TỔNG QUAN VỀ PHÂN CỤM DỮ LIỆU TRONG TIN SINH HỌC

Nguyễn Đức Tùng, Nguyễn Thị Lệ Thủy, Bùi Lê Thanh Nhàn
Khoa Cơ bản – Trường Đại học Y Dược Huế

Tóm tắt

Các phương pháp thực nghiệm phân tích hệ gen được quan tâm nghiên cứu trong những năm gần đây và đã thu được nhiều thành tựu quan trọng. Tuy nhiên, phần lớn các trình tự bộ gen hoàn chỉnh hiện nay có ít nhất một nửa số gen có chú thích không rõ ràng. Trong khi đó, cùng với sự bùng nổ của dữ liệu, một số xu hướng nghiên cứu mới đã xuất hiện trong tin học nhằm phân cụm để xử lý thông tin. Trong bài báo này, chúng tôi trình bày tổng quan về phân cụm dữ liệu ứng dụng cho phân tích protein, một bước khảo sát ban đầu rất có ý nghĩa đối với nghiên cứu thực nghiệm phân tích hệ gen, giúp giảm thiểu số lượng thí nghiệm nhận biết và từng bước hoàn thiện chức năng Protein.

Từ khoá: *Phân cụm dữ liệu, tin sinh học, dự đoán chức năng Protein.*

Abstract

AN INTRODUCTION TO DATA CLUSTERING IN BIOINFORMATICS

Nguyen Duc Tung, Nguyen Thi Le Thuy, Bui Le Thanh Nhan
Dept. of Basic Sciences, Hue University of Medicine and Pharmacy

Experimental methods for genome analysis are of crucial interest and have recently made a considerable progress. However, most complete orders of genomes have at least a half the number of gens with unexplicit note. There are some new trends of research in infomatics to deal with data clustering and treating. In this article, we introduce general data clustering and its application in Protein analysis, an initial step which is highly significant for the experimental study of gemone analysis. This method helps to reduce to number of prediction experiment and to perfect Protein function.

Keywords: *Cluster analysis, bioinformatics, Protein function prediction.*

1. MỞ ĐẦU

Rất nhiều nghiên cứu về trình tự gen tạo ra sự phát triển khổng lồ về cơ sở dữ liệu Protein. Những chú thích bằng tay các trình tự tìm được trong cơ sở dữ liệu thường rất đắt và khá bất tiện. Chính từ đó xuất hiện nhu cầu phát triển các thuật toán tin cậy để tự động hóa quá trình phân loại những trình tự này và nhận biết các họ Protein khác nhau. Hầu hết các phương pháp được sử dụng trong thực tế gần đây thực hiện các mối quan hệ tiến hóa giữa các chuỗi để dự đoán các đặc trưng về chức năng. Để thực hiện phân cụm Protein, trước hết chúng ta thu thập thông tin về các Protein từ các cơ sở dữ liệu, và tiến hành một phép đo thích hợp khoảng cách giữa hai chuỗi Protein để từ đó thu được một ma trận khoảng cách cho thuật toán phân cụm. Các thuật toán khác nhau sẽ cho kết quả không hoàn toàn như nhau tùy thuộc vào những ưu khuyết điểm của từng phương pháp. Chúng tôi sẽ giới thiệu khái quát về phân cụm dữ liệu trong tin học, một số hướng nghiên cứu của tin sinh học và ứng dụng của phân cụm dữ liệu Protein cũng như vai trò của nó trong dự đoán chức năng của Protein.

2. PHÂN CỤM DỮ LIỆU TRONG TIN HỌC

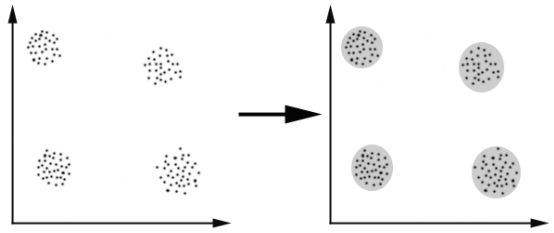
2.1. Khái niệm phân cụm dữ liệu

Phân cụm là một kỹ thuật quan trọng phân chia dữ liệu thành các nhóm đối tượng tương tự nhau. Mỗi nhóm (cụm) bao gồm các đối tượng mà các

 DOI: 10.34071/jmp.2013.1.11

đối tượng này là tương tự nhau trong cùng một nhóm và không tương tự với các đối tượng của các nhóm khác. Mục tiêu của phân cụm thực chất là gom các đối tượng dữ liệu thành từng nhóm [1].

