BƯỚC ĐẦU ÁP DỤNG PHƯƠNG PHÁP PHÂN CỤM VÀO VIỆC HỖ TRỢ CHẨN ĐOÁN BỆNH UNG THƯ

Nguyễn Thị Tân Tiến*, Trương Thị Hồng Thúy
Trường Đại học Y Dược - ĐH Thái Nguyên

TÓM TẮT

Sự mất cân bằng về số lượng và cấu trúc của nhiễm sắc thể là một trong những đặc trưng nổi bật nhất của các tế bào khối u. Trong vài thập kỷ qua, có hàng ngàn các công trình nghiên cứu di truyền học tế bào về bệnh ung thư ở người đã tìm hiểu sâu bên trong cơ chế di chuyển của sự phát triển các khối u và phát hiện các mục tiêu cho sự can thiệp của dược liệu đối với các khối u. Người ta cho rằng các mẫu khác thường của nhiễm sắc thể lặp đi lặp lại phản ánh quan hệ hợp tác của đa số các khối u liên quan đến gen trong hầu hết các bệnh ác tính [4].

Một phương pháp được dùng để đo sự khác thường của bộ gen là so sánh các bộ gen lai CGH. CGH là phương pháp phân tích các phần tử di truyền học tế bào để phát hiện ra các gen mất cân bằng – gen dị tật (gen mất hoặc thêm đoạn DNA). Dữ liệu CGH của một khối u có thể được xem như là một danh sách có thứ tự các giá trị rời rạc, ở đó mỗi giá trị tương ứng với một nhóm nhiễm sắc thể đơn và đánh dấu bởi một trong 3 trạng thái (thêm, mất, không đổi) [5].

Từ khoá: *Dữ liệu Comparitive Genomic Hybridization, khai phá dữ liệu, phương pháp phân cụm, các thuật toán trong khai phá dữ liệu...*

GIỚI THIỆU

Ngày nay, Công nghệ thông tin đang phát triển như vũ bão và có ứng dụng vô cùng to lớn vào hầu hết các lĩnh vực trong đời sống. Việc ứng dụng công nghệ thông tin nói chung và khai phá dữ liệu nói riêng sẽ có tác dụng lớn trong hỗ trợ y học chẩn đoán các căn bệnh. Đã có hàng ngàn các công trình nghiên cứu di truyền học tế bào về bệnh ung thư ở người đã tìm hiểu sâu bên trong cơ chế di truyền của sự phát triển các khối u và phát hiện các mục tiêu cho sự can thiệp của dược liệu đối với các khối u. Dựa trên kết quả nghiên cứu thực nghiệm, người ta phát hiện ra rằng các bệnh nhân mắc bệnh ung thư như nhau thì mẫu quang sai của họ sẽ tương đồng với nhau [1].

Ở đây, tôi nghiên cứu phương pháp phân cụm, mục đích là nhóm các bệnh nhân có mẫu quang sai giống nhau vào cùng một cụm, từ đó có thể hỗ trợ trong việc chẩn đoán các bệnh nhân mắc bệnh ung thư. Phương pháp phân cụm được tiến hành trên các tập dữ liệu lai gen so sánh Comparative Genomic Hybridization của cơ sở dữ liệu Progenetix.(Cơ sở dữ liệu Progenetix là một nguồn dữ liệu khổng lồ về cấu trúc gen của các bệnh nhân ung thư). CGH là phương pháp phân tích các phần tử di truyền học tế bào để phát hiện ra các gen mất cân bằng – gen dị tật (nguyên nhân gây ung thư).

PHƯƠNG PHÁP

Việc phân cụm dữ liệu CGH để hỗ trợ cho việc chẩn đoán bệnh nhân ung thư ở người là một bài toán khá phức tạp. Những kết quả mà tôi trình bày trong bài báo này mới chỉ là những tìm hiểu ban đầu để làm cơ sở cho quá trình nghiên cứu tiếp theo, với hi vọng sau này có thể cải tiến được kết quả phân cụm dự đoán chính xác bài toán đặt ra.

