THUẬT TOÁN LẬP BẢN ĐỒ GEN ĐỂ XÁC ĐỊNH VỊ TRÍ GEN MANG MẦM BỆNH

Huỳnh Thị Mỹ Trang, Trần Văn Lăng
Phân viện Công nghệ thông tin tại TP.HCM
(Bài nhận ngày 24 tháng 01 năm 2006, hoàn chỉnh sửa chữa ngày 17 tháng 04 năm 2006)

TÓM TẮT: *"Lập bản đồ gen" chính là việc lập bản đồ của các gen để xác định vị trí trên các nhiễm sắc thể. Đây là một bước then chốt trong việc hiểu về các bệnh di truyền. Có hai loại "lập bản đồ gen": lập bản đồ di truyền – sử dụng phân tích liên kết để xác định mối quan hệ của hai gen trên một nhiễm sắc thể; lập bản đồ vật lý – sử dụng các kỹ thuật hoặc các thông tin sẵn có để xác định vị trí tuyệt đối của gen trên một nhiễm sắc thể. Trong bài báo này chúng tôi đề xuất một hướng tiếp cận qua đó nâng cao hiệu suất của thuật toán sử dụng việc phân tích liên kết để lập bản đồ gen. Chúng tôi đã xây dựng thuật toán dùng phương pháp Haplotype Pattern Mining (HPM) và Density Based Spatial Clustering of Application with Noise (DBSCAN). Thuật toán này được thực hiện trên hệ thống tính toán lưới gồm các cluster của IOIT-HCM (Phân viện Công nghệ thông tin tại TP. Hồ Chí Minh) và của KISTI (Korea Institute of Science and Technology Information)..*

1. GIỚI THIỆU

Lập bản đồ gen thường dựa trên việc phân tích các trình tự di truyền gọi là haplotype. Một haplotype là một đại diện của DNA nằm dọc theo sợi nhiễm sắc thể. Trong các nghiên cứu quần thể bị bệnh và khoẻ mạnh, haplotype là chuỗi có chiều dài cố định. Khi thừa kế từ thế hệ này đến thế hệ khác, các haplotype được tái kết hợp bằng trao đổi chéo. Quá trình này làm tăng trạng thái biến dị của haplotype. Chính trạng thái biến dị này phản ảnh tính lịch sử của mỗi haplotype, hai haplotype có cùng một tổ tiên có khả năng chia sẻ chung một phân đoạn DNA của tổ tiên. Trong lập bản đồ kết hợp (association mapping), các nhà di truyền học tìm kiếm các phân đoạn tiêu biểu nhất của các bệnh nhân tương ứng với một loại bệnh nào đó. Vị trí của các phân đoạn này là vị trí của các gen ảnh hưởng đến bệnh, các phân đoạn này cũng như gen được kế thừa từ một tổ tiên chung.

Lập bản đồ kết hợp dựa trên giả định rằng nhiều người mang mầm bệnh của gen đã thừa kế từ tổ tiên của họ và vì vậy họ chia sẻ cùng đoạn mẫu haplotype. Tiêu chuẩn đánh giá độ tương đồng cho các cá thể là một công cụ hữu ích cho việc đánh giá các bệnh nhân có quan hệ huyết thống gần và giúp cho việc tìm ra các cấu trúc trong mẫu haplotype. Gom nhóm dựa trên tiêu chuẩn đánh giá độ tương đồng, có thể định vị các nhóm cá thể có khả năng chia sẻ cùng một gen mang mầm bệnh do di truyền. Hướng tiếp cận này tập trung nghiên cứu về khai phá dữ liệu các mẫu haplotype. Lập bản đồ gen được xem xét trong mối quan hệ giữa trạng thái bệnh và trạng thái khoẻ mạnh của các cá thể trong quần thể. Mục tiêu chính là đánh giá mức độ tương đồng giữa các haplotype và gom nhóm chúng bằng thuật toán DBSCAN **[1]**, sau đó dùng thuật toán Haplotype Pattern Mining **[2]** khám phá mẫu haplotype để tìm vị trí gen mang mầm bệnh. Tất cả các phương pháp áp dụng dựa trên các quy tắc khám phá tính tương đồng giữa các mẫu haplotype. Trong bài toán lập bản gen, tiêu chuẩn đánh giá tương đồng được xem xét trong mối quan hệ giữa hai trạng thái bệnh và khoẻ mạnh của các cá thể trong phạm vi nghiên cứu.