Hình 1.1 là một ví dụ về phân cụm trong đó chúng ta dễ dàng xác định được 4 cụm dữ liệu, tiêu chí "tương tự" ở đây là khoảng cách: hai hoặc nhiều đối tượng thuộc cùng cụm nếu nó là "gần gũi" nhau, theo một khoảng cách nhất định (trong trường hợp này là khoảng cách hình học). Đây được gọi là phân cụm dựa trên khoảng cách.



Hình 1.1. Ví dụ phân cụm dữ liệu.

Ngoài ra, còn có một số định nghĩa khác về phân cụm như "Cụm là một tập các điểm trong không gian mà khoảng cách giữa hai điểm bất kì trong nó luôn nhỏ hơn khoảng cách giữa một điểm bất kỳ bên trong nó và một điểm bên ngoài". Hai hoặc nhiều đối tượng được gọi là cùng một cụm nếu nó được định nghĩa cùng một khái niệm cho tất cả đối tượng. Nói cách khác, các đối tượng được nhóm lại để phù hợp với các khái niệm mô tả.

2.2. Các giai đoạn của phân cụm

Các hoạt động tiêu biểu của phân cụm các mẫu gồm các bước sau [2]:

- Bước 1: Biểu diễn mẫu (bao gồm chọn lựa và hay hoặc trích rút các đặc trưng) liên quan đến số lượng các phân lớp, số lượng các mẫu có ý nghĩa, và số lượng, kiểu, phạm vi của các tính năng có ý nghĩa cho thuật toán phân cụm.

- Bước 2: Định nghĩa một thước đo, sự "gần gũi" của các mẫu phù hợp với miền dữ liệu, thường là khoảng cách, chẳng hạn khoảng cách Euclide, hay Czekanowski-Dice,…

- Bước 3: Phân cụm hoặc phân nhóm, có thể được thực hiện theo một số phương pháp khác nhau.

- Bước 4: Trừu tượng hóa dữ liệu (nếu cần), là quá trình trích rút sự biểu diễn đơn giản và nhỏ gọn của một tập dữ liệu.

- Bước 5: Đánh giá đầu ra (nếu cần), và thuật toán phân cụm là "tốt" hay "nghèo".

2.3. Phân loại kỹ thuật phân cụm

Có nhiều phương pháp tiếp cận khác nhau để phân cụm dữ liệu [2],

- Vun đống hay phân chia: yếu tố này liên quan đến cấu trúc và hoạt động của thuật toán. Cách tiếp cận vun đống bắt đầu với mỗi mẫu thuộc một cụm riêng biệt (duy nhất), liên tục sát nhập các cụm lại với nhau và dừng khi thỏa mãn tiêu chí hoặc chỉ còn một cụm duy nhất. Phương thức phân chia bắt đầu với tất cả các mẫu nằm trong cùng một cụm và thực hiện chia tách cho đến khi thỏa mãn tiêu chí dừng, ngăn chặn.

- ***Đơn nguyên tắc*** (monothetic) ***hay đa nguyên tắc*** (polythetic): yếu tố này liên quan đến việc sử dụng tuần tự hoặc đồng thời các đặc trưng trong quá trình phân cụm. Hầu hết các thuật toán là đa nguyên tắc; nghĩa là, tất cả các đặc trưng đều tham gia vào việc tính toán các khoảng cách giữa các mẫu, và sự quyết định dựa trên những các khoảng cách đó.

- ***Cứng*** (hard) ***hay mờ*** (fuzzy): một thuật toán phân cụm cứng phân chia từng mẫu đến một cụm duy nhất trong thời gian thực hiện và lặp của nó. Phân cụm mờ gán độ đo thành viên của mỗi mẫu đầu vào trong vài cụm. Tùy thuộc vào giá trị độ đo thành viên để quyết định mẫu sẽ thuộc vào phân cụm nào. Một phân cụm mờ có thể được chuyển thành phân cụm cứng bằng cách phân định mỗi mẫu đến một phân cụm với việc gán giá trị độ đo thành viên là lớn nhất.