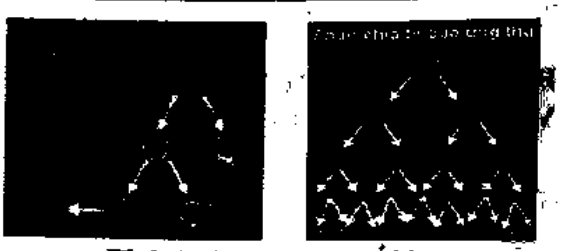
Trong phần này, tôi trình bày các kết quả tìm hiểu chính: tìm hiểu về cơ sở dữ liệu CGH và các bước tiền xử lý dữ liệu CGH thô thành dữ liệu rời rạc cho mục đích khai phá; phần tiếp theo tìm hiểu về phương pháp phân cụm với sự trợ giúp của các marker ; cuối cùng là trình bày kết quả chạy thực nghiệm thuật toán top - down sử dụng độ đo Rsim với sự trợ giúp của các marker. Các chương trình chạy thực nghiệm trên dữ liệu CGH được đưa ra, để so sánh hiệu quả khi sử dụng cùng phương pháp phân cụm nhưng trên hai độ đo khác nhau: độ đo Sim và độ đo Rsim.

* *Tel: 0988 905 105; Email. tantienyk@gmail.com*

Cơ sở dữ liệu CGH

Một trong những phương pháp đo quang sai gen là phương pháp lai gen so sánh CGH, đây là phương pháp được sử dụng để đo sự khác thường của bộ gen. CGH phân tích các phần tử di truyền học tế bào để phát hiện ra các gen mất cân bằng – gen dị tật (gen mất hoặc thêm đoạn DNA) trong bộ nhiễm sắc thể (phương pháp này không phân tích trường hợp mất cân bằng về số lượng nhiễm sắc thể). CGH là một phương pháp tiếp cận có hiệu quả để quét toàn bộ bộ gen cho các biến thể trong bản sao DNA. Lợi thế chính của dữ liệu CGH là số các bản sao DNA cho toàn bộ bộ gen có thể được đo trong các thí nghiệm đơn lẻ. CGH trên các mạch nhỏ (công nghệ chuỗi nhỏ) DNA phân tích phân tử di truyền học tế bào và có thể phát hiện đồng thời hàng nghìn gen mất cân bằng (thêm hoặc mất đoạn DNA). Trong công nghệ này, dữ liệu thô CGH được thể hiện như tỷ lệ các chất huỳnh quang được bình thường hóa của khối u và DNA liên quan.

Cơ thể con người được tạo nên từ các tế bào. Tế bào ở các cơ quan khác nhau của cơ thể có thể có hình dạng khác nhau, cách làm việc khác nhau, nhưng đều có chung cơ chế nhân bản. Tế bào cũng có tuổi thọ, già, rồi chết, và những tế bào mới được sinh ra để thay thế. Cơ thể có một cơ chế kiểm soát quy luật này một cách chặt chẽ và duy trì số lượng tế bào ở mỗi cơ quan, tổ chức ở mức ổn định. Bình thường, sự phân chia nhân bản tế bào được kiểm soát một cách có hệ thống. Trong quá trình phân chia tế bào, nhiễm sắc thể đơn thông thường sẽ được nhân đôi tạo ra một bản sao giống hệt với nó. Nếu bản sao giống hệt với nhiễm sắc thể gốc thì không có vấn đề gì, tuy nhiên trong quá trình phân ly tế bào vì một số lý do nào đó mà sự nhân đôi rồi lại phân ly của nhiễm sắc thể gây ra sự mất hay thêm đoạn DNA, chính là nguyên nhân gây ra các tế bào ung thư. Sự sai hỏng của DNA tạo nên các đột biến gen thiết yếu điều khiển quá trình phân bào cũng như các cơ chế quan trọng khác. Một hoặc nhiều đột biến được tích lũy gây ra sự tăng sinh không kiểm soát và tạo thành khối u



