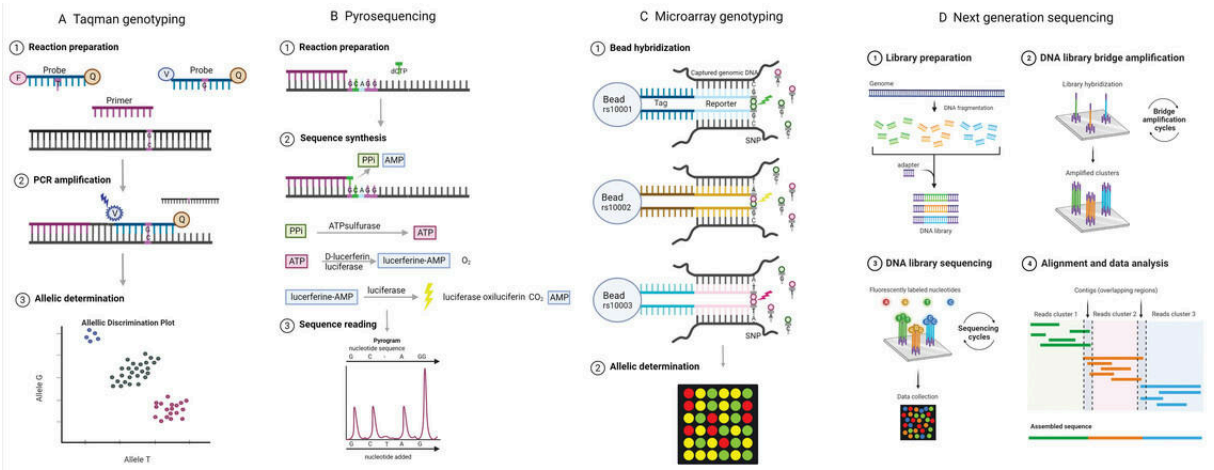
Thông tin di truyền cần thiết cho các nghiên cứu khám phá quy mô lớn, trong đó nhiều biến thể di truyền được khảo sát trong các nhóm đối tượng

lớn, khác biệt đáng kể so với nghiên cứu có mục tiêu về một gen cụ thể hoặc một tập hợp các locus. Trong chẩn đoán bệnh xơ nang, điều quan trọng là xác định các biến thể gây bệnh trong gen CFTR, đây là một ví dụ về điều tra có mục tiêu. Một ví dụ về nghiên cứu quy mô lớn để khám phá các biến thể ảnh hưởng đến nguy cơ mắc các bệnh phổ biến là việc xác định hơn 200 SNP nguy cơ đối với bệnh đa xơ cứng (International Multiple Sclerosis Genetics, 2019). Sự gia tăng các nghiên cứu sàng lọc toàn bộ bộ gen, trong đó toàn bộ bộ gen được định kiểu gen hoặc giải trình tự, cho các bệnh phổ biến trong những thập kỷ gần đây đã thúc đẩy việc sử dụng các vi mạch định kiểu gen, trong đó hàng trăm nghìn SNP được đánh giá cùng lúc. Kỹ thuật này tạo ra các kiểu gen cho các SNP đại diện cho một phần cụ thể của bộ gen. Mặc dù trình tự chưa được biết rõ, vùng di truyền được đại diện bởi SNP có thể được kiểm tra để liên kết với các bệnh. Mức độ thông tin này thường đại diện cho sự cân bằng tốt giữa độ bao phủ và chi tiết. Dữ liệu từ vi mạch cần được xử lý tính toán và đảm bảo chất lượng bổ sung; tuy nhiên, các phương pháp và công cụ tiêu chuẩn có sẵn để thực hiện các phân tích như vậy. Chi phí tính toán trên mỗi kiểu gen là thấp, nhưng tổng chi phí cao hơn hầu hết các xét nghiệm có mục tiêu do số lượng biến thể nhiều.

Trong khi đó, phương pháp PCR đặc hiệu alen hoặc PCR TaqMan phù với các nghiên cứu có mục tiêu và độ phân giải thấp, chẳng hạn như phân tầng di truyền, trong khi pyrosequencing hoặc giải trình tự thế hệ mới (NGS) phù hợp hơn nếu cần xác định đặc điểm di truyền của các locus cho từng cá nhân. Phân tầng di truyền so sánh các cá thể mang và không mang một biến thể hoặc haplotype cụ thể (ví dụ: biến thể HBB gây bệnh thiếu máu hồng cầu hình liềm). Nguy cơ di truyền của ung thư vú liên quan đến một số biến thể trong các gen BRCA1 và BRCA2, do đó pyrosequencing và

NGS là các phương pháp lựa chọn để điều tra nguy cơ cá nhân. Giải trình tự (cụ thể theo vị trí, exon, hoặc toàn bộ bộ gen) tạo ra thông tin di truyền chi tiết nhất, nhưng cũng đắt đỏ hơn và đòi hỏi tính toán cao, khiến các nhà nghiên cứu mới không dễ dàng áp dụng. Mức độ chi tiết này là không cần thiết cho nhiều nghiên cứu. Pyrosequencing có lợi thế trong việc định kiểu gen các locus có tính đa hình cao và thường được sử dụng trong các nghiên cứu di truyền biểu sinh, nhưng công nghệ này khá tốn công sức và thời gian. Pyrosequencing và các công nghệ định kiểu gen khác liên quan đến xử lý mẫu thủ công và kiểm tra kết quả, bao gồm PCR đặc hiệu alen và TaqMan, hoạt động tốt cho các nhóm đối tượng nhỏ hơn nhưng có sử dụng hạn chế khi định kiểu gen các nhóm đối tượng lớn hơn.

3.3 Một số phương pháp Genotyping



3.3.1 Phân lập DNA (DNA preparation)

Phân lập DNA bộ gen liên quan đến việc tách DNA nhiễm sắc thể từ nhân tế bào bằng chất tẩy và cắt cơ học, sau đó loại bỏ protein và mảnh vụn tế bào để thu được mẫu DNA tinh khiết. Có nhiều phương pháp có thể được sử dụng để tách DNA từ các mẫu sinh học, bao gồm máu, nước bọt hoặc mô sinh thiết được nhúng trong parafin. Phương pháp tách chiết hữu cơ DNA bộ gen liên quan đến việc tách DNA và protein vào các pha hữu

cơ khác nhau bằng cách thêm phenol/chloroform. Phương pháp này có chi phí thấp và yêu cầu rất ít thiết bị. Các bộ kit thương mại cung cấp một phương pháp nhanh hơn và dễ dàng hơn bằng cách sử dụng cột lọc để tách DNA. Các bộ kit này đắt hơn và khó khắc phục sự cố trong trường hợp có vấn đề vì chi tiết về thành phần đệm là bí mật. Hạt từ tính cũng có thể được sử dụng để tách và tinh chế DNA trong một bước, với các bộ kit thương mại có sẵn từ nhiều công ty.

3.3.2 Lai hoá (Hybridization)

Quá trình lai hóa giữa DNA bộ gen và đầu dò tương ứng là nguyên tắc cốt lõi đằng sau nhiều kỹ thuật định kiểu gen, bao gồm PCR, vi mạch (microarray), và giải trình tự thế hệ mới (NGS). Lai hóa là quá trình kết hợp hai đoạn DNA đơn sợi bổ sung để tạo thành một phân tử sợi kép, phụ thuộc vào sự khớp cặp bazơ giữa các đoạn. Sự liên kết không cộng hóa trị chặt chẽ giữa các sợi được đạt được khi nhiều cặp bazơ bổ sung, và các trình tự không đặc hiệu sẽ bị rửa trôi trước khi tiến hành các bước tiếp theo.

3.3.3 Các phương pháp PCR

Việc sử dụng PCR để khuếch đại một đoạn DNA nhất định đã dẫn đến tự động hóa nhanh chóng các phương pháp định kiểu gen. PCR là một phương pháp cell-free, nhanh chóng và tinh tế để tạo ra nhiều bản sao của một trình tự DNA, nhưng nó phụ thuộc vào việc biết trình tự xung quanh đoạn DNA quan tâm.

Phản ứng PCR có ba bước, được lặp lại trong nhiều chu kỳ. Bước đầu tiên là biến tính, khi nhiệt độ được tăng lên để các sợi DNA tan chảy và trở thành đơn sợi. Bước thứ hai là gắn mồi, khi nhiệt độ được giảm xuống để cho phép các đoạn DNA đơn sợi ngắn, được gọi là mồi, bổ sung với trình tự quan tâm, gắn vào. Bước thứ ba là tổng hợp DNA, khi DNA polymerase

chịu nhiệt mở rộng trình tự DNA từ mồi dọc theo DNA mục tiêu. Các bước này được lặp lại nhiều lần, tạo ra nhiều bản sao của vùng quan tâm. Nếu các mồi bao gồm biến thể quan tâm, có thể thiết kế các điều kiện khuếch đại để chỉ khuếch đại alen quan tâm và không phải các alen khác trong phương pháp PCR đặc hiệu alen.

Một cách khác để sử dụng PCR cho định kiểu gen là PCR-RFLP, trong đó vùng quan tâm được khuếch đại bằng PCR, sau đó được cắt bằng một enzyme hạn chế chọn lọc để nhận biết một trình tự DNA chỉ có trong một trong các alen; do đó, kích thước của các sản phẩm kết quả sẽ phân biệt các alen khác nhau. PCR có thể dễ dàng phát hiện các vi vệ tinh hoặc các đa hình lặp lại ngắn liên tiếp (STRP) vì chiều dài của đoạn khuếch đại sẽ thay đổi tùy thuộc vào số lần vi vệ tinh được lặp lại. Mặc dù PCR đặc hiệu alen và PCR-RFLP đã được thiết lập nhiều năm trước, chúng vẫn được sử dụng trong một số bối cảnh nhất định.

Bên cạnh đó, TaqMan-PCR là một phương pháp định kiểu gen được sử dụng rộng rãi để xác định kiểu gen của các SNP (đa hình nucleotide đơn) tiềm năng. Một vùng từ 100-150 bp xung quanh SNP quan tâm được khuếch đại bằng PCR trong sự hiện diện của hai đầu dò đặc hiệu alen, một cho mỗi alen thay thế. Trình tự đầu dò chứa SNP. Các đầu dò này có các nhãn huỳnh quang khác nhau ở đầu 5′ và một chất làm tắt huỳnh quang gắn vào đầu 3′ để không có tín hiệu từ đầu dò khi còn ở trong dung dịch. Đầu dò gắn vào DNA chứa alen bổ sung. Khi DNA polymerase mở DNA trong quá trình kéo dài, nhãn huỳnh quang được giải phóng vào dung dịch, tách rời khỏi chất làm tắt huỳnh quang, dẫn đến sự phát hiện huỳnh quang có thể quan sát được. Vì mỗi alen được đánh giá với đầu dò và nhãn huỳnh quang riêng, có thể phát hiện đồng thời hai alen thay thế với tiềm năng multiplex lên đến 100 locus di truyền khác nhau trên một mảng đơn. Vì

TaqMan PCR là một phương pháp định lượng, nó có thể được sử dụng để đánh giá sự biến đổi số lượng bản sao bằng cách đo số lượng bản sao được tạo ra của mục tiêu trong quá trình PCR và so sánh với một mẫu tham chiếu có số lượng bản sao đã biết.

3.3.4 Các phương pháp Microarray

Một trong những phương pháp phổ biến nhất để định kiểu gen là vi mạch (microarrays). Vi mạch là một bề mặt rắn trên đó các đốm DNA tổng hợp siêu nhỏ được áp dụng. Chúng có thể được sử dụng để định kiểu gen bằng cách thiết kế các đốm DNA chứa các đoạn DNA trùng lặp với các SNP mục tiêu. Các điều kiện lai hóa thích hợp đảm bảo rằng DNA mục tiêu chỉ lai hóa nếu nó bổ sung với DNA ở một đốm cụ thể.