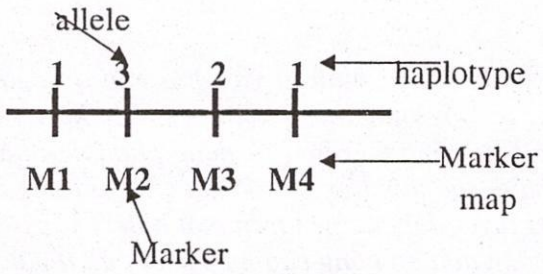
2. LẬP BẢN ĐỒ GEN

2.1.Một số khái niệm di truyền

Marker là nơi cung cấp thông tin về biến thể di truyền giữa cộng đồng người. Chúng là các vị trí đa hình trong bộ gen, các biến thể này thu được từ một cá thể mang bệnh được xác định

bằng phương pháp thực nghiệm. Vị trí của một marker thường được gọi là ***locus*** (nhiều locus gọi là loci). Các biến thể tại một marker được gọi các ***allele***. Người ta thường sử dụng số nguyên để đại diện cho các allele này. Tập các marker gọi là bản đồ marker.

Haplotype bao gồm một tập các allele nhận được tại các vị trí marker dọc theo một sợi nhiễm sắc thể. Hình 1 minh họa hình ảnh về marker, allele, haplotype.



Hình 1: Sơ đồ tổng quát các thành phần di truyền dùng trong bài toán.

Ví dụ: Gọi M1, M2, M3, M4 là các marker, định vị dọc theo một nhiễm sắc thể. Giả sử cho các allele tại 4 vị trí marker trên là 1, 3, 2, 1. Haplotype trên bốn marker trong nhiễm sắc thể này là [1 3 2 1], và haplotype trên marker M2 và M4 là [3 1].

Có hai loại marker chung, đó là marker dạng microsatellite và marker dạng SNP (*Single Nucleotide Polymorphism*). Marker dạng microsatellite (*STR – Short Tandem Repeats*) có khoảng 20 allele khác nhau, mỗi allele tương ứng với số nguyên chỉ số lần lặp lại trong trình tự DNA của cá thể. Đối với marker dạng SNP luôn luôn có 2 allele, nhưng marker loại SNP có tần số xuất hiện nhiều hơn trong bộ gen, vì vậy cho phép bản đồ marker dày đặc và thích hợp hơn cho việc lập bản đồ chính xác. Marker loại SNP ổn định hơn STR. Tốc độ đột biến của SNP được đánh giá khoảng là 10^{-8} trong quá trình phân bào giảm nhiễm, còn STR là 10^{-3}.

2.2.Đặt bài toán

Bài toán lập bản đồ gen được phát biểu như sau:

Gọi A là tập dữ liệu haplotype khỏe mạnh. Gọi C là tập dữ liệu haplotype bị bệnh. Thông qua ngưỡng thống kê x, để truy tìm tất cả các tập mẫu tiềm năng thoả ngưỡng thống kê. Từ đó suy ra kết quả dự đoán vị trí gen mang mầm bệnh dựa trên tần số xuất hiện cao nhất, hoặc kết quả dự đoán điểm dựa trên giá trị p-value thu được thông qua kiểm tra hoán vị.

3. PHƯƠNG PHÁP

Với tập tin dữ liệu đầu vào lớn, và sử dụng marker dạng microsatellite, sẽ tốn khá nhiều thời gian cho việc truy tìm mẫu trong thuật toán Haplotype Pattern Mining. Ngoài ra, tập dữ liệu đầu vào bao gồm tất cả cá thể trong phạm vi nghiên cứu, mỗi cá thể được đại diện bằng một haplotype. Để rút ngắn thời gian tìm mẫu tiềm năng, chúng tôi đề xuất phương pháp dùng thuật toán DBSCAN để lọc bỏ những cá thể đơn lẻ, không có khả năng mang mầm bệnh.