Hình 1: *Sự phân chia tế bào*

Dựa trên cơ chế nhân đôi của tế bào, trong kỹ thuật CGH người ta tiến hành lai (thực chất là ghép đôi để so sánh) bộ gen DNA của tế bào cần kiểm tra và bộ gen DNA của tế bào tham khảo. Nếu bộ gen của tế bào tham khảo là bình thường người ta sẽ kiểm định sự tăng giảm về kích thước của bộ DNA kiểm tra để phát hiện ra sự thiếu hụt hay thừa các đoạn DNA.

Cơ sở dữ liệu progenetix là nguồn tài nguyên quan trọng chứa dữ liệu CGH, là tập dữ liệu về di truyền học tế bào phân tử liên quan đến bệnh ung thư [14].

- Cơ sở dữ liệu này thống kê 21535 trường hợp mắc bệnh, 768 mô bệnh học về ung thư. Mỗi mẫu bao gồm một dãy dài các khoảng gen với các đoạn DNA bị thiếu hoặc bị thừa.
- Cơ sở dữ liệu CGH đưa ra cái nhìn tổng quan những dị thường về số lượng sao chép bộ gen liên quan đến bệnh ung thư ở người, có được từ kết quả thí nghiệm lai gen so sánh CGH. Cơ sở dữ liệu này được phát triển bởi nhóm của Michael Baudis, trường Đại học Zurich, Thụy Sĩ.
- Trong progenetix, CSDL CGH có số chiều là 862 (chính là số khoảng lượng khoảng gen). Các giá trị trong dữ liệu CGH là các giá trị rời rạc, rõ ràng (gồm 3 giá trị -1, 0, 1)

Phương pháp phân cụm sử dụng marker

Marker

Marker là một khoảng gen quan trọng đại diện cho một vùng biến đổi hồi quy. Mỗi marker được đại diện bởi 2 số <p, q>, trong đó p và q tương ứng biểu thị vị trí và loại quang sai.

Độ hỗ trợ

Cho S là một tập của N các mẫu CGH $\{S_1, S_2, \ldots, S_M\}$.

Cho x_d^j biểu thị giá trị đối với mẫu j tại khoảng gen d, $\forall d, 1 \le d \le D$, D là số lượng khoảng. Cho $s_j[u,v]$ là phân đoạn của s_j: bắt đầu từ khoảng thứ u và kết thúc ở khoảng thứ v. Cho $\{m_t = < p_t, q_t > | 1 \le t \le R\}$ là một tập các marker được đặt dọc theo các khoảng gen, Tức là: $p_1 < p_2 < \ldots\ldots < p_R$

Độ hỗ trợ của s_j tới m_t là $\sigma(s_j, m_t)$. $\sigma(s_j, m_t) = 1$ chỉ khi nó thỏa cả hai điều kiện sau đây:

1. Điều kiện cần: Tồn tại một phân đoạn $s_j[u,v]$ chồng chéo với m_t. Nghĩa là $u \le p_t \le v$ và kiểu của $s_j[u,v]$ giống như kiểu m_t, tức $x_u^j = q_t$

2. Điều kiện đủ: Không tồn tại marker $m_{t'}$, với $t'<t$, trong các nhiễm sắc thể giống nhau sao cho $u \le p_{t'} \le v$, $x_u^j = q_{t'}$ và $\sigma(s_j, m_{t'}) = 1$

Ngược lại, nếu không thỏa mãn 2 điều kiện trên thì $\sigma(s_j, m_t) = 0$.

Ta nói s_j hỗ trợ m_t hay m_t bao s_j nếu $\sigma(s_j, m_t) = 1$.

Như vậy:

- Điều kiện cần:

– Giá trị hỗ trợ của 1 marker đếm số các mẫu có tình trạng quang sai giống nhau và giống marker đó.

– Giá trị hỗ trợ của 1 marker lớn, điều đó tương đương marker đó quan trọng, có thể đặc trưng/ đại diện cho mô hình quang sai của mẫu.