Có nhiều loại vi mạch định kiểu gen, phổ biến nhất là vi mạch GeneChip của Affymetrix và các vi mạch hạt của Illumina. Công nghệ vi mạch hạt do Illumina giới thiệu là một phương pháp định kiểu gen dựa trên vi mạch thay thế. Các thư viện hạt được lắp ráp ngẫu nhiên vào một giá đỡ vi lỗ khắc. Mỗi hạt chứa nhiều bản sao của các oligonucleotide nhắm vào một locus cụ thể cũng như một oligonucleotide ánh xạ được thiết kế không có tính đồng nhất với bất kỳ trình tự nào từ loài đang được nghiên cứu. Điều này được sử dụng để ánh xạ vị trí mà một hạt cụ thể đã kết thúc bằng một loạt các bước lai hóa được thực hiện trong quá trình thiết kế vi mạch. Trên mỗi vi mạch, có thể kiểm tra từ vài trăm nghìn đến hơn một triệu kiểu gen cho một cá nhân. Trong bước định kiểu gen, DNA được định kiểu sẽ gắn vào các hạt với DNA bổ sung, dừng lại một cặp bazơ trước trình tự quan tâm, vị trí của SNP. Một bước mở rộng đơn cặp bazơ kết hợp một trong bốn nucleotide được đánh dấu diễn ra. Nucleotide nào được kết hợp sẽ được phát hiện dưới dạng một tín hiệu được phát ra khi bị kích thích

bởi tia laser. Cường độ tín hiệu cung cấp thông tin về tỷ lệ alen tại một locus cụ thể.

3.3.5 Next-generation

NGS, còn được gọi là giải trình tự thế hệ thứ hai, đang nhanh chóng thay thế các phương pháp cũ vì nó cho phép giải trình tự song song hàng triệu đoạn DNA. Đầu tiên, DNA (hoặc RNA) được phân tách thành các đoạn ngắn ngẫu nhiên. Trong quá trình chuẩn bị mẫu, các adaptor, các mô típ dùng để nhận diện mẫu (chỉ số), các cầu nối giải trình tự và các trình tự bổ sung với các oligos của ô dòng chảy (flow cell) được thêm vào các đoạn DNA. Mỗi đoạn DNA được khuếch đại đẳng nhiệt trên ô dòng chảy, một giá thể thủy tinh chứa hàng triệu vi lỗ tại các vị trí cố định. Mỗi lỗ chứa các đầu dò DNA được sử dụng để bắt giữ các sợi DNA trong quá trình lai hóa nhằm cho phép khuếch đại trong quá trình tạo cụm. Ô dòng chảy chứa hai oligos gắn vào các adaptor ở mỗi đầu của DNA mục tiêu. Khi DNA mẫu đi qua ô dòng chảy, nó lai hóa với một trong các oligos và một polymerase tổng hợp bản sao của DNA mục tiêu.

Tiếp theo, khuếch đại cầu (bridge amplification) được sử dụng để khuếch đại dòng DNA theo cách clonal. DNA mục tiêu đầu tiên được rửa đi, bản sao mới được tổng hợp uốn cong và đầu adaptor thứ hai gắn vào loại oligo thứ hai trên ô dòng chảy. Các bản sao mới của DNA mục tiêu được tạo ra với mỗi vòng khuếch đại. Các sợi ngược được rửa đi, chỉ để lại các sợi xuôi. Phát hiện trình tự sử dụng công nghệ giải trình tự bằng tổng hợp (SBS), theo dõi việc thêm các nucleotide được đánh dấu huỳnh quang, phát ra tín hiệu độc đáo cho mỗi nucleotide khi chúng được gắn vào. Sản phẩm đọc mới được tổng hợp được rửa đi và một mồi chỉ số được thêm vào để giải trình tự chỉ số cạnh một trong các adaptor. Cùng nhau, hai trình tự này

tạo thành lần đọc 1. Quá trình giải trình tự được lặp lại cho sợi ngược theo cùng cách thức, tạo ra lần đọc thứ hai ở đầu kia.

Quá trình này được thực hiện song song cho hàng triệu đoạn DNA, tạo ra một lượng lớn dữ liệu trình tự. Các trình tự chỉ số có thể được sử dụng để phân tách các trình tự DNA từ các mẫu khác nhau. Tiếp theo đó là việc lập bản đồ các đoạn DNA vào bộ gen tham chiếu, hoặc lắp ráp de novo nếu đang giải trình tự một loài mới hoặc mẫu từ một loài chưa biết. Đối với mục đích định kiểu gen, các khác biệt trong trình tự DNA giữa các mẫu tại cùng vị trí được xác định. Các phương pháp NGS có thể được áp dụng cho các vùng mục tiêu, toàn bộ exon, hoặc toàn bộ bộ gen. Chúng cũng có thể định lượng biểu hiện gen nếu mẫu là RNA thay vì DNA. Một lợi thế của phương pháp định kiểu gen này là nó cho phép xác định các đa hình/đột biến mới, điều không thể thực hiện được với các phương pháp genotyping khác.

3.4 Ứng dụng thực tiễn

3.4.1 Liên kết bệnh và đặc điểm

Bằng cách so sánh các biến thể gen giữa các cá thể của một loài, các nhà nghiên cứu có thể xác định các dấu hiệu di truyền có thể di truyền, hoặc các đánh dấu, liên quan đến các đặc điểm cụ thể. Những khác biệt độc đáo này có thể được sử dụng như các đánh dấu trong các nghiên cứu liên kết và liên kết.

Các nghiên cứu liên kết toàn bộ gen, hoặc GWAS, so sánh các khác biệt gen trên toàn bộ gen của hai cá thể hoặc dân số. Ví dụ, các gen của một nhóm người mắc bệnh có thể được so sánh với các trình tự gen từ một nhóm người tương tự mà không mắc bệnh. Bất kỳ SNP hoặc haplotype nào phổ biến hơn ở những người mắc bệnh được gọi là một đánh dấu gen

liên quan. Mặc dù các nghiên cứu như vậy thường đo lường từ 100,000 đến 2,000,000 biến thể, điều này chỉ mô tả một phần nhỏ của toàn bộ gen. Tuy nhiên, do Trạng thái mất cân bằng liên kết (LD), một tập hợp lựa chọn các điểm đánh dấu có thể được sử dụng để xác định các hiệu ứng di truyền ngay cả khi biến thể gây ra không được xác định. LD cũng làm khó khăn trong việc xác định các biến thể gây ra, vì các SNP trong LD thường được liên kết thành các khối. Do đó, các mối quan hệ với các biến thể gen thường không chỉ ra tính chất gây ra của chúng.

3.4.2 Di truyền học y khoa

Di truyền học đã đóng vai trò quan trọng trong chẩn đoán lâm sàng kể từ nửa sau của thế kỷ 20, bắt đầu với việc sử dụng xét nghiệm mô HLA để đánh giá sự tương thích mô trong quá trình cấy ghép cơ quan. Tuy nhiên, với các công nghệ genotyping dễ tiếp cận hơn, di truyền ứng dụng đã trở nên nổi bật hơn trong các cơ sở chăm sóc sức khỏe. Các gen hoặc đa hình cụ thể có thể được đánh giá trong di truyền học lâm sàng để chẩn đoán các bệnh di truyền đã biết rõ, như bệnh Huntington hoặc thiếu máu hồng cầu hình liềm. Đối với các trường hợp khám phá, có thể thực hiện giải trình tự toàn bộ exome và genome theo trio (trio ám chỉ một cá nhân và cha mẹ sinh học của họ) để xác định nguyên nhân gốc rễ của các rối loạn di truyền nghi ngờ, chẳng hạn như đột biến điểm. Việc xác định đột biến gen gây bệnh, đặc biệt là protein không hoạt động và cơ chế sinh học liên quan, tạo điều kiện thuận lợi cho quản lý bệnh và đánh giá các phương án điều trị.
Gần đây, các tổ chức thương mại như 23andMe và selfDecode đã thúc đẩy việc sử dụng các bộ xét nghiệm di truyền để cung cấp tự đánh giá về nguy cơ di truyền đối với bệnh tật. Những bộ xét nghiệm này thường bao gồm các gen có nguy cơ ung thư đã được xác định rõ như BRCA1/BRCA2 và nhiều bệnh di truyền như đa xơ cứng, xơ nang, và thiếu máu hồng cầu

hình liềm. Tuy nhiên, các vấn đề đạo đức đã được nêu ra do nhu cầu phải truyền đạt kết quả một cách đúng đắn, đặc biệt với các biện pháp ít được xác lập hơn liên quan đến các đặc điểm và bệnh phổ biến (ví dụ: cân nặng, tiểu đường)

3.4.3 Phả hệ và khoa học pháp y

Di truyền học cũng rất quan trọng trong việc xác định và thiết lập quan hệ họ hàng, điều này đã trở nên đặc biệt phổ biến với sự xuất hiện của các bộ kit xét nghiệm di truyền thương mại. Nó cũng trở thành một công cụ quan trọng trong lĩnh vực tư pháp hình sự và khoa học pháp y. Tính đặc hiệu của genotyping cung cấp một "dấu vân tay" độc nhất vô nhị để liên kết cá nhân với bằng chứng sinh học. Tuy nhiên, xét nghiệm di truyền không phải lúc nào cũng không thể bác bỏ, vì việc xử lý không đúng cách và ô nhiễm mẫu có thể ảnh hưởng đến độ tin cậy của bằng chứng DNA. Ô nhiễm cũng có thể gây ra vấn đề trong các môi trường y tế hoặc nghiên cứu, thường được phát hiện qua tỷ lệ thất bại genotyp cao hơn và tỷ lệ dị hợp tử. Chất lượng của cả việc xử lý mẫu và xét nghiệm có thể được sử dụng để biện minh cho việc bác bỏ bằng chứng DNA hoặc là căn cứ để kháng cáo. Hơn nữa, DNA có thể vô tình hoặc do ý đồ xấu mà buộc tội sai những người vô tội.

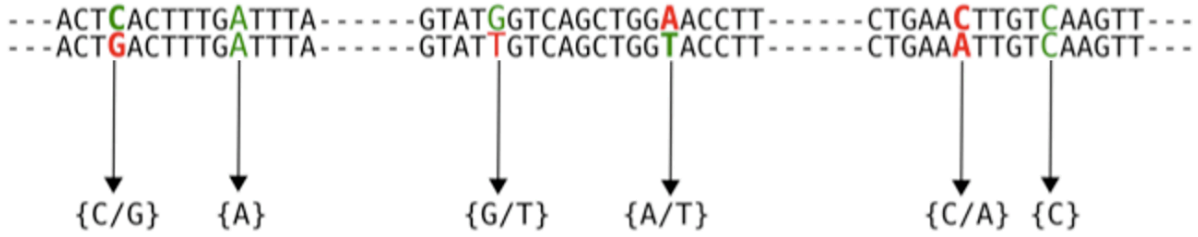
3.4.4 Phát triển nông nghiệp

Ứng dụng của genotyping trong nông nghiệp đã mang lại nhiều tiến bộ vượt bậc trong việc cải thiện chất lượng và năng suất cây trồng cũng như vật nuôi. Bằng cách xác định các biến thể di truyền liên quan đến các đặc tính mong muốn như khả năng chống chịu sâu bệnh, hạn hán, năng suất cao, và chất lượng dinh dưỡng, các nhà khoa học và nông dân có thể chọn lọc và lai tạo những giống cây trồng và vật nuôi có ưu thế di truyền. Genotyping giúp tối ưu hóa quá trình chọn lọc giống, giảm thiểu thời gian

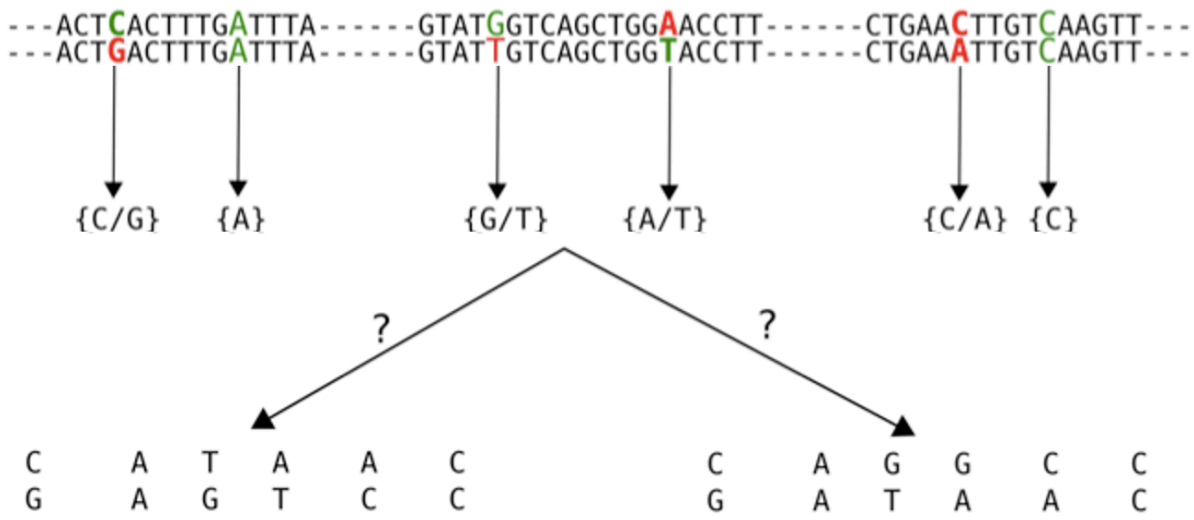
và chi phí so với các phương pháp truyền thống. Ngoài ra, công nghệ này còn hỗ trợ trong việc theo dõi và quản lý sự đa dạng di truyền trong quần thể, giúp bảo tồn các nguồn gen quý hiếm và duy trì sự bền vững của hệ sinh thái nông nghiệp. Ví dụ, việc sử dụng genotyping để xác định các gen liên quan đến khả năng chịu mặn đã giúp tạo ra các giống lúa mới có thể trồng ở vùng đất mặn, góp phần tăng cường an ninh lương thực trong bối cảnh biến đổi khí hậu.