- ***Xác định*** (deterministic) ***hay ngẫu nhiên*** (stochastic): phù hợp nhất với cách tiếp cận phân vùng, được thiết kế để tối ưu hóa hàm lỗi bình phương, bằng cách sử dụng các kỹ thuật truyền thống hoặc thông qua quá trình tìm kiếm ngẫu nhiên trong không gian trạng thái gồm tất cả các nhãn có thể có của nó.

- ***Gia tăng*** (incremental) ***hay bất gia tăng*** (non-incremental): phát sinh khi các mẫu được thiết lập bởi sự phân cụm lớn và ràng buộc về thời gian thực hiện hoặc không gian bộ nhớ ảnh hưởng đến kiến trúc của thuật toán. Ban đầu, có ít thuật toán phân cụm để thao tác trên các tập dữ liệu lớn, sau đó, sự ra đời của khai phá dữ liệu đã thúc đẩy

sự phát triển của các thuật toán phân cụm này để giảm thiểu số lượng các lần duyệt các tập mẫu, giảm số lượng các mẫu kiểm tra trong quá trình thực hiện hoặc giảm kích thước của cấu trúc dữ liệu được sử dụng trong khi thuật toán hoạt động.

2.4. Một số thuật toán phân cụm

- Phân cụm phân cấp (Hierarchical clustering)

Thuật toán dựa trên sự hợp nhất giữa hai cụm gần nhất. Ban đầu, thuật toán xem mỗi mẫu là một cụm, sau một số lần lặp nó đạt đến các cụm cuối cùng theo mong muốn. Phân cụm phân cấp thường tạo một cây các cụm phân cấp, được gọi là *dendrogram* [1]. Các lá của cây biểu diễn các đối tượng riêng lẽ. Các nút trong của cây biểu diễn các cụm. Dendrogram có thể bị cắt ở các cấp độ khác nhau để có thể tạo ra các phân cụm dữ liệu khác nhau. Tiêu biểu của phân cụm phân cấp là thuật toán liên kết đơn (single-link), liên kết đầy đủ (complete-link), và phương sai nhỏ nhất.

- Phân cụm phân hoạch (Partitional clustering)

Phân cụm phân hoạch cho kết quả là các phân vùng tách biệt của dữ liệu thay vì một cấu trúc phân cấp (chẳng hạn như dendrogram được tạo ra bởi một kỹ thuật phân cấp). Phương pháp này có ý nghĩa trong các ứng dụng liên quan đến bộ dữ liệu lớn trong đó các dendrogram không được phép xây dựng. Lựa chọn số các phân cụm đầu ra mong muốn là thao tác quan trọng khi sử dụng các thuật toán này. K - mean là thuật toán thường được sử dụng thuộc nhóm này [2].

- Thuật toán phân cụm kiểu pha trộn (Mixture-resolving and Mode-seeking)

Thuật toán này được phân tích cụm dựa trên giả thiết cơ bản là các mẫu phân cụm được rút ra từ một trong một số loại phân phối xác suất, và mục tiêu là xác định các tham số và giá trị của các tham số đó. Hầu hết các thuật toán trong cách tiếp cận này đều sử dụng mật độ trộn các thành phần cá thể (individual) là phân phối Gaussian trong đó các tham số của Gaussians sẽ được ước tính.

- Phân cụm láng giềng gần nhất (Nearest neighbor clustering)

Khoảng cách láng giềng gần nhất có thể được dùng làm cơ sở cho phân cụm. Thuật toán này gán nhãn cho các mẫu không có nhãn từ nhãn của mẫu láng giềng gần nhất cho đến khi tất cả các mẫu đều có nhãn hoặc không có thêm sự gán nhãn xảy ra.

- Phân cụm mờ (Fuzzy Clustering)

Phương pháp phân cụm theo cách tiếp cận truyền thống tạo các phân vùng, trong một phân vùng, mỗi mẫu thuộc về một và chỉ một cụm nào đó. Vì vậy, các mẫu trong cụm là phân chia cứng. Thuật toán mờ mở rộng khái niệm này để các mẫu liên kết với mỗi cụm, sử dụng hàm thành viên. Đầu ra của thuật toán là một phân cụm thay vì là một phân vùng.