– Một mẫu hỗ trợ một marker khi và chỉ khi có một quang sai giống marker tại vị trí xác định.

- Điều kiện đủ:

– Quang sai trong cùng một phân đoạn có thể tương ứng với một quang sai đơn mà nó trải ra tới khoảng gen lân cận 10.

– Phân đoạn chồng chéo/trùng với bội số marker của quang sai cùng loại để chỉ hỗ trợ một trong những marker đó.

Độ đo Rsim

Cho D biểu thị số khoảng gen của mỗi mẫu

Cho $s_i = x_1^i, x_2^i, \ldots, x_D^i$ và $s_j = x_1^j, x_2^j, \ldots, x_D^j$ là 2 mẫu CGH. Trong đó, x_d^i và x_d^j tương ứng là giá trị hoặc trạng thái của khoảng gen d của s_i, s_j

Rsim là phép đo sự chồng chéo giữa các phân đoạn giao với một điểm marker nào đó mà $x_d^i = x_d^j$.

Sử dụng marker để phân cụm

Thuật toán dựa trên từng cặp tương tự phân các mẫu thành các cụm sao cho sự giống nhau giữa các mẫu trong cùng một cụm lớn hơn sự giống nhau của mẫu với các cụm khác. Điều này thường đòi hỏi tính toán độ tương tự giữa hai mẫu.

Ta có thể sử dụng marker để loại bỏ sự ảnh hưởng của nhiễu đến từng cặp giống nhau. Từ độ đo Sim, phát triển một độ đo mới gọi là Rsim , cụ thể như sau:

Cho $M = \{m_1, m_2, \ldots, m_R\}$, $p_1 < p_2 < \ldots < p_R$ là tập các marker được xác định trên tất cả các mẫu. Các marker là các khoảng gen quan trọng có liên quan tới các mô hình quang sai của các mẫu.

Cho D là số lượng khoảng gen của mỗi mẫu. Cho 2 mẫu CGH s_i, s_j, với $s_i = x_1^i, x_2^i, \ldots, x_D^i$; $s_j = x_1^j, x_2^j, \ldots, x_D^j$.

Trong đó x_d^i, x_d^j tương ứng là giá trị hay trạng thái của s_i, s_j.

Rsim đo độ tương tự giữa chúng như là số lượng của cặp phân đoạn chồng chéo, như vậy các phân đoạn này phải thỏa mãn 2 điều kiện sau:

1. Ít nhất một trong marker của M chứa cả hai phân đoạn.

2. Kiểu của marker giống kiểu quang sai của cả hai phân đoạn chồng chéo đó.

Cho $x_u^i, \ldots, x_v^i$ và $x_{u'}^j, \ldots, x_{v'}^j$ là 1 cặp tương ứng từ mẫu si và sj. RSim đếm cặp của phân đoạn là một nếu:

1. Tồn tại marker $m_t = \langle p_t, q_t \rangle m_t \in M, u \le p_t \le v; u' \le p_t \le v'$

2. Loại quang sai của cả 2 phân đoạn giống nhau như của mt, tức là $x_u^i = x_{u'}^j = q_t$

KẾT QUẢ THỰC NGHIỆM

kết quả chạy thực nghiệm thuật toán top-down với độ đo Rsim

Bài toán

Input: Bộ dữ liệu Carsinomas với 23 thực thể lâm sàng, với hơn 6000 mẫu dữ liệu

Output: Hình ảnh phân thành các cụm với số lượng marker được chỉ định trước.

Cài đặt thực nghiệm

Dữ liệu và công cụ

Dữ liệu của bài toán là file có dạng *.txt, được lấy tại cơ sở dữ liệu progenetix 8 (trang web http://www.progenetix.net). Đây là bộ dữ liệu gồm 31925 mẫu CGH. Mỗi mẫu có 862 khoảng gen được lấy ra từ 24 nhiễm sắc thể.