4 Haplotyping

Con người thuộc bộ lưỡng bộ có hai nhiễm sắc thể tương đồng ở mỗi cặp nhiễm sắc thể. Trong đó một chiếc được di truyền từ bố, còn lại được di truyền từ mẹ.

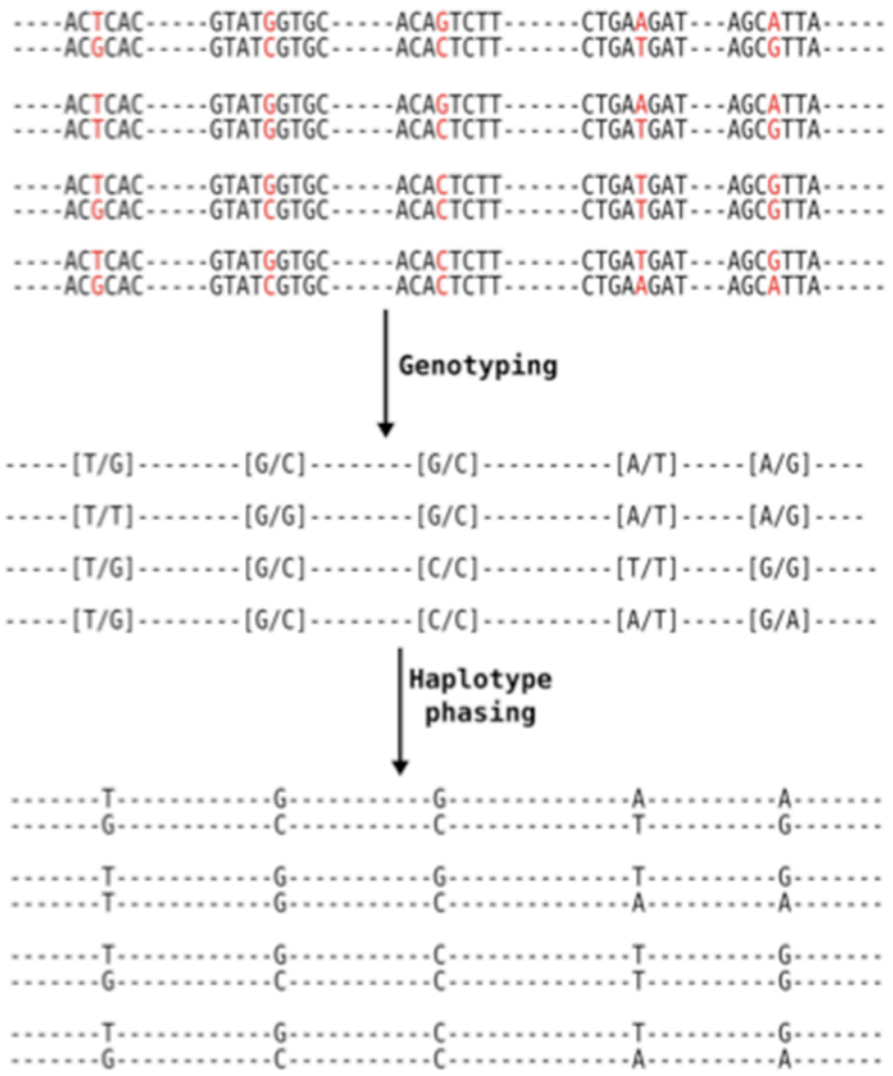


Tuy nhiên công nghệ genotyping lại không maintain phase này:



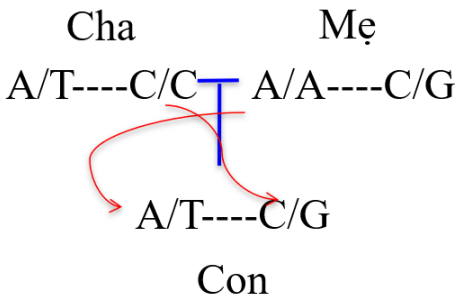
Nên ta chưa thể biết được allen đến từ bố hay mẹ, từ đó dẫn đến sự ra đời của haplotyping.

Haplotyping (còn gọi là haplotype phasing) là quá trình quyết định haploid DNA sequence (halotypes) từ unodered (unphased) genotype data hay đơn giản hơn **chính là quá trình xác định được allen đến từ bố hay mẹ.**



Hình 4: Từ SNP đến haplotype

Vậy haplotype phasing thực sự được diễn ra như thế nào? Hãy xem ví dụ đơn giản sau:



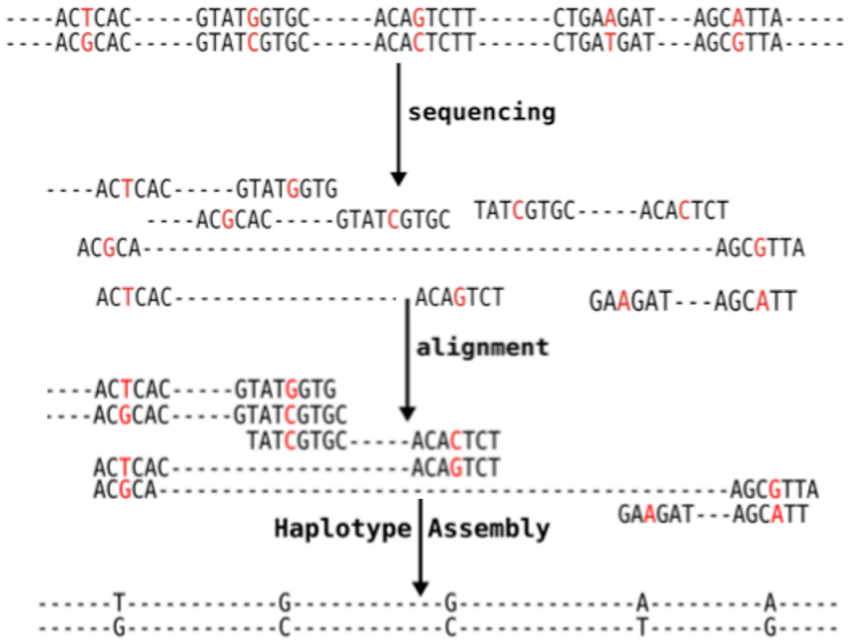
Nếu con là dị hợp còn cha mẹ là đồng hợp thì ta sẽ biết được alen nào đến từ cha hay mẹ

- Vị trí 1: A từ mẹ, T từ cha (Vì chỉ có cha mới có T)
- Vị trí 2: C từ cha, G từ mẹ (Vì chỉ có mẹ mới có G)

→ Haplotype của con sẽ là (TC, AG)

4.1 Haplotype assembly

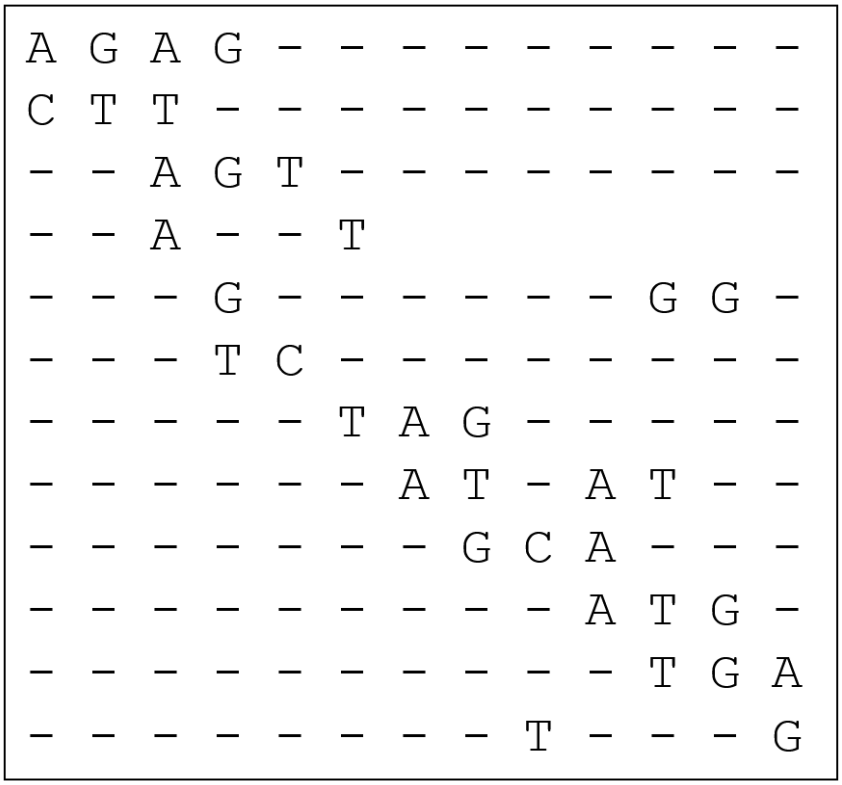
Haplotype assembly là quá trình tính toán để merge các đoạn DNA ngắn hơn thành một đoạn haplotype dài duy nhất



Hình 6: DNA sequencing → alignment → haplotype assembly

Quá trình nối này dựa vào những đoạn overlap với nhau có chứa một hoặc nhiều nucleotide.

4.2 Xem xét bài toán

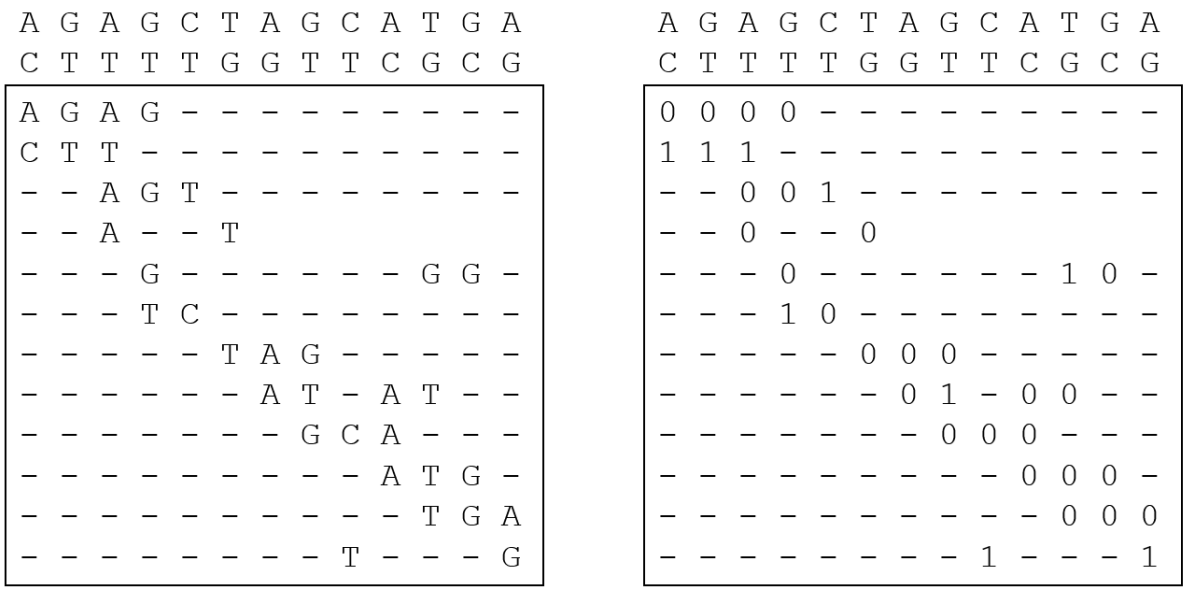


A	G	A	G	–	–	–	–	–	–	–	–	–
C	T	T	–	–	–	–	–	–	–	–	–	–
–	–	A	G	T	–	–	–	–	–	–	–	–
–	–	A	–	–	T							
–	–	–	G	–	–	–	–	–	–	G	G	–
–	–	–	T	C	–	–	–	–	–	–	–	–
–	–	–	–	–	T	A	G	–	–	–	–	–
–	–	–	–	–	–	A	T	–	A	T	–	–
–	–	–	–	–	–	–	G	C	A	–	–	–
–	–	–	–	–	–	–	–	–	A	T	G	–
–	–	–	–	–	–	–	–	–	–	T	G	A
–	–	–	–	–	–	–	–	T	–	–	–	G

