LẬP CHỈ MỤC CƠ SỞ DỮ LIỆU CẤU TRÚC PROTEIN

Phan Mạnh Thường[1], Lâm Thị Hoà Bình[1], Đặng Như Toàn[1], Đoàn Thiện Minh[1]
Trần Văn Lăng[2]

[1] *Khoa Công nghệ thông tin, Trường Đại học Lạc Hồng*
10 Huỳnh Văn Nghệ, Biên Hòa, Đồng Nai
{thuong,binh,dangnhutoan,dtminh}@lhu.edu.vn

[2] *Viện Khoa học và Công nghệ Việt Nam*
01 Mạc Đĩnh Chi, Quận 1, TP. Hồ Chí Minh
tvlang@vast-hcm.ac.vn

Tóm tắt. Tìm kiếm sự tương đồng về cấu trúc bậc ba của các protein trong cơ sở dữ liệu cấu trúc protein lớn là một bài toán phức tạp và đòi hỏi nhiều thời gian xử lý. Số lượng các cấu trúc protein được khám phá ngày càng gia tăng nhanh chóng và trong các cơ sở dữ liệu về cấu trúc protein việc lập chỉ mục cho các protein sẽ giúp thao tác tìm kiếm, so sánh cấu trúc thực hiện nhanh hơn và hiệu quả hơn. Trong bài báo này trình bày một phương pháp lập chỉ mục cho cơ sở dữ liệu cấu trúc protein thông qua việc phân tích cấu trúc, từ đó rút ra vector đặc trưng và xây dựng một cấu trúc cây dựa trên các vector đặc trưng để lập chỉ mục cho cấu trúc protein. Với cơ sở dữ liệu đã được lập chỉ mục, việc tìm kiếm một cấu trúc protein hoặc một cấu trúc con trong protein trở nên nhanh chóng và chính xác hơn.

Từ khoá: Cấu trúc protein bậc ba, lập chỉ mục, cơ sở dữ liệu protein.

1. Đặt vấn đề

Protein là một chuỗi polypeptide được tạo thành từ các axít amin. Nghiên cứu protein đóng vai trò quan trọng, vì chúng hoạt động trong tất cả các quá trình sinh học, bao gồm cả xúc tác enzym (tất cả các phản ứng hóa học trong tế bào sống được xúc tác

bởi enzyme protein), vận chuyển các chất khác nhau như dưỡng khí, các ion …, và tín hiệu. Để hiểu được mối quan hệ giữa cấu trúc và chức năng của protein, các nhà nghiên cứu cần phải lấy từ cơ sở dữ liệu cấu trúc protein và phân loại chúng thành các họ protein khác nhau.Vấn đề quan trọng trong việc gom nhóm các protein dựa trên sự tương đồng cấu trúc nhằm mục tiêu:

- Phát hiện các mối quan hệ tiến hóa
- Xác định các motif (đoạn lặp), là những cấu trúc được hình thành bởi sự sắp xếp của các axit amin trong không gian ba chiều
- Phát hiện mối quan hệ giữa cấu trúc và chức năng của protein
- Hỗ trợ trong việc thiết kế thuốc trị bệnh
- Phát hiện các trình tự có liên quan đến bệnh ung thư và các bệnh khác.

Với sự đổi mới công nghệ và phát triển nhanh chóng của các phương pháp xác định cấu trúc protein như: phương pháp X-quang tinh thể, kỹ thuật phân tích quang phổ NMR…, một số lượng lớn các cấu trúc 3 chiều của các phân tử protein mới đã được xác định. Các cấu trúc này hiện đang được lưu trữ tại nhiều cơ sở dữ liệu trên internet và cung cấp miễn phí cho các nhà nghiên cứu, có thể kể đến:

- Ngân hàng dữ liệu protein PDB [1] (Protein Data Bank) thuộc phòng thí nghiệm RCSB (Research Collaboratory for Structural Bioinformatics): bao gồm 73153 cấu trúc
- SCOP Structural Classification of Proteins [2]: bao gồm 38221 cấu trúc
- CATH Protein Structure Classification [3]: bao gồm 104238 cấu trúc
- ModBase Database of Comparative Protein Structure Models (Sali Lab, UCSF): bao gồm 41140 cấu trúc

Tìm kiếm sự tương đồng về cấu trúc bậc ba của một protein hoặc một cấu trúc con của protein bất kỳ trong cơ sở dữ liệu cấu trúc protein ngày càng lớn là một nhiệm vụ khó khăn và tốn thời gian. Vì vậy các nhà sinh học đang cần một phương tiện để tìm kiếm cơ sở dữ liệu cấu trúc protein nhanh chóng và hiệu quả, tương tự như cách BLAST [5] tìm kiếm trong cơ sở dữ liệu trình tự. Bài toán tìm kiếm và phân loại protein thường trải qua hai giai đoạn: rút trích đặc trưng mô tả cho protein và đo sự giống nhau về đặc trưng của các protein để phân loại chúng.