Phần mềm chạy chương trình: Matlab

Chạy chương trình thực nghiệm

Trong công cụ có chạy hàm PhyloMain như sau:

PhyloMain (dataset, outfile, nMarkers, nMkperClust, MaxnClust, MinCases)

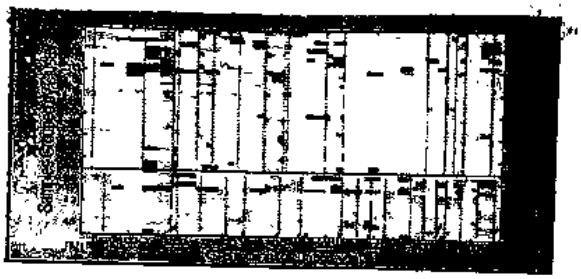
- Cách sử dụng:

Sử dụng phương pháp phân cụm với độ đo RSim để chia từng thực thể bệnh ung thư vào 1, 2, 4 cụm, ... (Nếu số lượng các cụm là 1, nghĩa là không phân cụm). Ở đây số lượng các cụm phụ thuộc tham số MinCases trong PhyloMain.m. Trong mỗi trường hợp mà số cụm khác nhau thì sẽ sử dụng chương trình nhận dạng các marker để xác định một tập hợp các marker cho mỗi cụm. Số lượng các marker được thiết lập bởi tham số nMarkers trong PhyloMain.m.

- Kết quả:

Sau thời gian 3h30 chạy thuật toán, với hàm các tham số trong PhyloMain lần lượt là 'carcinomas_informativematrix.txt', 'outfile1.txt', 20, 10, 8, 50, thu được kết quả là hình ảnh sau

Hình 2 - Hình ảnh phân cụm trên tập mẫu thực nghiệm, có hai chiều: một chiều nằm ngang biểu thị 862 khoảng gen, một chiều thẳng đứng biểu thị 45 mẫu bệnh. Hình ảnh được minh họa bằng các màu: xanh, đỏ. Màu xanh biểu thị trạng thái gia tăng (thừa) số lượng sao chép DNA, màu đỏ biểu thị trạng thái thiếu (mất mát) số lượng sao chép DNA. Các marker là các đường kẻ dọc.



Hình 2: *Hình ảnh phân cụm trên tập mẫu thực nghiệm*

Đánh giá kết quả thực nghiệm

So sánh số marker trong một cụm

Sử dụng số lượng marker khác nhau 10, 20, 40, 60 và 80 các marker cho thuật toán phân cụm top - down, sử dụng độ đo Rsim sẽ cho chất lượng cụm khác nhau. Ở đây, nhập số lượng marker, giá trị chất lượng cụm là giá trị trung bình trong 20 lần chạy thực nghiệm, cụ thể:

Bảng 1: *Chất lượng cụm khi thay đổi số marker*

Số lượng marker	Chất lượng cụm	Thời gian
10	455.6411	3
20	510.6452	3h30
40	624.3025	4h
60	628.7182	5h
80	737.851	9h

Quan sát bảng 1 ta đưa ra được nhận xét sau: số lượng marker khác nhau thì chất lượng cụm thu được cũng khác nhau. Khi số lượng marker càng nhiều thì chất lượng cũng tốt hơn nhưng lại mất nhiều thời gian hơn, nếu 60 marker chất lượng là ~ 629 hơn phân cụm dùng 40 marker là ~634 nhưng lại tốn thời gian hơn 1h.

KẾT LUẬN

Phân cụm dữ liệu là một phương pháp phổ biến, đơn giản và hiệu quả trong khai phá dữ liệu. Các phương pháp phân cụm được áp dụng rộng rãi trong rất nhiều bài toán thực tế, nhất là trong lĩnh vực tin-sinh học. Trong bài báo này, tôi đã giới thiệu về một loại dữ liệu

sinh học CGH được sử dụng trong việc phân tích phân tử gen di truyền của bệnh nhân ung thư cho mục đích nghiên cứu. Bên cạnh đó bài báo đã trình bày phương pháp phân cụm sử dụng các marker trên thuật toán topdown. Hiện tại tôi đang tiếp tục nghiên cứu cải tiến các marker, nghiên cứu phương pháp phân cụm trên các định dạng CGH khác như dữ liệu array CGH, và tập hợp dữ liệu khác như mảng dữ liệu biểu hiện gen, dữ liệu SNP và dữ liệu protein.