A	G	A	G	C	T	A	G	C	A	T	G	A
C	T	T	T	T	G	G	T	T	C	G	C	G

Cho 1 dãy nucleotide `X` của cha và `Y` của mẹ và một matrix mà ở đó:

- Các đoạn được align theo unphased genotyping
- Loại bỏ các đoạn và các cột rỗng
- Đã xóa các cột tri-allelic SNP
- Gắn lại mác cho 2 alen bằng 0 hoặc 1:

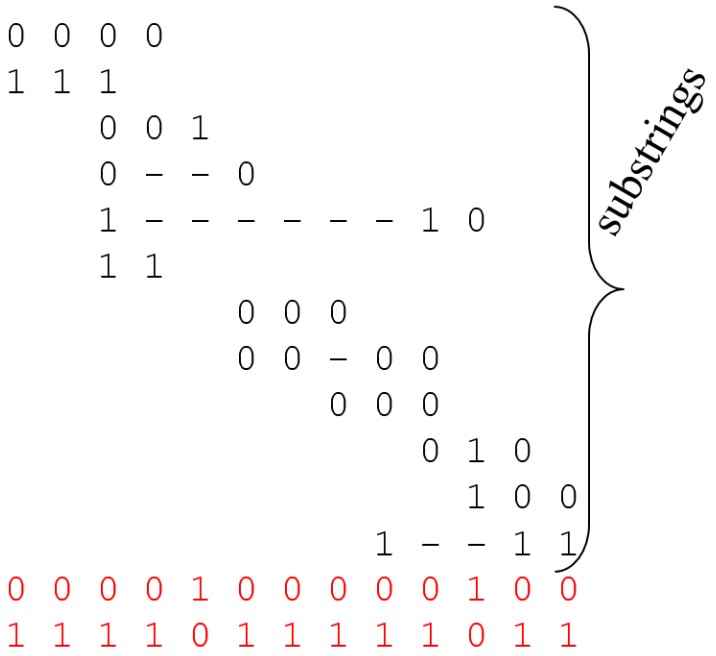


Hình 8: Haplotype assembly relabel

Nếu nucleotide trùng với cha thì ta gắn mác 0 còn lại sẽ gắn 1, hay formerly:

```
if a[i][j] == X[j]: a[i][[j] = 0
else if a[i][j] == Y[j]: a[i][j] = 1
```

Mục đích của chúng ta là construct lại phased binary string, biết các substrings (Các substring sẽ bắt đầu từ index nhỏ nhất khác rỗng đến index lớn nhất khác rỗng trên cùng một hàng:



Gọi unordered pair haplotype là (H1, H2), không mất tổng quát coi H1 là dãy của cha, H2 là của mẹ. Một substrings được coi là hợp lệ nếu mỗi substring match một trong hai H1 hoặc H2. Ký hiệu `-` được coi như wildcard matching:

E.g.

`- - 1 - - - - - - 1 0 - -` matches `1 1 1 1 0 1 1 1 1 1 0 1 1` (H2)

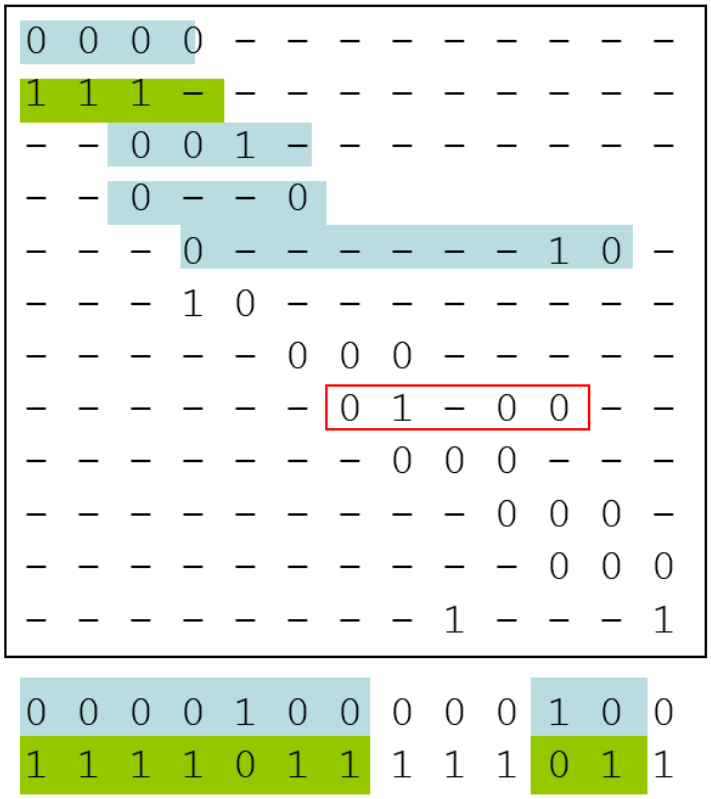
`- - - - - - - - - - - 1 0 0` matches `0 0 0 0 1 0 0 0 0 0 1 0 0` (H1)

Bài toán này trở nên khó hơn nhiều nếu các substrings có errors, điều này sẽ được em đề cập đến chi tiết hơn ở chương sau.

4.3 Các thuật toán tìm haplotype

4.3.1 Greedy heuristic

4.3.1.1 Simple approach



Hình 10: Extending greedy

Vì haplotype là unique nên nếu ta có các substrings error free thì có thể đơn giản là duyệt qua các đoạn rồi extend mình H1 rồi flip các bit lại sẽ có H2

```
#include <bits/stdc++.h>
using namespace std;
using ll = long long;

int32_t main() {
  cin.tie(0)->sync_with_stdio(0);
  int n;
  cin >> n;
  vector a(n, vector(n, -1));
  for (int i = 0; i < n; i++) {
    for (int j = 0; j < n; j++) {
      char x;
      cin >> x;
      if (x != '-') a[i][j] = x - '0';
    }
  }
  vector<int> h(n, -1);
  for (int i = 0; i < n; i++) {
    bool ok = 1;
    for (int j = 0; j < n; j++) {
      if (a[i][j] != -1) {
        if (h[j] != -1 && a[i][j] != h[j]) {
          ok = 0;
          break;
        }
      }
    }
```

```
    if (ok) {
      for (int j = 0; j < n; j++) {
        if (a[i][j] != -1) {
          h[j] = a[i][j];
        }
      }
    }
  }
  for (int j = 0; j < n; j++) {
    cout << h[j] << " \n"[j == n - 1];
  }
  for (int j = 0; j < n; j++) {
    cout << 1 - h[j] << " \n"[j == n - 1];
  }
}
```

Input

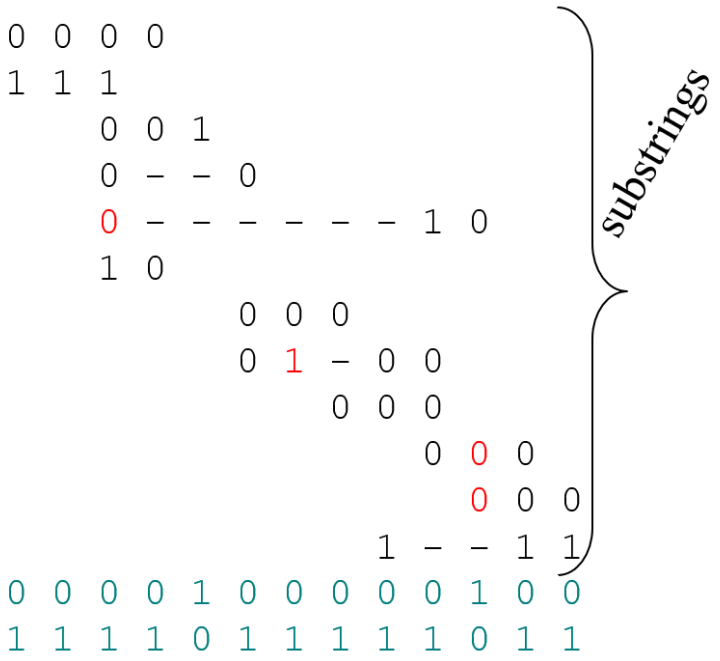
```
13
0 0 0 0 - - - - - - - - - -
1 1 1 - - - - - - - - - - -
- - 0 0 1 - - - - - - - - -
- - 0 - - 0 - - - - - - - -
- - 1 - - - - - - - 1 0 - -
- - 1 1 - - - - - - - - - -
- - - - - 0 0 0 - - - - -
- - - - - 0 0 - 0 0 - - -
- - - - - - - 0 0 0 - - -
- - - - - - - - - - 0 1 0 -
- - - - - - - - - - - 1 0 0
- - - - - - - - - 1 - - 1 1
```

Output

```
0 0 0 0 1 0 0 0 0 0 1 0 0
1 1 1 1 0 1 1 1 1 1 0 1 1
```

4.3.1.2 Substrings error

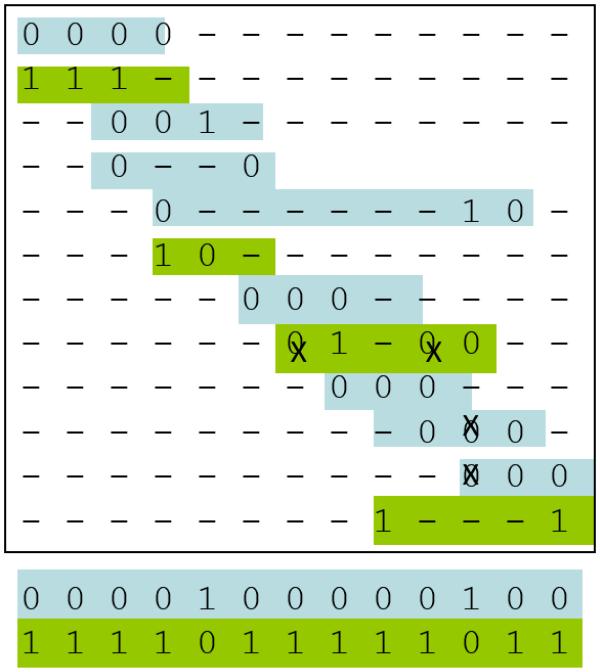
Tuy nhiên trong thực tế thì greedy dễ dẫn đến suboptimal solutions vì các substrings có thể có error rate



Hình 11: Ví dụ với 4 substrings có error

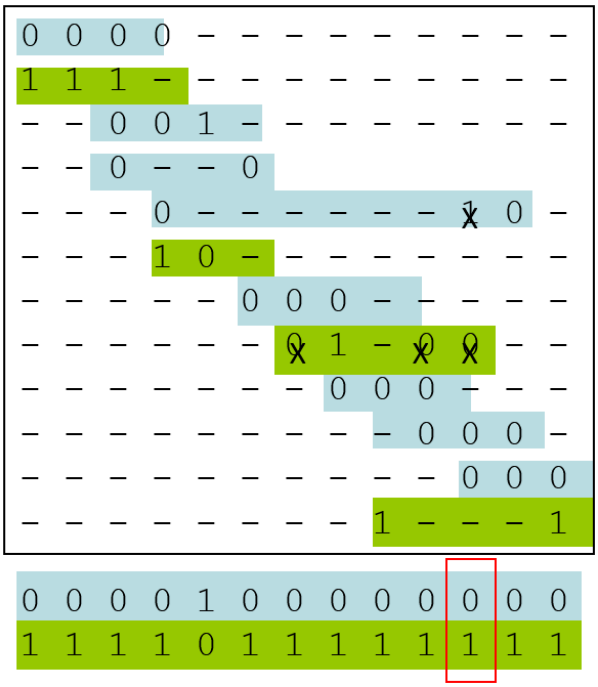
Để "giải quyết" vấn đề này người ta đã define `MEC` (Minimum error correction) là số lần correct lại substrings ít nhất để match tất cả các patterns với haplotype pair