Ưu điểm của thuật toán DBSCAN là tìm nhóm dựa theo mật độ tương đồng. Phương pháp thống kê χ^2 hoặc Z-score, căn cứ trên số mẫu bị bệnh và mẫu khoẻ mạnh trong một nhóm, từ đó đi đến quyết định là nhóm này có kết hợp mạnh với bệnh hay không. Với những nhóm thỏa mãn điều kiện thống kê, dùng thuật toán Haplotype Pattern Mining truy tìm các mẫu tiềm năng, thật sự kết hợp mạnh với bệnh. Bước cuối cùng của hướng tiếp cận này, sẽ tổng hợp dữ liệu mẫu gởi về từ các nhóm. Với tập mẫu tiềm năng tổng hợp, vị trí gen mang mầm bệnh được dự đoán dựa trên tần số xuất hiện và p-value tại mỗi marker.

Hiệu quả của hướng tiếp cận này phụ thuộc vào tiêu chuẩn đánh giá tương đồng haplotype sử dụng trong thuật toán gom nhóm. Trong phần tiêu chuẩn đánh giá tương đồng sẽ xử lý vị trí mang allele không xác định trong mẫu haplotype, và nó được xem như một allele mới.

3.1.Tiêu chuẩn đánh giá tương đồng

Cho tập G chứa n cá thể, mỗi cá thể đại diện một chuỗi haplotype với m maker. Hàm đánh giá tương đồng *sim:* G x $G \rightarrow [0,1]$, nếu hàm nhận giá trị bằng 0, tất cả các allele trong hai haplotype không tương đồng, và bằng 1, tất cả các allele trong hai haplotype tương đồng. Các haplotype có quan hệ họ hàng có mức độ tương đồng càng cao và chia sẻ nhiều di truyền IBD (identical by descent).

Gọi H_1, H_2 là hai haplotype thuộc tập G, so sánh từng cặp allele tại vị tri các marker. Gọi vector $S_{H1,H2} = (s_1, \ldots, s_m)$, với $s_i = 1$ nếu $H_1(i) = H_2(i)$, ngược là $s_i = 0$, $1 \le i \le m$.

3.2. Phương pháp 1 [3]

Đầu tiên, xem xét kỹ thuật phân chia cửa sổ, một cửa sổ có chiều dài $w \in N$. Với mỗi marker thứ k, tính $a_k = \sum_{i=k}^{k+w-1} s_i$, với $s_i = 0$ cho $i \notin \{1,\ldots,m\}$.

Tiếp theo, tính độ tương đồng $a = \sum_{k=(-w+2)}^{m} a_k^{\alpha}$, với $\alpha \ge 1$. Giá trị số mũ α mang lại trọng số mong muốn trong đoạn quan sát mất cân bằng liên kết (LD – Linkage disequilibrium), với giá trị α lớn hơn thì chiều dài $s_i = 1$ trong vector S sẽ dài hơn.

Tiếp theo tính giá trị *score*(C) lớn nhất có thể $C = (m-w+1)w^{\alpha} + 2\sum_{k=1}^{w-1} k^{\alpha}$

Hàm tương đồng là $sim(H_1, H_2) = a/C$

Hàm khoảng cách là $1 - sim(H_1, H_2)$.

Phương pháp này mang lại kết quả tốt trong trường hợp dữ liệu haplotype nguồn có đột biến mất dữ liệu và đột biến điểm.

3.3. Phương pháp 2 [3]

Tìm tất cả những chuỗi con $s_i, \ldots, s_k$ mang giá trị 1 trong vector $S_{H1,H2}$, sao cho $1 \le i < k \le m$, và $s_{i-1} = s_{k+1} = 0$.

Gọi S là tập các chuỗi con này, ta có, $a = \sum_{s \in S} |s|^{\alpha}, \alpha \ge 1$, $|s|$ là chiều dài chuỗi con. Khi đó, hàm tương đồng là $sim(H_1, H_2) = a/m^{\alpha}$.