Ngoài các kỹ thuật phân cụm còn có nhiều thuật toán khác như sử dụng mạng Neural, phân cụm dựa trên lưới, ...

2.5. Ứng dụng của phân cụm

Thuật toán phân cụm đã được ứng dụng lớn trong nhiều lĩnh vực khác nhau như:

- Tiếp thị: tìm kiếm các nhóm khách hàng với các hành vi tương tự trong một dữ liệu khách hàng lớn bao gồm các thuộc tính và hồ sơ mua trong quá khứ,…

- Sinh học: phân loại thực vật và động vật theo các tính năng của nó, xây dựng cây phát sinh loài, dự đoán tương tác hoặc chức năng cấu trúc của protein.

- Thư viện: phân loại sách, tài liệu, văn bản,…

- Bảo hiểm: xác định nhóm chủ sở hữu bảo hiểm với một yêu cầu bồi thường chi phí trung bình là cao; xác định gian lận.

- Lập kế hoạch thành phố: xác định cụm nhà ở theo kiểu nhà, giá trị và vị trí địa lý.

- Nghiên cứu động đất: phân cụm quan sát tâm chấn của động đất để xác định khu vực nguy hiểm.

- WWW: phân loại tài liệu; dữ liệu phân cụm weblog để khám phá các nhóm có cùng kiểu truy cập tương tự.

- Phân đoạn hình ảnh: các phân đoạn của hình ảnh được biểu diễn cho một hệ thống phân tích hình phụ thuộc nhiều vào phương của cảnh, hình dạng ảnh, cấu hình; sau đó sử dụng bộ cảm biến để chuyển đổi cảnh vào ảnh kỹ thuật số; và cuối cùng là mục tiêu mong muốn của hệ thống.

Ngoài ra phân cụm còn có nhiều ứng dụng khác như nhận dạng đối tượng chuyển động, chữ viết tay, truy vấn thông tin, khai phá dữ liệu hoặc phát hiện tri thức,…

3. TIN SINH HỌC

3.1. Khái niệm

Công nghệ sinh học ngày nay rất phát triển và đã tạo ra một khối lượng dữ liệu khổng lồ, bởi vậy phân tích dữ liệu bằng tay là điều khó có thể thực hiện được. Do đó việc kết hợp các khoa học khác như toán học, thống kê, thuật toán và khoa học máy tính vào công nghệ sinh học là rất cần thiết. Tin sinh học (bioinformatics), là một ngành khoa học kết hợp giữa các ngành khoa học là tin học, toán học và công nghệ sinh học, ra đời nhằm giải quyết vấn đề này. Tin sinh học đôi khi còn được gọi là *sinh học tính toán* (computational biology). Tuy nhiên tin sinh học thiên về phát triển các giải thuật, lý thuyết và các kỹ thuật thống kê và tính toán để giải quyết các bài toán bắt nguồn từ nhu cầu quản lý và phân tích dữ liệu sinh học. Trong khi đó, sinh học tính toán thiên về kiểm định các giả thiết đặt ra trong sinh học và nhờ máy tính thực nghiệm trên dữ liệu mô phỏng như dự đoán mối quan hệ tương tác giữa các protein, dự đoán cấu trúc bậc 2 của protein,…

Mối quan tâm chính của tin sinh học và sinh học tính toán là việc sử dụng các công cụ toán học để trích rút các thông tin hữu ích từ các dữ liệu hỗn độn được thu thập từ các kỹ thuật sinh học với lưu lượng mức độ lớn.

3.2. Các nhiệm vụ cơ bản của tin sinh học

- Xây dựng, bổ sung, tổ chức quản lý, khai thác cơ sở dữ liệu đa dạng, toàn diện trên quy mô toàn cầu liên quan đến sinh học và lĩnh vực khoa học liên quan.

- Xây dựng và phát triển các chương trình xử lý dữ liệu ứng dụng, dưới dạng các chương trình xử lý dữ liệu độc lập hay tích hợp ngay trong các thiết bị phân tích hiện đại.

- Đào tạo và cập nhật thường xuyên cho các nhà sinh học có kỹ năng tư duy và năng lực khai thác hai nội dung trên vào hoạt động khoa học và công nghệ tạo ra bước chuyển biến đột phá trong phương pháp tiếp cận và nghiên cứu khám phá thế giới sống.