Để thực hiện rút trích đặc trưng của cấu trúc protein có rất nhiều thuật toán, thuật toán CTSS [6] xấp xỉ cấu trúc các Cα xương sống của protein bằng một đường spline mịn với độ cong tối thiểu, sau đó lưu trữ đường cong, góc xoắn và cấu trúc bậc hai của mỗi nguyên tử Cα trong một mục chỉ số dựa trên phép băm.

ProGreSS [5] là một phương pháp mới, thực hiện rút trích đặc trưng từ cấu trúc kết hợp với trình tự thông qua một cửa sổ trượt trên cấu trúc xương sống của protein. Đặc trưng về cấu trúc của nó tương tự như các đặc trưng rút ra từ CTSS (độ cong, góc xoắn, và thông tin cấu trúc bậc hai); các chuỗi đặc trưng được tính toán từ việc sử dụng ma trận điểm như PAM hoặc BLOSUM. Giống như CTSS, các đặc trưng rút ra từ ProGreSS không phải là đặc trưng cục bộ.

Thuật toán PSIST[7] là một trong số các thuật toán hiệu quả vì có độ chính xác tương đối cao, cách tiếp cận của thuật toán PSIST là biến đổi các thông tin cấu trúc cục bộ của một protein thành một "trình tự" và dựa trên tập các "trình tự" đó xây dựng một cây hậu tố phục vụ cho việc tìm kiếm. So với cách rút trích các đặc trưng cục bộ từ một axit amin duy nhất, thì cách rút trích đặc trưng theo cửa sổ trượt trong hướng tiếp cận của thuật toán PSIST là tốt hơn vì vector đặc trưng hàm chứa cả hai thông tin tịnh tiến và xoay ở bên trong. Sau khi các vectơ đặc trưng được chuẩn hóa, cấu trúc protein được chuyển thành một chuỗi (gọi là trình tự đặc trưng-cấu trúc) của các ký hiệu được rời rạc hoá.

Tuy nhiên việc tìm kiếm trên cây hậu tố thực sự chưa đạt hiệu quả cao về tốc độ, thuật toán PSISA[8] sử dụng hướng tiếp cận trích vector đặc trưng giống PSIST, nhưng thay vì dùng cây hậu tố thì thuật toán này sử dụng mảng hậu tố trong phương pháp đánh chỉ mục nhằm tăng tốc độ tìm kiếm. Kết quả thực nghiệm trong PSISA chỉ ra rằng, đánh chỉ mục bằng mảng hậu tố giúp tăng tốc độ tìm kiếm nhưng đồng thời cũng làm gia tăng khả năng sử dụng bộ nhớ với hệ số lên đến hơn 35% so với cây hậu tố như trong PSIST.

Trong bài báo này, trình bày một phương pháp lập chỉ mục cho cơ sở dữ liệu cấu trúc protein thông qua việc kế thừa thuật toán PSIST để rút ra vector đặc trưng và từ tập các vector đặc trưng bài báo đề xuất xây dựng một cấu trúc cây chỉ mục dựa trên việc ghép nhánh các chuỗi vector đặc trưng, cấu trúc cây này vừa giúp hạn chế việc sử dụng bộ nhớ và vừa cho phép tìm kiếm trên không gian của toàn bộ các cấu trúc thuộc

các họ protein khác nhau, điều này giúp cho việc tìm kiếm một cấu trúc protein hoặc một tiểu cấu trúc trong protein trở nên nhanh chóng và chính xác hơn.