Mục đính mới tất nhiên sẽ là xây dựng lại haplotype mà giảm thiểu `MEC`



Hình 12: MEC corrections

Local flipping haplotype thường giúp ta cải thiện được MEC



Hình 13: Flipping bits

Em sẽ define hàm MEC là tổng số error của min pattern match với H1 hoặc H2. Sau đó đơn giản là duyệt qua các index rồi check xem nếu flip bit sẽ làm giảm `MEC`

```cpp
#include <bits/stdc++.h>
using namespace std;
using ll = long long;

int MEC(vector<vector<int>> a, vector<int> h) {
  int n = a.size();
  int errors = 0;
  for (int i = 0; i < n; i++) {
    int cnt1 = 0, cnt2 = 0;
    for (int j = 0; j < n; j++) {
      if (a[i][j] != h[j]) cnt1++;
      if (a[i][j] != 1 - h[j]) cnt2++;
    }
    errors += min(cnt1, cnt2);
  }
  return errors;
}

int32_t main() {
  cin.tie(0)->sync_with_stdio(0);
  int n;
  cin >> n;
  vector a(n, vector(n, -1));
  for (int i = 0; i < n; i++) {
    for (int j = 0; j < n; j++) {
      char x;
      cin >> x;
      if (x != '-') a[i][j] = x - '0';
    }
  }
  vector<int> h(n, -1);
```

```
  for (int i = 0; i < n; i++) {
    bool ok = 1;
    for (int j = 0; j < n; j++) {
      if (a[i][j] != -1) {
        if (h[j] != -1 && a[i][j] != h[j]) {
          ok = 0;
          break;
        }
      }
    }
    if (ok) {
      for (int j = 0; j < n; j++) {
        if (a[i][j] != -1) {
          h[j] = a[i][j];
        }
      }
    }
  }
  int pre_mec = MEC(a, h);
  for (int i = 0; i < n; i++) {
    vector<int> nh(h);
    nh[i] = 1 - nh[i];
    int cur_mec = MEC(a, nh);
    if (cur_mec < pre_mec) {
      h[i] = 1 - h[i];
      pre_mec = cur_mec;
    }
  }
  for (int j = 0; j < n; j++) if (h[j] == -1) h[j] = 0;
  for (int j = 0; j < n; j++) {
    cout << h[j] << " \n"[j == n - 1];
```

```
  }
  for (int j = 0; j < n; j++) {
    cout << 1 - h[j] << " \n"[j == n - 1];
  }
}
```

Input

```
13
0 0 0 0 - - - - - - - - - -
1 1 1 - - - - - - - - - - -
- - 0 0 1 - - - - - - - - -
- - 0 - - 0 - - - - - - - -
- - - 0 - - - - - - 1 0 -
- - - 1 0 - - - - - - - - -
- - - - - 0 0 0 - - - - - -
- - - - - - - 0 1 - 0 0 - -
- - - - - - - - 0 0 0 - - -
- - - - - - - - - - 0 0 0 -
- - - - - - - - - - - 0 0 0
- - - - - - - - - 1 - - - 1
- - - - - - - - - - - - - -
```

Output

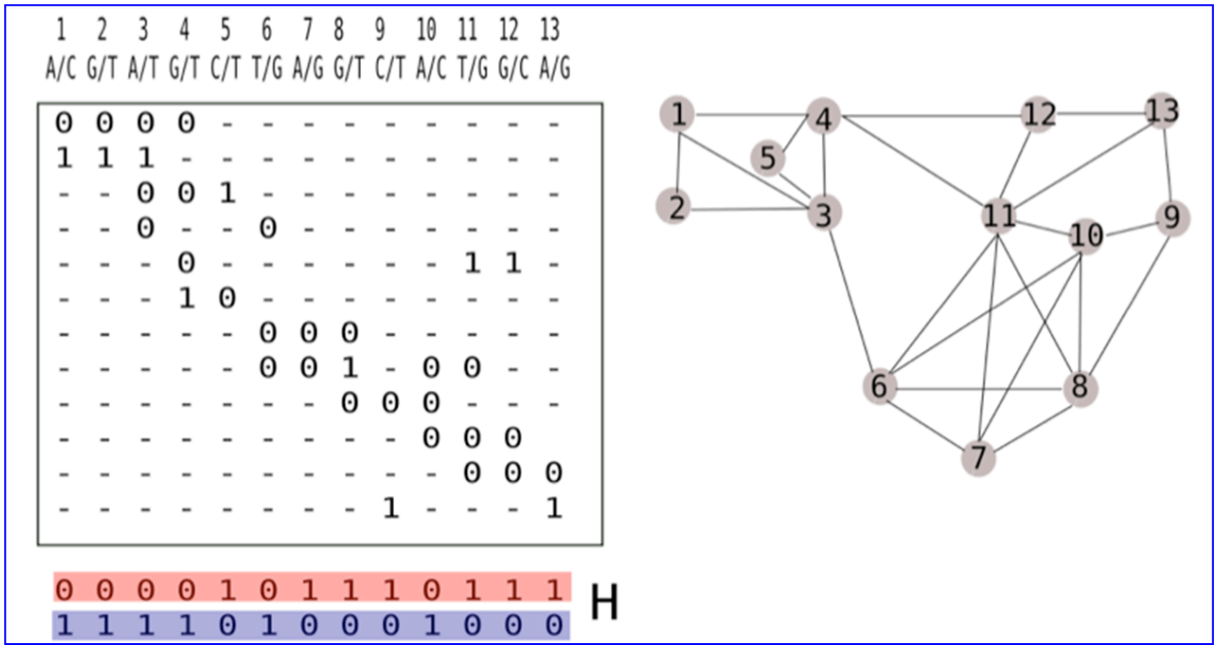
```
0 0 0 0 1 0 0 0 0 0 0 0 0
1 1 1 1 0 1 1 1 1 1 1 1 1
```

4.3.2 Hapcut

Ngoài tìm kiến tham lam thì ta còn có một thuật toán mạnh hơn rất nhiều đó là hapcut.

4.3.2.1 Xây dựng read-haplotype graph

Ý tưởng của chúng ta sẽ là xây dựng một đồ thị từ matrix ban đầu:



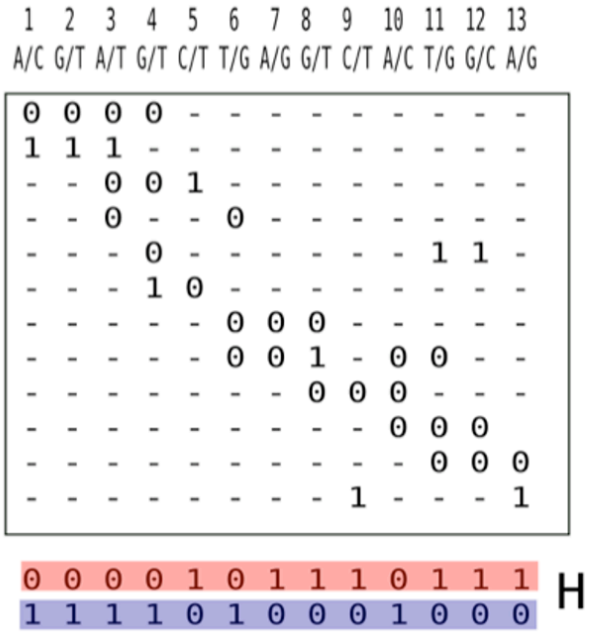
Với 2 đỉnh x, y nếu tồn tại i sao cho `a[i][x] == 0/1 && a[i][y] == 0/1` thì sẽ có cạnh vô hướng từ x đến y. Ta có thể for 2 đỉnh x và y rồi kiểm tra xem bitset hàng x AND bitset hàng y có khác 0 hay không trong $O\left(\frac{n^3}{64}\right)$

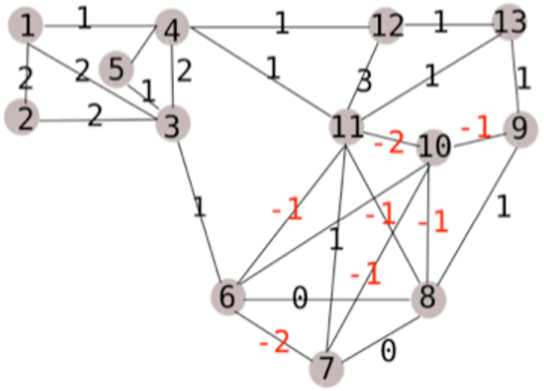
E.g.

- Cạnh 1,3 có cạnh nối vì có hàng 1, 2 thỏa mãn ({0, 0} và {1, 1})
- 1,4 có hàng 1 thỏa mãn ({0, 0})
- 4,11 có hàng 5 thỏa mã ({0, 1})

4.3.2.2 Định nghĩa trọng số cạnh

Gọi `W(i, j)` là số cạnh matching H trừ đi số cạnh mismatching H

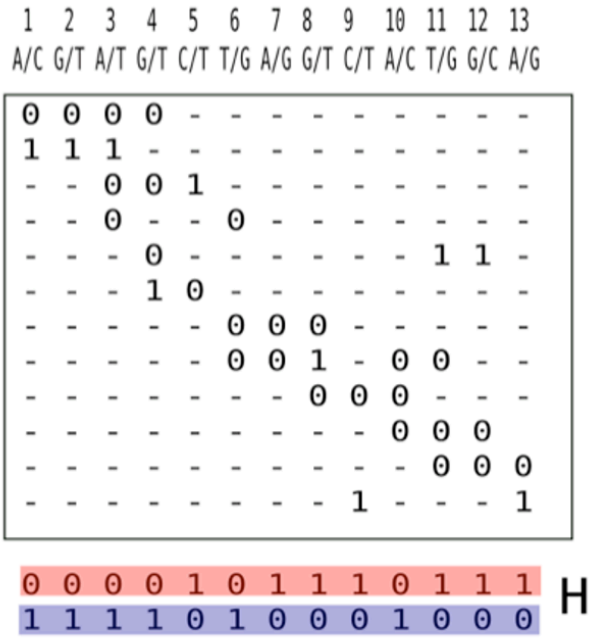


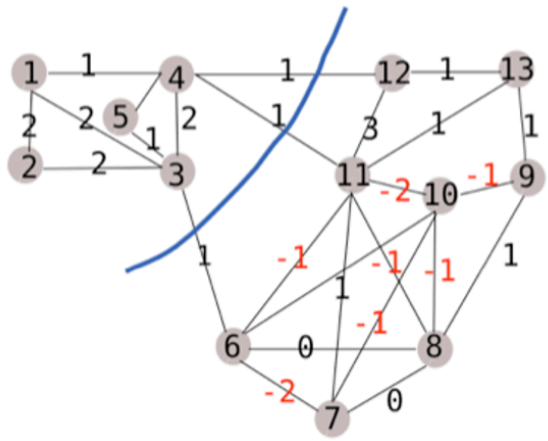


E.g.

- Cạnh 6,7 có 2 pairs {0, 0} đều mismatch với H → W(6,7) = −2
- Cạnh 11,12 có 2 {0, 0} và 1 {1,1} đều match với H → W(11, 12) = 3

4.3.2.3 Good cuts?





S = { 1,2,3,4,5}

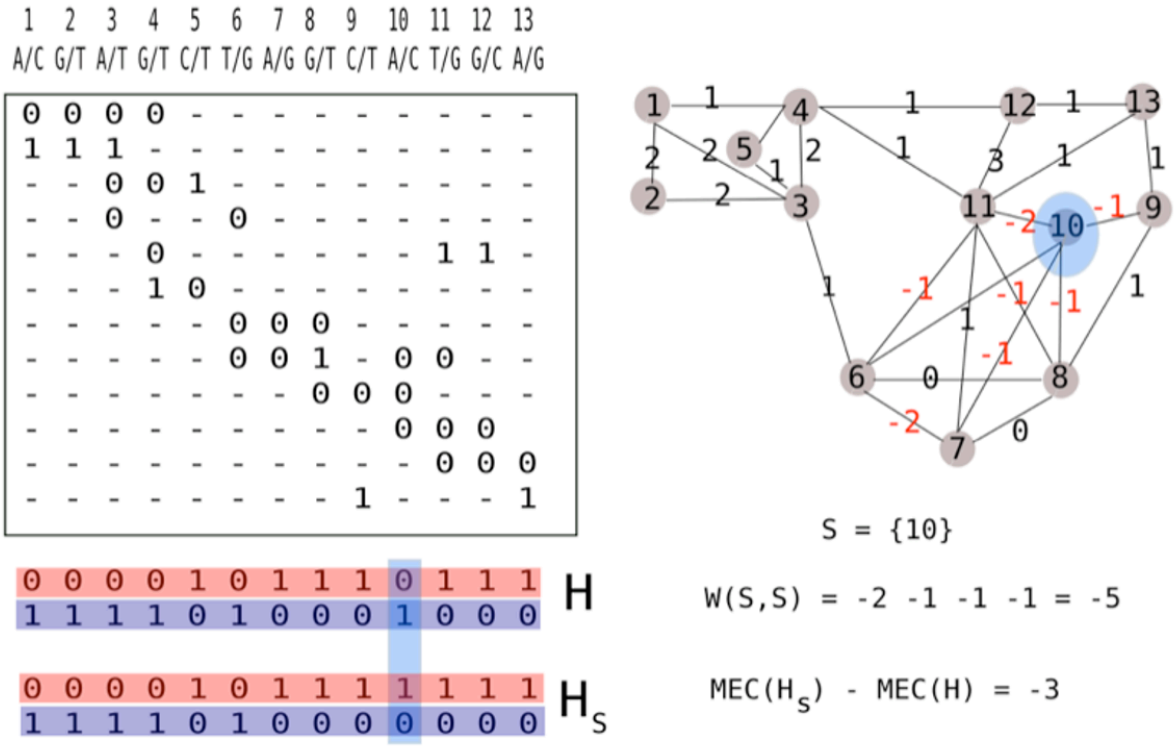
W(S,$\overline{S}$) = 1 + 1 + 1 = 3

Một lát cắt S là một phần của đồ thị sao khi chia đồ thị ra làm đôi.