3.3. Các lĩnh vực nghiên cứu chính của tin sinh học

- Hệ gen học (genomics) gồm phân tích trình tự của DNA để tìm gen cấu trúc hay quy luật của những trình tự protein tương đồng, và chỉ định gen hay dò tìm đột biến gen.

- Sinh học tiến hoá gồm có phân loại phân tử nhằm theo dõi sự tiến hoá của các loài dựa trên những thay đổi trong trình tự DNA, và bảo tồn đa dạng sinh học

- Phân tích chức năng gen gồm có mức độ thể hiện gen, nhận diện protein và dự đoán cấu trúc protein.

Chúng tôi sẽ trình bày cụ thể trong phần tiếp theo một số hướng nghiên cứu liên quan có ứng dụng cộng cụ thuật toán phân cụm.

4. THUẬT TOÁN PHÂN CỤM VÀ DỰ ĐOÁN CHỨC NĂNG PROTEIN

4.1. Một số thuật toán phân cụm Protein

- Thuật toán kết nối láng giềng (NJ, viết tắt của Neighbor-joining) [3]: là một thuật toán tái xây dựng cây phát sinh loài từ dữ liệu khoảng cách tiến hóa, và tính toán độ dài của các nhánh trong cây. Thuật toán này dựa trên lược đồ vun đống, bắt đầu với một cây hình sao và lặp đi lặp lại việc chọn cặp đơn vị phân loại hoạt động OTU (operational taxonomic unit) sao cho tổng chiều dài của nhánh bắt đầu từ các OTU ở từng giai đoạn phân cụm là nhỏ nhất, đồng thời rút gọn ma trận khoảng cách bằng cách thay thế các đơn vị phân loại được chọn bởi một nút mới. Độ dài các nhánh cũng như topo của cây có thể nhanh chóng thu được bằng cách sử dụng thuật toán này.

Thuật toán NJ có dữ liệu vào là ma trận khoảng cách (D_{ij}) có kích thước n x n, với n là số đơn vị phân loại, dữ liệu ra là cây cộng và khoảng cách của các nhánh trong cây. Thuật toán đầu tiên gồm 4 bước trong đó bước đầu tiên là nhập thông tin ma trận khoảng cách D_{ij}, tiếp theo tính tổng độ dài S_{ij} các nhánh giữa hai OUT i và j. Ở bước thứ ba, một nút mới X được thêm vào rồi xác định khoảng cách giữa các nút X và phần nút còn lại, và nhập các khoảng cách đó vào ma trận khoảng cách. Loại bỏ các nút 1 và 2 từ ma trận khoảng cách. Đồng thời tính toán chiều dài cho các nhánh đã được tham gia, đây là những nhánh 1-X và 2-X. Bước cuối cùng là quá trình lặp đi lặp lại từ bước 2 - một lần nữa tìm 2 nút gần nhất, và tiếp tục làm như vậy cho đến khi cây chỉ còn 2 nút thì thu được một cây phân cấp và độ dài các nhánh của nó. Thuật toán bảo đảm về độ tin cậy của các ước tính độ dài nhánh, tuy nhiên độ phức tạp của thuật toán khá lớn, $O(n^5)$.

Để giảm độ phức tạp cho thuật toán xuống còn $O(n^3)$, Studier & Keppler [4, 5] đã điều chỉnh thuật toán NJ trong đó các tham số được sử dụng để chọn hai nút láng giềng có tổng độ dài nhánh nhỏ nhất, S_{ij}. Studier và Keppler đã cung cấp một tham số thay thế. Tham số này, có tên là M_{ij}, thực sự là một chuyển đổi của *Sij*. *Mij* làm giảm độ phức tạp còn $O(n^3)$. Đồng thời, Studier và Keppler chứng minh rằng S và M là các tiêu chí liên quan: giảm thiểu S cũng giảm thiểu M và ngược lại ([4]).