Hàm khoảng cách là $1 - sim(H_1, H_2)$

Các phương pháp trên khá đơn giản và dễ cài đặt, độ phức tạp của thuật toán là $\Theta(m)$. Tuy nhiên, việc xác định hai hằng số w và α không được đề cập rõ ràng. Chúng phải có một giá trị đủ lớn để phân biệt giữa chia sẻ IBD và chia sẻ ngẫu nhiên. Mặt khác, các tham số này không được quá lớn, nếu không sẽ dẫn đến tình huống các giá trị tương đồng là không đáng kể và phương pháp haplotype riêng lẻ không thoả đáng. Để an toàn, nên chọn giá trị w và α trong khoảng $1 \le w \le 5$ và $1 \le \alpha \le 2$ cho bản đồ marker dài.

3.4.Đánh giá độ kết hợp mạnh với tính trạng bệnh của nhóm

Gom nhóm là một thuật toán mạnh dùng trong việc khai phá tập dữ liệu lớn. Trong nghiên cứu lập bản đồ gen, việc gom nhóm các haplotype nhằm mục tiêu tìm ra các haplotype có quan hệ họ hàng, các nhóm có thể tương ứng với các đột biến gây bệnh khác nhau. Giả thiết, cách ly một số nhỏ cá thể và số cá thể này phát triển thành một quần thể. Giả định rằng đột biến gen mang mầm bệnh đang quan tâm ở thế hệ đầu tiên. Và sau nhiều thế hệ các cá thể mang các đột biến khác nhau có thể được tìm thấy thông qua thuật toán gom nhóm. Mục đích không phân

chia tất cả các haplotype vào trong các nhóm, vì có những haplotype từ cá thể nhiễm bệnh hoặc khoẻ mạnh không cần thiết có trong nhóm, và các haplotype từ cá thể mang bệnh sẽ có độ tương đồng cao hơn haplotype từ cá thể khoẻ mạnh.

Đã ra đời nhiều thuật toán gom nhóm, hầu hết chúng đều có điểm chung là sử dụng tiêu chuẩn đánh giá tương đồng giữa các mẫu. Trong hướng tiếp cận của chúng tôi, thuật toán DBSCAN được áp dụng để gom nhóm các haplotype theo tiêu chuẩn đánh giá tương đồng (phương pháp 1).

Có hai tham số sử dụng trong thuật toán DBSCAN. Một là bán kính ε của vùng lân cận giữa các haplotype quan tâm, hai là ngưỡng số phần tử tối thiểu *MinPts* xung quanh một haplotype đang xét. Một haplotype được gọi là haplotype lõi (core haplotype) nếu có số phần tử trong bán kính lân cận nhiều hơn *MinPts*. Các haplotype trong bán kính lận cận ε thì đạt được một cách trực tiếp từ haplotype lõi.

Định nghĩa 1: Một haplotype đến được từ một haplotype lõi khi và chỉ khi có một dây chuyền các haplotype lõi ở giữa hai haplotype đó.

Định nghĩa 2: Hai haplotype được gọi là density-connected khi và chỉ khi có một haplotype lõi mà nó đến được cả hai haplotype đó.

Định nghĩa 3: Một nhóm haplotype là tập hợp các haplotype density-connected với khả năng lan rộng cực đại.

Thuật toán:

- Bước đầu tiên chọn một haplotype bất kỳ làm haplotype lõi.
- Tiếp theo, tìm các haplotype trong vùng lân cận có thể đạt đến một cách trực tiếp từ haplotype lõi đó,
- Xác lập các nhóm khi cần thiết.
- Quá trình kết thúc khi tất các haplotype được kiểm tra.

Kết quả thu được là các nhóm thỏa điều kiện thống kê Z-score và loại bỏ các haplotype không thuộc bất kỳ nhóm nào.

Giả định rằng:

- Số haplotype từ cá thể bị bệnh là m
- Số haplotype từ cá thể khoẻ mạnh là n
- Số haplotype từ cá thể khoẻ mạnh trong một nhóm là m'
- Số haplotype từ cá thể bị bệnh trong một nhóm là n'

Ta có công thức Z-score **[5]** như sau:

$$Z = \frac{m'/m - n'/n}{\sqrt{\frac{m'+n'}{m+n}\left(1-\frac{m'+n'}{m+n}\right)\left(\frac{1}{m}+\frac{1}{n}\right)}}$$

Một giá trị Z-score lớn mang ý nghĩa độ kết hợp giữa tính trạng bệnh và nhóm haplotype là mạnh.