E.g. Với $S = \{1, 2, 3, 4, 5\}$ ta phải cắt ít nhất là 3 cạnh 3,6 4,11 và 4,12 tức $W\left(S, \overline{S}\right) = 1 + 1 + 1 = 3$

Mục tiêu của ta là tìm ra lát cắt S sao cho:

$$W\left(S, \overline{S}\right) < 0 \rightarrow \text{MEC}(H_S) < \text{MEC}(H)$$



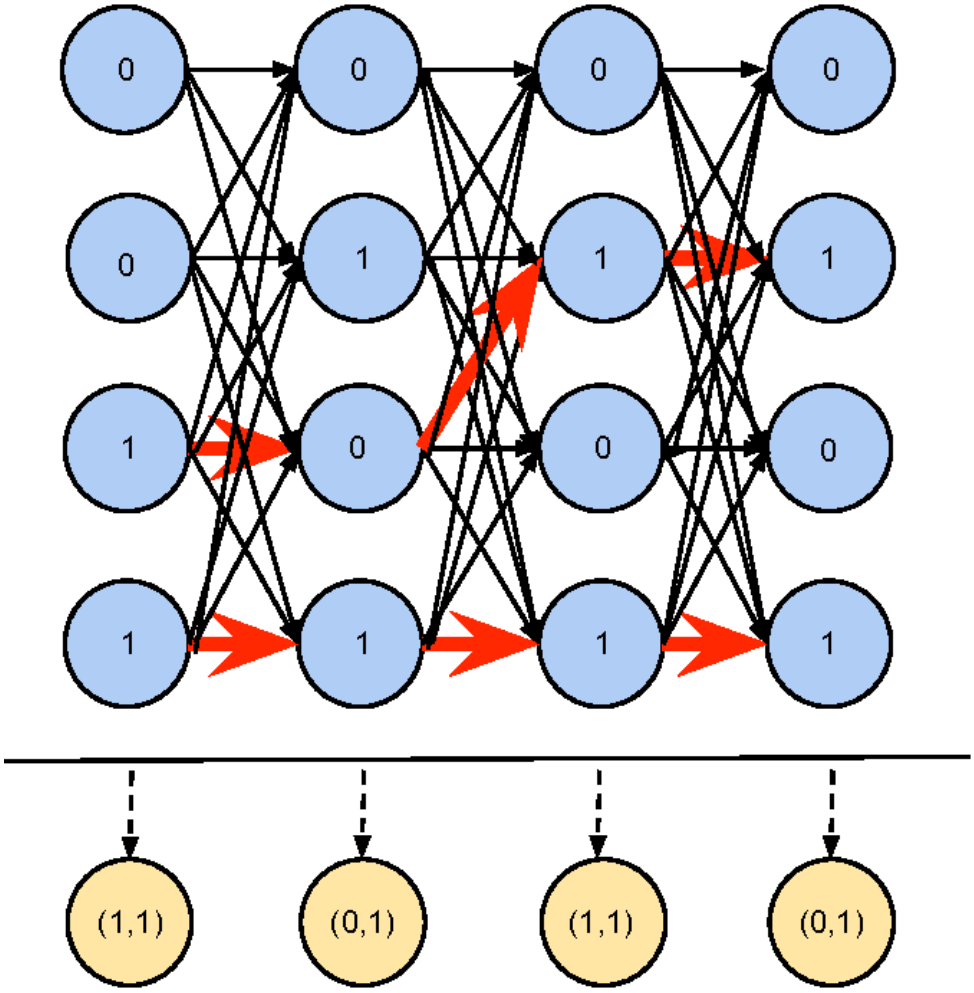
Vì vậy negative cuts là good cuts

4.3.2.4 Thuật toán hapcut

Khởi tạo: Chọn một haplotype H^1 bất kỳ

Phép lặp: For t = 1 $\rightarrow \infty$

1. Xây dựng read-haplotype graph $G(H^1)$
2. Tìm một cut $\left(S, \overline{S}\right)$ trong $G(H^t)$ sao cho W(S, $\overline{S}$) < 0
3. If MEC(H^t_S) <= MEC(H^t): $H^{t+1} = H^t_S$
4. Else $H^{t+1} = H^t$

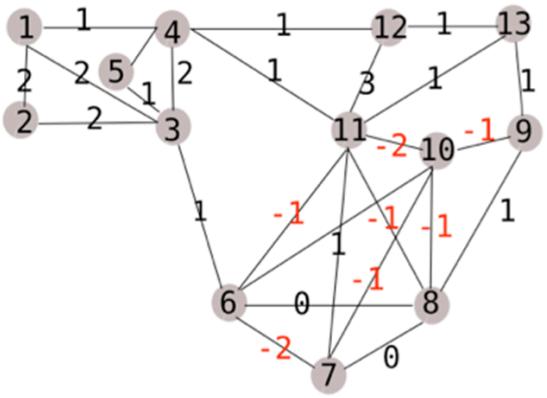


Hình 14: Ảnh minh họa HAPCUT

4.3.2.5 Example

4.3.2.5.1 Khởi tạo haplotype bằng greedy

1	2	3	4	5	6	7	8	9	10	11	12	13
A/C	G/T	A/T	G/T	C/T	T/G	A/G	G/T	C/T	A/C	T/G	G/C	A/G
0	0	0	0	-	-	-	-	-	-	-	-	-
1	1	1	-	-	-	-	-	-	-	-	-	-
-	-	0	0	1	-	-	-	-	-	-	-	-
-	-	0	-	-	0	-	-	-	-	-	-	-
-	-	-	0	-	-	-	-	-	-	1	1	-
-	-	-	1	0	-	-	-	-	-	-	-	-
-	-	-	-	-	0	0	0	-	-	-	-	-
-	-	-	-	-	0	0	1	-	0	0	-	-
-	-	-	-	-	-	-	0	0	0	-	-	-
-	-	-	-	-	-	-	-	-	0	0	0	
-	-	-	-	-	-	-	-	-	-	0	0	0
-	-	-	-	-	-	-	-	1	-	-	-	1

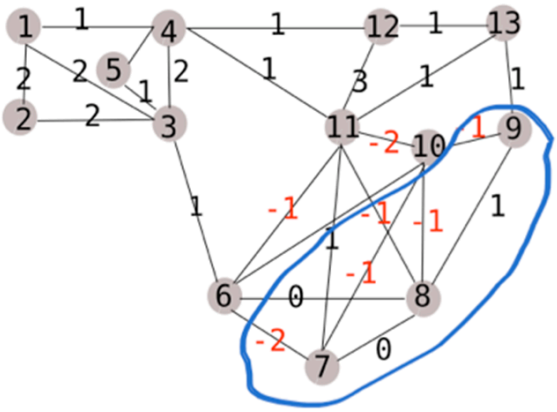


0	0	0	0	1	0	1	1	1	0	1	1	1
1	1	1	1	0	1	0	0	0	1	0	0	0

H

4.3.2.5.2 Tìm negative cut bằng mincut

1	2	3	4	5	6	7	8	9	10	11	12	13
A/C	G/T	A/T	G/T	C/T	T/G	A/G	G/T	C/T	A/C	T/G	G/C	A/G
0	0	0	0	-	-	-	-	-	-	-	-	-
1	1	1	-	-	-	-	-	-	-	-	-	-
-	-	0	0	1	-	-	-	-	-	-	-	-
-	-	0	-	-	0	-	-	-	-	-	-	-
-	-	-	0	-	-	-	-	-	-	1	1	-
-	-	-	1	0	-	-	-	-	-	-	-	-
-	-	-	-	-	0	0	0	-	-	-	-	-
-	-	-	-	-	0	0	1	-	0	0	-	-
-	-	-	-	-	-	-	0	0	0	-	-	-
-	-	-	-	-	-	-	-	-	0	0	0	
-	-	-	-	-	-	-	-	-	-	0	0	0
-	-	-	-	-	-	-	-	1	-	-	-	1



S = {7,8,9}

0	0	0	0	1	0	1	1	1	0	1	1	1
1	1	1	1	0	1	0	0	0	1	0	0	0

H

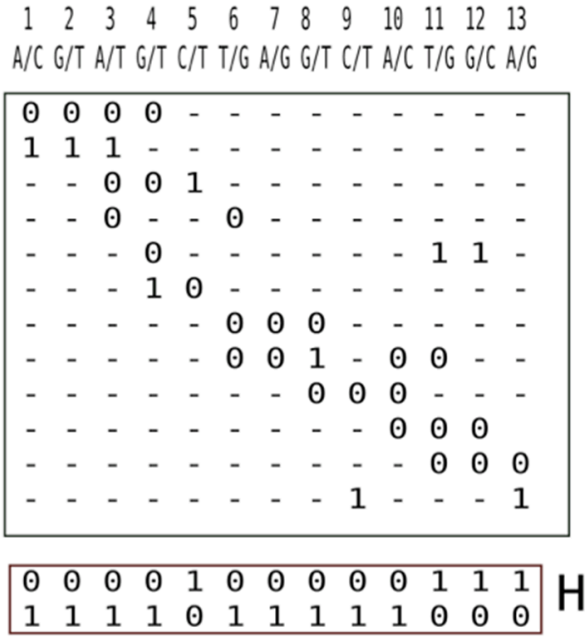
$W(S,\bar{S})$ = -2 + (-3) + 0 = -5

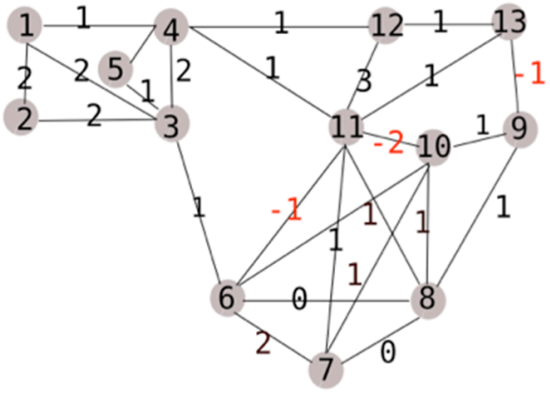
0	0	0	0	1	0	0	0	0	0	1	1	1
1	1	1	1	0	1	1	1	1	1	0	0	0