- Thuật toán BIONJ là một phiên bản cải tiến của thuật toán NJ bằng cách xem xét lại một số công thức của NJ do có tính đến yếu tố sinh học. Giống như NJ, thuật toán này cũng sử dụng phân cụm theo kiểu vun đống, bao gồm sự lặp đi lặp lại việc chọn một cặp đơn vị phân loại, tạo ra một nút mới đại diện cho cụm các đơn vị phân loại này, và giảm ma trận khoảng cách dần dần. Tuy nhiên, BIONJ sử dụng mô hình đơn bậc nhất (simple first-order model) của phương sai và hiệp phương sai để ước lượng khoảng cách tiến hóa. Tại mỗi bước nó cho phép chọn lựa, từ các lớp rút gọn chấp nhận được, rút gọn làm tối thiểu phương sai của ma trận khoảng cách mới. Bằng cách này, chúng ta có thể ước lượng tốt hơn việc chọn cặp đơn vị phân loại để vun đống trong các bước tiếp theo. Hơn nữa, so với ước lượng của NJ, những ước lượng này trở nên ngày càng tốt hơn.

Về cơ bản, so với thuật toán NJ, thuật toán BIONJ tốn thêm chi phí tính toán về thời gian, không gian; và nhu cầu không gian bộ nhớ cũng gấp đôi. Tuy nhiên điều này thực tế không quan trọng đối với các máy tính hiện đại. Ưu điểm của thuật toán này là đơn giản hơn, nhưng có độ chính xác cao hơn với cùng một thời gian tính toán. BIONJ được sử dụng rộng rãi khi có một khoảng cách tiến hóa đáp ứng các giả thuyết của thuật toán.

- ***Thuật toán FNJ*** (Fast Neighbor Joining) được Isaac Elias and Jens Lagergren xây dựng [6] để cải thiện độ phức tạp của thuật toán NJ. Thuật toán này có bán kính xây dựng tối ưu và độ phức tạp về thời gian là $O(n^2)$. Các thực nghiệm ban đầu cho thấy FNJ gần chính xác như NJ, chứng tỏ rằng bán kính xây dựng lại tối ưu.

4.2. Dự đoán chức năng protein dựa trên mạng tương tác protein

Sự bùng nổ của dữ liệu sinh học mở đường cho việc nghiên cứu chú thích chức năng protein cùng với sự xuất hiện và phổ biến các dự đoán chức năng tự động. Nhiều phương pháp tiếp cận như vậy đã được nghiên cứu, bao gồm cả việc sử dụng các chuỗi tương đồng, tương tác protein - protein, cấu trúc protein, dạng thể hiện, hồ sơ phát sinh loài. Các công cụ phát triển từ dự đoán chức năng tự động sẽ cung cấp cho hệ thống như tài liệu tiềm năng của các chú thích được xác minh bằng thực nghiệm. Điều này làm cho chú thích chức năng của protein ngày càng nhiều hơn.

Để thu được kết quả chính xác trong dự đoán chức năng protein, các dữ liệu cần phải đầy đủ, hoặc không bị nhiễu (chứa nhiều dương tính giả, do các protein dính có thể kích hoạt các gen của các protein không tương tác), và cần được cung cấp một lược đồ chú thích chuẩn có ý nghĩa cũng như một công ước đặt tên chung.

Bài toán dự đoán chức năng được xây dựng dựa trên mạng tương tác protein (TTP) bởi vì các protein không tồn tại rời rạc hay độc lập nhau. TTP được xem là nguồn quan trọng của thông tin liên quan đến quá trình sinh học và chức năng trao đổi chất phức tạp của tế bào. Từ mạng tương tác protein, sử dụng một mức đo khoảng cách (Czekanowski-Dice) [7] để tính giá trị khoảng cách giữa tất cả các cặp protein và áp dụng các thuật toán phân cụm NJ, BIONJ, FNJ trên ma trận khoảng cách protein để xây dựng một cây phân cấp. Các lớp chức năng được xác định theo topo cây và số lượng protein chia sẻ các chú thích chức năng. Các lớp kết quả được gán một chức năng sinh học theo chú thích chức năng của các thành viên của nó theo một quy luật đa số cổ điển. Các dự đoán chức năng cho protein chưa biết đặc trưng sau đó sẽ được đề xuất dựa trên lớp chức năng cụ thể.