3.5. Xử lý các allele không xác định

Công việc này được thực hiện khi tập dữ liệu haplotype nguồn bao gồm các giá trị allele xác định và không xác định tại các vị trí marker đang xét, đây là nguyên nhân dẫn đến việc thống kê không chính xác. Một mẫu tiềm năng P được xem là phù hợp không chắc chắn với mẫu haplotype thứ i khi và chỉ khi có ít nhất một marker thứ j mà tại đó $D_{ij} = 0$ và $P_j \neq 0$. Thuật toán HPM **[4]** đã không xét đến trường hợp này.

Gọi π_{AU} là số mẫu phù hợp với bệnh không chắc chắn, và π_{CU} là số mẫu phù hợp khoẻ mạnh không chắc chắn. Khi đó tổng số haplotype mang bệnh phù hợp với mẫu tiềm năng P là

π_{AP} sẽ bằng tổng của π_{AP} và π_{AU}, tương tự cho tổng số haplotype khoẻ mạnh phù hợp với mẫu tiền năng P là π_{CP} sẽ bằng tổng π_{CP} và π_{CU}. Với hai giá trị này sẽ làm cho công thức thống kê χ^2 trong [4],[5] thay đổi.

3.6.Thuật toán trên hệ thống phân tán

Chúng tôi xây dựng thuật toán ClusterHPM trên N tiến trình xử lý được thực hiện như sau:

- Đọc tập tin dữ liệu haplotype, và tập tin tham số dùng trong thuật toán DBSCAN và HPM
- Thực hiện gom nhóm theo thuật toán DBSCAN
- Tiến trình chủ gửi dữ liệu từng nhóm cho các tiến trình con.
- Tiến trình con nhận dữ liệu, thực thi tìm mẫu phù hợp.
- Tiến trình chủ nhận dữ liệu từ các tiến trình con, tổng hợp dữ liệu.
- Tiến trình chủ thực hiện phép hoán vị ngẫu nhiên và tính p-value
- Tiến trình chủ ghi kết quả gồm các mẫu phù hợp và dự đoán vị trí gen mang mầm bệnh lên tập tin.

4. KẾT QUẢ THỬ NGHIỆM

Các kết quả thử nghiệm trên Cluster gồm 4 máy, cài đặt GT3, MPICH-G2 1.2.27 và Condor 6.7.

4.1. Tập dữ liệu thật thứ nhất: liên quan đến bệnh Friedreich Ataxia (FA - bệnh Thất điều – di truyền ở trẻ từ 8-12 tuổi, nguyên nhân là do suy hoá hệ thần kinh trung ương, hệ thần kinh ngoại biên và tim), được cung cấp từ [6], và được phân tích lại bởi Molitor [7], dữ liệu bao gồm 58 haplotype từ cá thể mắc bệnh, 69 haplotype từ cá thể khoẻ mạnh với 12 marker dạng microsatellite, tất cả dữ liệu được nhận từ quần thể Acadian. 12 marker dạng microsatellite chiếm một vùng dài 15cM với khoảng cách giữa các marker là: 3, 6.5, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, 0.125, và 4.5 cM. Kết quả chương trình dự đoán gen mang mầm bệnh ở giữa marker thứ 5 và marker thứ 6, với 3000 lần hoán vị, p-value tại marker 5 và 6 bằng 0.01 *(hình 2.a)*.