H_S

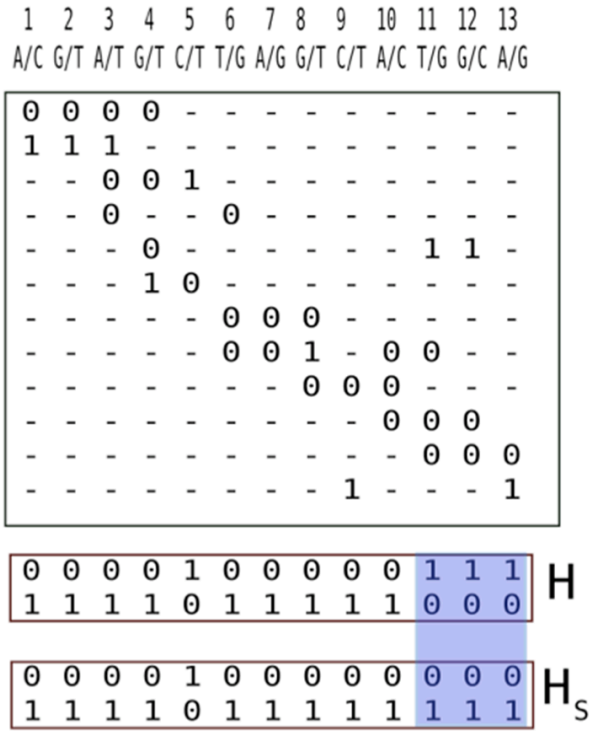
$MEC(H_S)$ - MEC(H) = -3

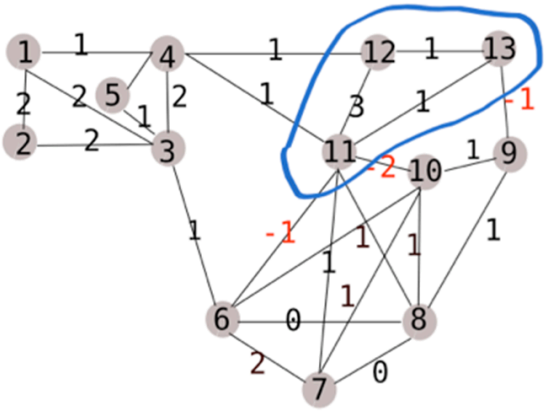
4.3.2.5.3 Flip lại các bit trong S nếu MEC giảm





4.3.2.5.4 Xây lại graph H_S mới





S = {11,12,13}

W(S,$\overline{S}$) = -2 + 1 + (-1) = -2

MEC(H_S) - MEC(H) = -2

Lặp lại bước trên cho đến khi không thể giảm `MEC` được nữa

4.3.2.6 Code

Sau đây là implementation của em cho HAPCUT. Vì code hapcut dài hơn rất nhiều code greedy em nên xin phép được chia ra thành các phần khác nhau

4.3.2.6.1 Template

```cpp
#include <bits/stdc++.h>
using namespace std;
using ll = long long;

int MEC(vector<vector<int>> a, vector<int> h) {
  int n = a.size();
  int errors = 0;
  for (int i = 0; i < n; i++) {
    int cnt1 = 0, cnt2 = 0;
    for (int j = 0; j < n; j++) {
      if (a[i][j] != h[j]) cnt1++;
      if (a[i][j] != 1 - h[j]) cnt2++;
    }
    errors += min(cnt1, cnt2);
  }
  return errors;
}

template<class T>
pair<T, vector<int>> global_min_cut(vector<vector<T>> adjm){
  int n = (int)adjm.size();
  vector<int> used(n);
  vector<int> cut, best_cut;
  T best_weight = -1;
  for(auto phase = n - 1; phase >= 0; -- phase){
```

```
    vector<T> w = adjm[0];
    vector<int> added = used;
    int prev, k = 0;
    for(auto i = 0; i < phase; ++ i){
      prev = k, k = -1;
       for(auto j = 1; j < n; ++ j) if(!added[j] && (k == -1 ||
w[j] > w[k])) k = j;
      if(i == phase - 1){
        for(auto j = 0; j < n; ++ j) adjm[prev][j] += adjm[k][j];
         for(auto j = 0; j < n; ++ j) adjm[j][prev] = adjm[prev]
[j];
        used[k] = true, cut.push_back(k);
         if(best_weight == -1 || w[k] < best_weight) best_cut =
cut, best_weight = w[k];
      }
      else{
        for(auto j = 0; j < n; ++ j) w[j] += adjm[k][j];
        added[k] = true;
      }
    }
  }
  return {best_weight, best_cut};
```

Em sẽ sử dụng global min cut (Wager algorithm) để tìm negative cut, thuật toán này chạy trong $O(n^3)$

4.3.2.6.2 Chọn một haplotype bất kỳ

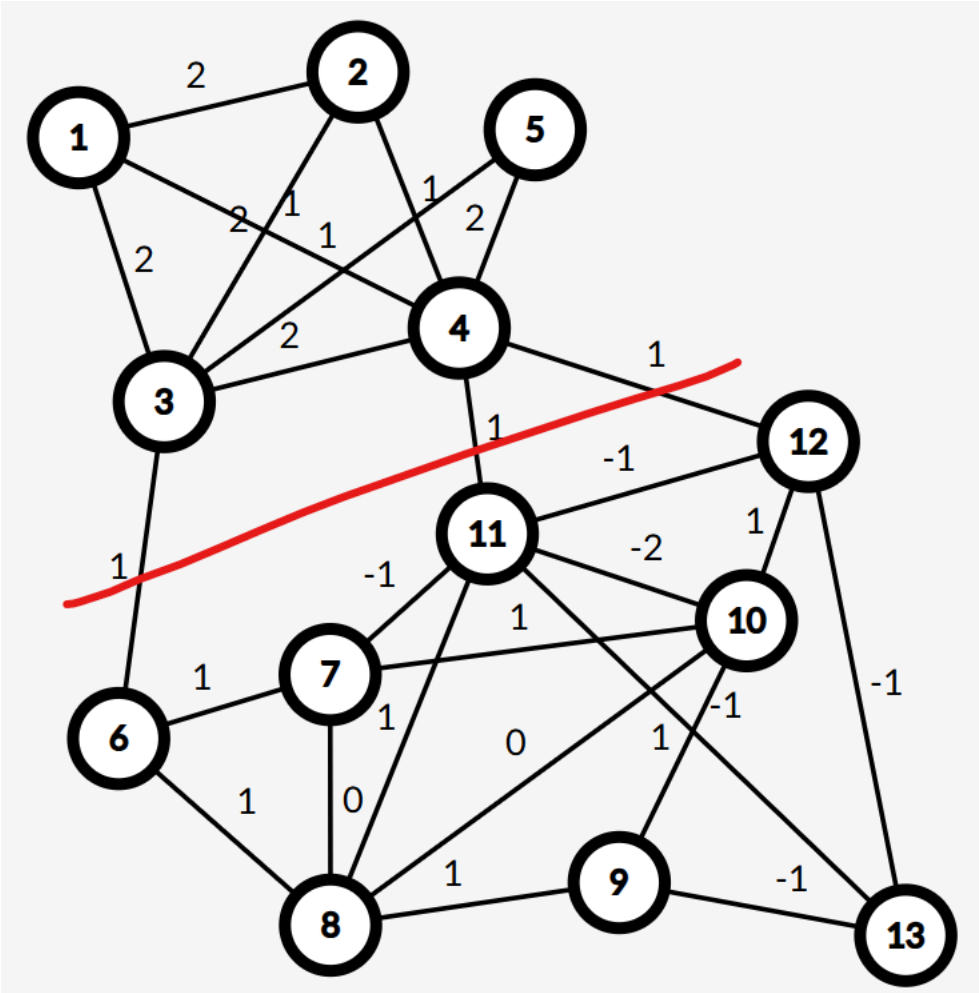
Tương tự với greedy chúng ta sẽ chọn haplotype bất kỳ bằng các extend dần dần

```
int32_t main() {
  cin.tie(0)->sync_with_stdio(0);
  int n;
  cin >> n;
  vector a(n, vector(n, -1));
  for (int i = 0; i < n; i++) {
    for (int j = 0; j < n; j++) {
      char x;
      cin >> x;
      if (x != '-') a[i][j] = x - '0';
    }
  }
  for (int i = 0; i < n; i++) {
    for (int j = 0; j < n; j++) {
      cerr << a[i][j] << ' ';
    }
    cerr << endl;
  }
  // Chọn haplotype h bất kỳ
  vector<int> h(n, -1);
  for (int i = 0; i < n; i++) {
    bool ok = 1;
    for (int j = 0; j < n; j++) {
      if (a[i][j] != -1) {
        if (h[j] != -1 && a[i][j] != h[j]) {
          ok = 0;
          break;
```

```
        }
      }
    }
    if (ok) {
      for (int j = 0; j < n; j++) {
        if (a[i][j] != -1) {
          h[j] = a[i][j];
        }
      }
    }
  }
  for (int j = 0; j < n; j++) if (h[j] == -1) h[j] = 0;
```

4.3.2.6.3 Xây dựng đồ thị

```
  // Xây dựng đồ thị
  auto build_graph = [&]() {
    vector mat(n, vector<int>(n));
    for (int x = 0; x < n; x++) {
      for (int y = x + 1; y < n; y++) {
        bool edge = false;
        int match = 0, mismatch = 0;
        for (int i = 0; i < n; i++) {
          if (a[i][x] != -1 && a[i][y] != -1) {
            edge = true;
            if ((a[i][x] == h[x] && a[i][y] == h[y]) || (a[i][x]
== 1 - h[x] && a[i][y] == 1 - h[y])) {
              match++;
            } else {
              mismatch++;
            }
          }
        }
        if (edge) {
          int w = match - mismatch;
          mat[x][y] += w;
          mat[y][x] += w;
          cerr << x + 1 << ' ' << y + 1 << ' ' << w << endl;
        }
      }
    }
    return move(mat);
  };
  build_graph();
```



4.3.2.6.4 Tìm min cut để tối ưu MEC

Xây dựng đồ thị xong chúng ta cần tìm cách để tìm minimum cut. Em đã sử dụng matrix để lưu cạnh đồ thị để chạy wagner algo

```
int pre_mec = MEC(a, h);
for (int i = 0; i < n; i++) {
  vector<vector<int>> mat = build_graph();
  auto [weight, cuts] = global_min_cut<int>(mat);
  cerr << "Weight: " << weight << endl;
  cerr << "Cuts: ";
  for (auto x : cuts) cerr << x << ' ';
  cerr << endl;
  if (weight < 0) {
    vector<int> nh(h);
```

```
      for (auto cut : cuts) {
        nh[cut] = 1 - nh[cut];
      }
      int cur_mec = MEC(a, nh);
      if (cur_mec < pre_mec) {
        for (auto cut : cuts) {
          h[cut] = 1 - h[cut];
        }
        pre_mec = cur_mec;
      }
    }
  }
  for (int j = 0; j < n; j++) {
    cout << h[j] << " \n"[j == n - 1];
  }
  for (int j = 0; j < n; j++) {
    cout << 1 - h[j] << " \n"[j == n - 1];
  }
}
```

Tổng độ phức tạp của HAPCUT là $O(\text{iter} * n^3)$ với iter là số vòng lặp (càng cao càng thì MEC càng nhỏ) và n^3 là độ phức tạp của Wager algorithm

Input

```
13
0 0 0 0 - - - - - - - - - -
1 1 1 - - - - - - - - - - -
- - 0 0 1 - - - - - - - - -
- - 0 - - 0 - - - - - - - -
- - - 0 - - - - - - - 1 0 -
- - - 1 0 - - - - - - - - -
- - - - - 0 0 0 - - - - - -
- - - - - - - 0 1 - 0 0 - -
- - - - - - - - 0 0 0 - - -
- - - - - - - - - - 0 0 0 -
- - - - - - - - - - - 0 0 0
- - - - - - - - - 1 - - - 1
- - - - - - - - - - - - - -
```

Output

```
0 0 0 0 1 0 0 0 0 0 0 0 0
1 1 1 1 0 1 1 1 1 1 1 1 1
```

4.4 So sánh haplotyping và genotyping

Genotyping và haplotyping là hai phương pháp quan trọng trong nghiên cứu di truyền học, mỗi phương pháp có những đặc điểm và ứng dụng riêng biệt. Genotyping là quá trình xác định các alen cụ thể tại các vùng trên nhiễm sắc thể của một cá thể, tập trung vào việc nhận diện sự khác biệt di truyền giữa các cá thể dựa trên các alen cụ thể. Phương pháp này thường được sử dụng trong các nghiên cứu liên kết gen-bệnh, xác định kiểu gen của các cá thể trong nghiên cứu di truyền học, và trong các nghiên cứu liên quan đến chọn lọc giống. Mặc dù genotyping đơn giản hơn và ít tốn kém

hơn, nó có thể bỏ qua thông tin về liên kết giữa các biến thể, làm giảm độ nhạy của các nghiên cứu liên kết.

Ngược lại, haplotyping là quá trình xác định tập hợp các biến thể DNA tại các vùng kề nhau trên một nhiễm sắc thể, những biến thể này có xu hướng được di truyền cùng nhau. Phương pháp này cung cấp thông tin chi tiết hơn về cấu trúc di truyền và sự liên kết giữa các biến thể, thường được sử dụng trong các nghiên cứu liên kết haplotype-bệnh, lập bản đồ liên kết di truyền, và trong các nghiên cứu về di truyền học dân số. Haplotype-based association methods có thể mạnh hơn vì chúng nắm bắt được thông tin mất cân bằng liên kết (LD), nhưng cũng có thể gặp khó khăn về việc giải quyết pha giao tử và tăng bậc tự do khi số lượng locus tăng lên.