Sơ đồ thực hiện dự đoán chức năng protein từ mạng tương tác [8,9,10,11], áp dụng cho các thuật toán NJ, BIONJ, FNJ nói riêng và các thuật toán phân cụm kiểu vun đống (tạo cây phân cấp) nói chung gồm có 4 bước. Đầu tiên, từ mạng tương tác chúng ta chuyển thành ma trận khoảng cách, sau đó áp dụng các thuật toán phân cụm phân cấp như NJ, BIONJ, FNJ để tạo ra cây phân cấp. Từ đây, dựa vào danh sách các chức năng của protein chúng ta tạo phân lớp chức năng thỏa mãn tiêu

chuẩn tạo cụm. Và cuối cùng là dự đoán chức năng từ các cụm nhờ vào việc phân tích các cụm đã tạo ra ở bước trên để xem xét cho các protein chưa biết chức năng.

Với sơ đồ này, chúng tôi đã thử xây dựng một chương trình demo ứng dụng thuật toán NJ, BIONJ, FNJ, sau đó áp dụng nghiên cứu với protein Cerevisiae. Kết quả cho thấy các thuật toán phân cụm tốt các protein theo chức năng và thuật toán BIONJ, FNJ có nhiều ưu điểm so với thuật toán NJ. Chúng tôi sẽ trình bày những nghiên cứu này ở bài sau.

5. KẾT LUẬN

Trong bài báo này chúng tôi đã trình bày tổng quan về phân cụm dữ liệu và tin sinh học. Phân cụm dữ liệu là một chủ đề được nghiên cứu tích cực trong nhận dạng và học máy, đã được áp dụng để phân nhóm đối tượng trong nhiều lĩnh vực khác nhau. Tin sinh học cũng là ngành khoa học áp dụng tin học, thống kê để xử lý các vấn đề sinh học. Chúng tôi cũng khái quát một số ứng dụng của các thuật toán phân cụm dữ liệu trong dự đoán chức năng protein thông qua việc phân cụm mạng tương tác giữa các protein. Các protein có cùng chức năng có thể có cấu trúc protein tương đồng nhau, từ đó có thể suy ra các mối quan hệ tiến hóa giữa các loài.

Các thuật toán có các ưu điểm, hạn chế riêng, và dựa vào phân tích kết quả thu được ta có thể đánh giá đó là một thuật toán tốt hay nghèo. Chúng tôi đã nghiên cứu áp dụng một số thuật toán trong phân cụm chức năng Protein trên cơ sở dữ liệu của loài Cerevisiae (nấm bánh mì), và các kết quả chính sẽ được trình bày trong bài báo tiếp theo.

TÀI LIỆU THAM KHẢO

1. http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/index.html.
2. Jain A.K., Murty M.N., Flynn P.J. (1999), *Data Clustering: A Review*, ACM Computing Surveys, Vol. 31, No. 3.
3. Olivier, Mike Steel (2006), *Neighbor-Joining Revealed*, Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
4. James A. Studier, Karl J. Keppler (1988), *Letter to the Editor: A Note on the Neighbor-Joining Algorithm of Saitou and Neil,* Mol. Biol. Evol. 5(6):729
5. Naruya Saitou, Masatoshi Nei (1987), *The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees*, Center for Demographic and Population Genetics, The University of Texas Health Science Center at Houston.
6. Isaac Elias and Jens Lagergren (2005), *Fast Neighbor Joining*, Dept. of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, Sweden. Springer-Verlag Berlin Heidelberg.
7. Chuang Peng (2006), *Distance Based Methods In Phylogentic Tree Construction*, Department Of Mathematics Morehouse College Atlanta, Ga 30314.
8. Anais Baudot, Bernard Jacq and Christine Brun (2004), *A scale of functional divergence for yeast duplicated genes revealed from analysis of the protein-protein interaction network*.
9. Anaıs Baudot, David Martin, Pierre Mouren, Francois Chevenet, Alain Gue noche, Bernard Jacq and Christine Brun (2005), *PRODISTIN Web Site: a tool for the functional classification of proteins from interaction networks*, Published by Oxford University Press.
10. Christine Brun, Carl Herrmann and Alain Guénoche (2004), *Clustering proteins from interaction networks for the prediction of cellular functions*, BMC Bioinformatics 2004, 5:95 doi:10.1186/1471-2105-5-9.
11. Christine Brun, François Chevenet, David Martin, Jérôme Wojcik, Alain Guénoche and Bernard Jacq (2003), *Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network*, Genome Biology 2003, 5:R6.