4.2.Tập dữ liệu thật thứ hai, liên quan đến bệnh cơ gan Cystic Fibrosis (CF- bệnh xơ nang) của Kerem (1989). Dữ liệu bao gồm 94 haplotype từ cá thể mang bệnh và 92 haplotype từ cá thể khoẻ mạnh trên 23 marker RFLP (Restriction Fragment Length Polymorphism). Khoảng cách giữa các marker là 0.009, 0.0158, 0.5, 0.01, 0.02, 0.015, 0.025, 0.02, 0.005, 0.035, 0.03, 0.025, 0.035, 0.035, 0.08, 0.01, 0.02, 0.015, 0.055, 0.6, 0.07 và 0.1 cM. Được biết vị trí đột biến trong khoảng 0,88 cM tính từ marker đầu tiên. Có 67% nhiễm sắc thể mang bệnh. Gen mang mầm bệnh được dự đoán điểm tại marker thứ 19 (0,9 cM tính từ marker đầu tiên) với tần số xuất hiện là 78, p-value = 0.0026. Kết quả này cho bởi 5000 lần hoán vị và không sử dụng gap trong mẫu tiềm năng *(hình 2.b)*.

4.3.Tập dữ liệu thật thứ ba do phòng thí nghiệm JDRF/WT (Juvenile Diabetes Research Foundation/Wellcome Trust) cung cấp, bao gồm 385 gia đình hạt nhân (sib-pair) mắc bệnh tiểu đường loại 1, gồm bố mẹ và 2 con. Có 25 marker loại microsatellite (D6S1641, D6S1548, D6S1576, D6S291, D6S439, D6S1629, D6S1568, D6S1560, D6S2445, D6S2444, HLA-DQB1, HLA-DRB1, D3A, 9N-1, D6S273, 82-1, TNFd, TNFe, TNFc, 62b, C1-2-A, C1-2-C, HLA-B, HLA-C, C1-4-4) trải dài trong vùng 14Mb trên nhiễm sắc thể thứ 6. Các mẫu haplotype được suy ra từ kiểu di truyền gia đình. Mỗi một gia đình, một haplotype từ bốn haplotype của bố mẹ được xem như haplotype mang bệnh (case haplotype) nếu nó xuất hiện trong bất kỳ đứa con mang bệnh nào. Ngược lại, các haplotype không truyền là haplotype khoẻ mạnh. Tổng cộng có 213 haplotype mang bệnh và 143 haplotype khoẻ mạnh. Kết quả chương trình dự đoán gen mang mầm bệnh ở tại marker 10, gen D6S2444. Với tần số xuất hiện tại

marker 10 là 6, p-value = 0.184667 của 3000 lần hoán vị, có sử dụng gap trong mẫu tiềm năng *(hình 3.a).*

4.4.Tập dữ liệu thật thứ tư, của nhóm nghiên cứu Mark J. Daly về bệnh Crohn, được công bố tại website http://www.broad.mit.edu/humgen/IDB5/ raw_data.txt. Dữ liệu haplotype được suy ra từ kiểu di truyền của 129 trẻ em Châu Âu trong mối quan hệ gia đình bộ ba gồm bố mẹ và con. Tập dữ liệu gồm 147 haplotype, trên 103 marker, với 20% allele không xác định. Kết quả dự đoán gen mang mầm bệnh với 3000 lần hoán vị ở gần marker thứ 74, p-value tại marker 74 bằng 0.019, tần số xuất hiện là 232 *(hình 3.b)*

Để minh họa thêm hiệu suất của thuật toán, ClusterHPM đã sử dụng tập dữ liệu mô phỏng trong thuật toán Haplotype Pattern Mining của Toivonen **[1]**. Một tập tin dữ liệu mô phỏng tương ứng với một quần thể được cô lập với một cặp nhiễm sắc thể tương đồng có chiều dài 100cM. Tổng số mẫu haplotype là 400, với 200 haplotype mang bệnh và 200 haplotype khoẻ mạnh. Khoảng cách của các marker dọc thể nhiễm sắc thể là 1cM đối với marker dạng microsatellite và bằng 1/3 cM đối với marker dạng SNP. Số lần hoán vị là 500, không có gap trong mẫu tiềm năng.