TÀI LIỆU THAM KHẢO

1. Anil K.Jain, Richard C. Dubes (1988), *"Algorithms for clustering data"* Michigan State Univ East Lansing

A Fritz, C Percy, A. Jack, L Sobin and M. Parkin, editors, (2000) International Classification of Diseases for Oncology (ICD-O), Third Edition World Health Organization, Geneva

2 C.Rouvenol, N Stranskym P Hup, P. La Rosa, E Viara, E Barillot, and F Radvanyi, (January 2006) *Computation of recurrent minial genomic alterations from array - cgh data* Bioinformatics

3 Jun Liu, (2008), *mining comparative genomic hybridization data*, University of Florida

4 Jun Liu, S Ranka, and T Kahveci, (2007) Markers improve clustering of CGH data Bioinformatics

5 Jun Liu, J Mohammed [illegible] Kahveci, and M Baudis, (2[illegible] clustering of CGH data Bi[illegible]

6. Jiawei Han, Micheline K[illegible] mining concepts and techniques

[illegible] M Baudis, (March 2006) O[illegible] bioinformatics toolbox to support data [illegible] cancer cytogenetics Biotechniques 40(3)

8 M Baudis and M L Cleary, (200[illegible]) Progenetix net: an online repository for molecular cytogenetic aberration data. Bioinformatics.

9 P. Berkhin (2001), Survey of Clustering Data Mining Techniques Research paper Accrue Software, Inc. http www accrue com

10 P.-N Tan, M Steinbach, and V Kumar (2005) Introduction to Data Mining Addison-Wesley Longman Publishing Co . Inc

11 S Solinas-Toldo, S Lampel, S Stilgenbauer J Nickolenko, A Benner, H Dohner, T Cre-mer and P Lichter, (1997) Matrix-based comparative genomic hybridization biochips to screen for genomic imbalances Genes Chromosomes Cancer.

12. Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth, (2002) *"From Data Mining To Discory Knowledge in Database"*

13 http progenetix net index shtml

14 http en wikipedia org wiki Cluster_analysis

15 http. en wikipedia org wiki Comparative_genomic_hybridization

SUMMARY

INITIALLY APPLIED CLUSTERING METHODS IN ASSISTING CANCER DIANOSIS

Nguyen Thi Tan Tien*, Truong Thi Hong Thuy

College of Medicine and Pharmacy – TNU

Numerical and structural chromosomal imbalances are one of the most prominent features of neoplastic cells. Thousands of (molecular-) cytogenetic studies of human neoplasias have searched for insights into genetic mechanisms of tumor development and the detection of targets for pharmacologic intervention. It is assumed that repetitive chromosomal aberration patterns reflect the supposed cooperation of a multitude of tumor relevant genes in most malignant diseases[4]. One method for measuring genomic aberrations is Comparative Genomic Hybridization (CGH). CGH is a molecular-cytogenetic analysis method for detecting regions with genomic imbalances (gains or losses of DNA segments). CGH data of an individual tumor can be considered as an ordered list of discrete values, where each value corresponds to a single chromosomal band and denotes one of three aberration statuses (gain, loss and no change) [5]

Key words: *Data Comparative Genomic Hybridization data mining clustering Algorithms for clustering data*

Ngày nhận bài 12 11 2015 Ngày phản biện 30 12 2015 Ngày duyệt đăng 30 5 2016

Phản biện khoa học: *TS Hạc Văn Vinh – Trường Đại học Y Dược - ĐHTN*

* *Tel [illegible] Email: tantienvk@gmail.com*