Mặc dù haplotyping phức tạp hơn và đòi hỏi nhiều dữ liệu hơn để xác định chính xác các haplotype, nó cung cấp nhiều thông tin hơn trong các nghiên cứu liên kết chi tiết. Trong khi đó, genotyping có thể đơn giản hơn và tiết kiệm chi phí, nhưng sức mạnh phân tích của nó có thể bị giảm do thông tin LD chứa trong các dấu hiệu kề bên bị bỏ qua. Tóm lại, cả hai phương pháp đều có vai trò quan trọng trong nghiên cứu di truyền học và được lựa chọn tùy theo mục tiêu cụ thể của từng nghiên cứu.

Tiêu chí	**Genotyping**	**Haplotyping**
Định nghĩa	Xác định các alen cụ thể tại các locus trên nhiễm sắc thể của một cá thể.	Xác định tập hợp các biến thể DNA (SNPs và indels) tại các locus kề nhau trên một nhiễm sắc thể.
Thông tin thu thập	Các alen cụ thể tại các vị trí xác định trên nhiễm sắc thể.	Các tập hợp alen liên kết với nhau trên một đoạn nhiễm sắc thể.

Tiêu chí	**Genotyping**	**Haplotyping**
Ứng dụng	• Nghiên cứu liên kết gen-bệnh • Xác định kiểu gen của cá thể • Nghiên cứu chọn lọc giống	• Nghiên cứu liên kết haplotype - bệnh • Lập bản đồ liên kết di truyền • Nghiên cứu di truyền học dân số
Độ chính xác và phức tạp	• Đơn giản hơn • Ít tốn kém hơn	• Cung cấp thông tin chi tiết hơn • Phức tạp hơn • Đòi hỏi nhiều dữ liệu hơn)
Sức mạnh phân tích	• Có thể bỏ qua thông tin liên kết giữa các biến thể • Giảm độ nhạy của các nghiên cứu liên kết	• Nắm bắt thông tin mất cân bằng liên kết (LD) • Cung cấp nhiều thông tin hơn • Gặp khó khăn trong việc giải quyết pha giao tử • Bậc tự do tăng khi số lượng locus tăng lên)

Tiêu chí	**Genotyping**	**Haplotyping**
Sức mạnh phân tích	• Có thể bỏ qua thông tin liên kết giữa các biến thể • Giảm độ nhạy của các nghiên cứu liên kết	• Nắm bắt thông tin mất cân bằng liên kết (LD) • Cung cấp nhiều thông tin hơn • Gặp khó khăn trong việc giải quyết pha giao tử • Bậc tự do tăng khi số lượng locus tăng lên)
Nhược điểm	• Bỏ qua thông tin LD từ các dấu hiệu kề bên • Độ nhạy thấp hơn	• Phức tạp và tốn kém hơn • Yêu cầu dữ liệu lớn và tính toán phức tạp
Ví dụ về ứng dụng thực tiễn	• Phân tích kiểu gen cho các bệnh di truyền đơn gen • Xác định alen trong các chương trình chọn giống	• Lập bản đồ liên kết chi tiết cho các bệnh phức tạp • Nghiên cứu nguồn gốc và sự di cư của các quần thể người

4.5 Ứng dụng thực tiễn

4.5.1 Y học cá nhân

Trong lĩnh vực y học cá nhân, phân tích tổ hợp gen đóng vai trò quan trọng trong việc cải thiện chăm sóc sức khỏe cá nhân. Một trong những ứng dụng tiêu biểu của tổ hợp gen là dự đoán nguy cơ mắc bệnh. Ví dụ, một nghiên cứu đã chỉ ra rằng các tổ hợp gen liên quan đến gen BRCA1 và BRCA2 có thể được sử dụng để đánh giá nguy cơ mắc ung thư vú và ung thư buồng trứng ở phụ nữ. Các phân tích tổ hợp gen này giúp xác định nhóm người có nguy cơ cao hơn và đưa ra các biện pháp phòng ngừa hoặc theo dõi sớm để giảm thiểu nguy cơ mắc bệnh.

Ngoài ra, haplotype cũng đóng vai trò quan trọng trong việc cá nhân hóa điều trị thuốc. Ví dụ, một số bệnh nhân có tổ hợp gen đặc biệt có thể phản ứng khác biệt với các loại thuốc. Nghiên cứu đã chỉ ra rằng gen CYP2D6 ảnh hưởng đến việc chuyển hóa một số thuốc chống trầm cảm và thuốc giảm đau. Bằng cách phân tích tổ hợp gen này, bác sĩ có thể điều chỉnh liều lượng và loại thuốc sử dụng cho từng bệnh nhân, tăng khả năng điều trị và giảm thiểu các tác dụng phụ.

Hơn nữa, trong việc chẩn đoán bệnh di truyền, tổ hợp gen đóng vai trò quan trọng trong việc xác định các bệnh di truyền như bệnh Huntington, xơ nang (cystic fibrosis), và bệnh thalassemia. Phân tích tổ hợp gen giúp chẩn đoán chính xác bệnh và đưa ra kế hoạch điều trị phù hợp. Ví dụ, kiểm tra tổ hợp gen trước khi sinh con có thể giúp các cặp vợ chồng đánh giá nguy cơ và đưa ra quyết định sinh sản phù hợp, giảm thiểu nguy cơ mắc bệnh di truyền cho con cháu.

4.5.2 Nghiên cứu dịch tễ học

Trong lĩnh vực nghiên cứu dịch tễ học, phân tích tổ hợp gen (haplotype) đóng vai trò then chốt trong việc phát hiện và tìm hiểu sự phát triển và lây lan của các bệnh truyền nhiễm. Từ việc xác định tổ hợp gen của các tác nhân gây bệnh đến việc đánh giá mối quan hệ giữa ýếu tố di truyền và sự lây lan của dịch bệnh, phương pháp này chứa đựng nhiều kỹ thuật quan trọng trong việc xác định và ứng phó với các vấn đề dịch tễ học hiện đại.

Một ví dụ cụ thể về ứng dụng của phân tích tổ hợp gen trong nghiên cứu dịch tễ học là việc nghiên cứu về đại dịch COVID-19. Trong bối cảnh này, các nhà nghiên cứu đã sử dụng phương pháp này để phân loại và theo dõi các biến thể của virus SARS-CoV-2. Bằng việc phân tích tổ hợp gen của virus từ các mẫu được thu thập từ các ca nhiễm, họ đã có thể theo dõi sự biến đổi của virus theo thời gian, từ đó đưa ra các thông tin cần thiết về tốc độ lây lan, mức độ nguy hiểm, và độ phản ứng của virus đối với các biện pháp kiểm soát.

Ngoài ra, phân tích tổ hợp gen cũng giúp cho việc hiểu rõ hơn về cơ chế di truyền của các bệnh truyền nhiễm và đặc điểm di truyền của các nguồn gốc dịch bệnh trờ nên dễ dàng hơn. Bằng cách theo dõi sự biến đổi của các tổ hợp gen của vi khuẩn và virus, nhà nghiên cứu đã có thể xác định các yếu tố góp phần vào sự lan truyền nhanh chóng và hiệu quả của dịch bệnh. Thông tin này đóng vai trò quan trọng trong việc phát triển các chiến lược kiểm soát và phòng ngừa hiệu quả hơn.

Những ứng dụng của phân tích tổ hợp gen trong nghiên cứu dịch tễ học không chỉ là cơ sở cho việc hiểu rõ hơn về cơ chế lây lan của các bệnh truyền nhiễm mà còn cung cấp các thông tin quan trọng để phát triển các chiến lược đối phó và điều trị, góp phần vào việc kiểm soát và ngăn chặn sự lây lan của các dịch bệnh.

4.5.3 Sinh học tiến hóa

.Việc nghiên cứu haplotype cho phép các nhà khoa học theo dõi sự di chuyển và trao đổi của các gen trong quần thể, giúp hiểu rõ hơn về lịch sử tiến hóa của loài. Các mẫu haplotype có thể được so sánh giữa các nhóm cá thể hoặc loài để xác định nguồn gốc chung và mối quan hệ di truyền. Hơn nữa, những thay đổi trong tần suất các haplotype theo thời gian có thể tiết lộ dấu hiệu của chọn lọc tự nhiên, từ đó nhận diện được các gen ưu việt hơn trong quá trình tiến hóa. Sự khác biệt về haplotype cũng có thể được sử dụng để xác định các sự kiện lai ghép trong quá khứ, như lai tạo giữa các loài gần gũi. Ngoài ra, haplotype liên quan đến bệnh tật hoặc phản ứng với thuốc còn có ứng dụng trong y sinh, giúp dự đoán nguy cơ mắc bệnh hoặc hiệu quả của một loại thuốc.

4.5.4 Nông nghiệp

Việc nghiên cứu và phân tích haplotype có nhiều ứng dụng quan trọng trong lĩnh vực nông nghiệp. Đầu tiên, các mẫu haplotype có thể được sử dụng để xác định các gen ưu việt liên quan đến các đặc tính mong muốn như năng suất, chất lượng, khả năng chịu bệnh của cây trồng và vật nuôi. Điều này cho phép các chuyên gia thiết kế chương trình chọn lọc hiệu quả hơn nhằm cải thiện các đặc tính quan trọng.

Ngoài ra, phân tích haplotype cũng giúp theo dõi sự đa dạng di truyền trong các quần thể cây trồng và vật nuôi, cung cấp thông tin quan trọng để xác định những nguồn gen cần được bảo tồn nhằm duy trì sự phong phú di truyền. Các mẫu haplotype còn có thể được sử dụng để xác định nguồn gốc và định danh các giống cây trồng và vật nuôi, ứng dụng này rất hữu ích trong công tác quản lý và bảo hộ các giống.

Thêm nữa, phân tích haplotype có thể giúp xác định các gen liên quan đến khả năng chịu bệnh của cây trồng và vật nuôi, hỗ trợ việc lựa chọn và di

truyền các gen kháng bệnh để cải thiện sức khỏe. Các mẫu haplotype liên quan đến các đặc tính như hương vị, kích thước, màu sắc của sản phẩm nông nghiệp cũng có thể được sử dụng để chọn lọc và cải thiện các đặc tính mong muốn.

4.5.5 Phát hiện và bảo tồn đa dạng sinh học

Các mẫu haplotype có thể được sử dụng để theo dõi sự đa dạng di truyền của các quần thể động vật và thực vật hoang dã. Điều này cung cấp thông tin quan trọng về tình trạng sức khỏe và khả năng sinh sản của các quần thể, giúp các nhà bảo tồn xác định những loài và quần thể cần được ưu tiên bảo vệ.

Ngoài ra, phân tích haplotype còn có thể được sử dụng để xác định nguồn gốc và định danh các cá thể động vật và thực vật hoang dã. Ứng dụng này rất hữu ích trong việc giám sát các hoạt động buôn bán và săn bắt trái phép các loài động vật và thực vật quý hiếm.

Các mẫu haplotype cũng có thể cung cấp thông tin về các đặc tính di truyền liên quan đến khả năng chịu bệnh, khả năng thích nghi với môi trường và các đặc tính quan trọng khác của các loài hoang dã. Những thông tin này hỗ trợ việc lựa chọn các cá thể có đặc tính ưu việt để bảo tồn và phục hồi các quần thể.

Hơn nữa, việc nghiên cứu haplotype còn giúp phát hiện các loài mới chưa được biết đến hoặc các biến thể di truyền độc đáo, cung cấp thông tin quý giá cho các nghiên cứu về tiến hóa và sinh học bảo tồn.

5 Kết luận

Chúng em đã tìm hiểu về genotyping và haplotyping cũng như thử code các thuật toán để tìm kiếm haplotype set. Mong là trong tương lai gần những kiến thức của lập trình thi đấu sẽ được mở rộng sang áp dụng tin sinh học và có chỗ đứng trong phân tích gen nói riêng và lĩnh vực bioinformatics nói chung. Nhóm 3 chúng em xin cân thành cảm ơn thầy TS. Lê Sỹ Vinh đã theo dõi, truyền đạt và bổ sung kiến thức về Tin sinh học trong suốt học kỳ qua cho chúng em. Cảm ơn thầy đã theo dõi bài báo cáo của bọn em!