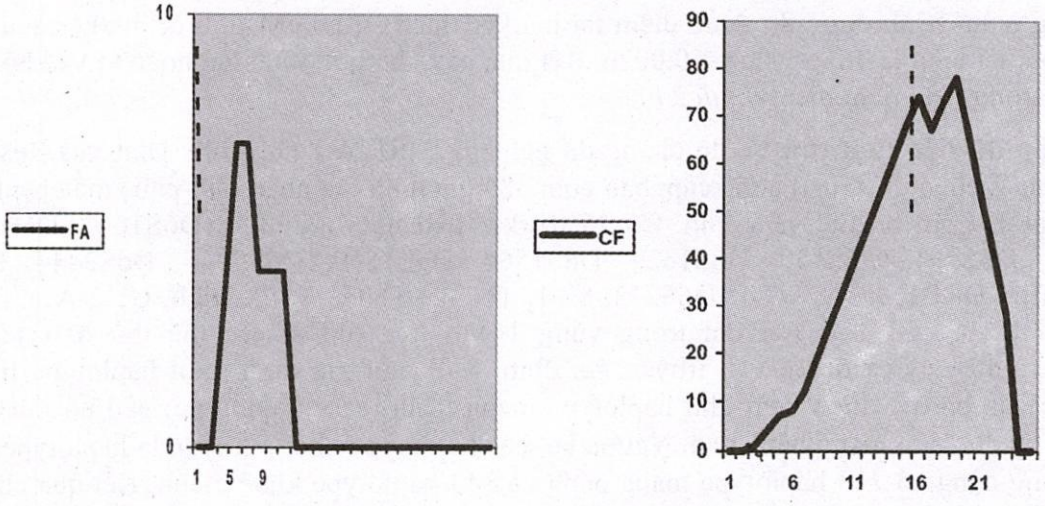
Bảng so sánh kết qủa như trong hình 4, và bảng so sánh thời gian thực thi của thuật toán HPM với thuật toán ClusterHPM như hình 5.

5. KẾT LUẬN

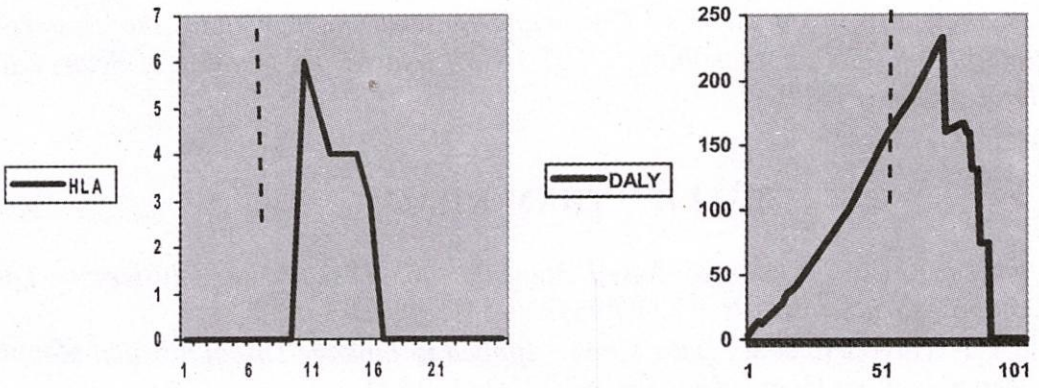
Ngành sinh tin học là một ngành mới và hấp dẫn rất nhiều nhà nghiên cứu trong và ngoài nước tham gia. Với các khám phá trong cấu trúc gen đã mở ra nhiều hướng nghiên cứu mới trong đó có y sinh học.

Bài toán xác định vị trí gen mang mầm bệnh là một trong bài toán của hướng y sinh học đặt ra. Trong báo cáo này, chúng tôi đề xuất sử dụng kỹ thuật khai phá dữ liệu trong bài toán lập bản đồ gen. Từ đó khám phá dữ liệu mẫu haplotype để giải quyết vấn đề xác định vị trí gen mang mầm bệnh. Với tập dữ liệu haplotype lớn, thực thi chương trình trên một máy với thuật toán đệ quy tuần tự, thời gian thực thi thu được rất lâu. Để cải tiến về tốc độ thực thi, trước tiên sẽ thanh lọc dữ liệu bằng cách tính mức độ tương đồng giữa các haplotype trong thuật toán gom nhóm, và loại bỏ những nhóm không có khả năng kết hợp mạnh với bệnh.

Với tập kết quả thu được từ thuật toán gom nhóm, dữ liệu sẽ được thực thi song song với thuật toán HPM. Tất cả chương trình sẽ thực thi trong môi trường tính toán lưới.. Kết quả tính toán không sai khác (hình 4), nhưng thời gian thực thi được cải tiến rất lớn (hình 5).



Hình 2. (a) Biểu đồ kết quả của tập dữ liệu FA. (b) Biểu đồ của tập dữ liệu kết quả CF.



Hình 3. (a) Biểu đồ của tập kết quả HLA. (b) Biểu đồ của tập kết quả Daly.

Bảng 1. Bảng so sánh kết quả giữa các chương trình

	FA	CF	HLA	Daly
HPM (2000) **[4]**	/	/	D6S2444	/
HapMiner (2005) **[5]**	M5	M18 (0.89cM)	D6S2444	/
HPM_*	M5-M6	/	D6S2444	/
ClusterHPM	M5-M6	M18-M19(0.9cM)	D6S2444	M74

Bảng 2. Bảng so sánh thời gian thực thi giữa thuật toán

	FA	CF	HLA	Daly
HPM_*	15 s	/	500 s	/
ClusterHPM	2.6 s	37 s	34 s	1391 s

Chú thích: Các ô (/) không được tác giả đề cập trong bài báo **[4]** và **[5]**. HPM_* là chương trình được hiện thực theo thuật toán HPM.

Bài báo này được thực hiện dưới sự hỗ trợ kinh phí từ đề tài "Tính toán hiệu năng cao và tính toán lưới trong một số bài toán tin sinh học" thuộc chương trình nghiên cứu cơ bản.

THE GENE MAPPING ALGORITHMS FOR SPECIFICATION THE LOCATIONS OF GENETIC DISEASES

Huynh Thi My Trang, Tran Van Lang
HCMC Institute of Information Technology

ABSTRACT: *The "Gene mapping" refers to mapping of genes to specific locations on chromosomes. It is a critical step in the understanding of genetic diseases. There are two types of gene mapping: genetic mapping – using linkage analysis to determine the relation between two genes on a chromosome; physical mapping – using all available techniques or information to determine the absolute position of a gene on a chromosome. In this paper we propose an approach to improve the efficiency of the algorithms using linkage analysis on gene mapping.*

We build the algorithms by using the Haplotype Pattern Mining (HPM) Algorithms and Density Based Spatial Clustering of Application with Noise Algorithm (DBSCAN). These algorithms are implemented on the Grid Computing System which includes the clusters of HCMC Institute of Information Technology (IOIT-HCM) and Korea Institute of Science and Technology Information (KISTI).

TÀI LIỆU THAM KHẢO

[1]. Martin Ester, et al., *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise,* Proc KDD'96, 226-231, 1996.

[2]. Hannu T.T. Toivonen, et al., *Data Mining Applied to Linkage Disequilibrium Mapping*, the American Society of Human Genetics, 67:133-145, 2000.

[3]. Hannu T.T. Toivonen, et. al , *Data Mining for Gene Mapping*, Next Generation of Data Mining Applications by Mehmed Kantardzic and Jozef Zurada (Eds.), Wiley-IEEE Press, 263-293., 2005

[4]. Hannu T.T. Toivonen, et al. , *Gene Mapping by Haplotype Pattern Mining*, IEEE International Symposium on Bio-Informatics and Biomedical Engineering, 99-108, 2000.

[5]. Jing Li, Tao Jiang , *Haplotype-based linkage disequilibrium mapping via direct data mining*, Department of Computer Science and Engineering - University of California at Riverside, Bioinformatics, 21:4384-4393, 2004.

[6]. Liu J.S., et al. , *Bayesian analysis of haplotypes for linkage disequilibrium mapping.* Genome.Res., 10:1716-1724., 2001.

[7]. Molitor J., Marjoram P., Thomas D., . *File-Scale Mapping of Disease Genes with Multiple Mutations via Spatial Clustering Techniques*, Am.J.Hum.Genet, 73:1368-1384, 2000.