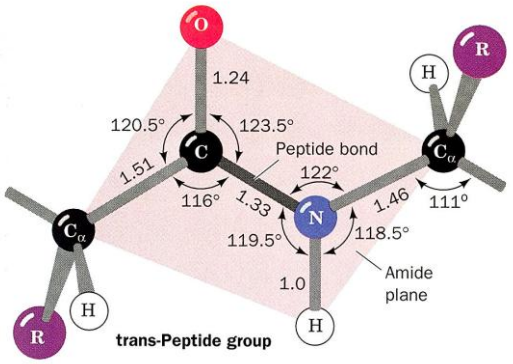
Các nội dung còn lại của bài báo được trình bày như sau: phần thứ hai trình bày phương pháp lập chỉ mục dữ liệu cấu trúc protein, cách thức rút trích vector đặc trưng, chuẩn hóa vector đặc trưng, cũng như việc xây dựng cây chỉ mục; phần thứ ba nêu lên một số thử nghiệm từ nguồn dữ liệu cấu trúc protein, việc truy vấn trên nguồn dữ liệu này; phần cuối cùng trình bày một số đánh giá và kết luận.

2. Lập chỉ mục dữ liệu cấu trúc protein

a) Rút trích vector đặc trưng

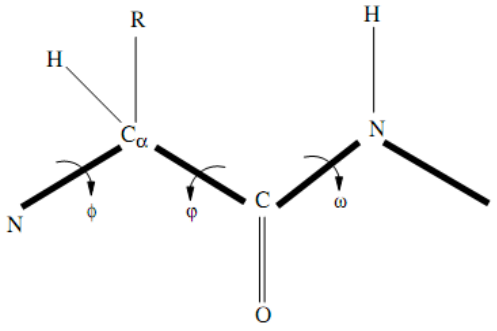
Mỗi protein là một tổ hợp của một chuỗi có thứ tự các axit amin (residue) được liên kết với nhau bởi các liên kết peptide. Mỗi residue gồm một C_α, các N và C khác. Chiều dài của liên kết, góc liên kết và các góc xoắn hoàn toàn xác định cấu tạo và hình học của protein.

Độ dài liên kết là khoảng cách giữa các nguyên tử được nối kết được tính bằng đơn vị Amstrong ($\mathring{A}$), và góc liên kết là góc giữa hai liên kết cộng hoá trị của cùng một nguyên tử. Ví dụ, độ dài liên kết giữa cặp nguyên tử N-C là 1.33 $\mathring{A}$, góc liên kết giữa C_α-N và N-C là 122^0.



Hình 1. Độ dài liên kết và các góc liên kết giữa các nguyên tử

Góc xoắn dùng để mô tả các cấu trúc có thể xoay quanh các liên kết. Giả sử ta có bốn nguyên tử được kết nối thông qua ba liên kết B_{i-1}, B_i và B_{i+1}, thì góc xoắn của mối liên kết B_i được định nghĩa bằng góc nhỏ nhất của các hình chiếu B_{i-1} và B_{i+1} lên mặt phẳng vuông góc với B_i



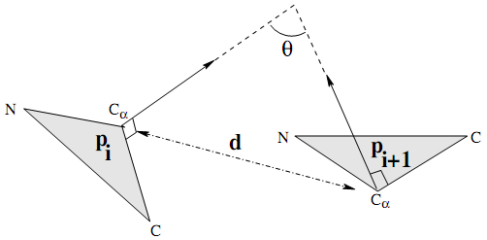
Hình 2. Các góc xoắn ϕ, φ và ω giữa các nguyên tử

Để chụp được các đặc trưng cục bộ một cách chính xác hơn, cần phải trích xuất các đặc trưng từ một tập các residue cục bộ. Để tạo ra vector đặc trưng cục bộ, đầu tiên mô tả từng residue riêng biệt và xác định sự liên hệ giữa một cặp residue và giữa một tập các residue với nhau. Với mỗi residue, độ dài liên kết C_α-N là 1.46 $\mathring{A}$, liên kết C_α-C là 1.51 $\mathring{A}$ và góc giữa C_α-N và C_α-C là 116^0. Như vậy tất cả các tam giác tạo nên từ các nguyên tử N-C_α-C của mỗi residue là tương đương như nhau và mỗi residue có thể đại diện bởi một tam giác.

Khoảng cách d giữa một cặp residue được xác định dựa trên khoảng cách Euclide giữa hai nguyên tử C_α của chúng. Công thức (1) được sử dụng để tính toán khoảng cách giữa hai residue

$$d(x,y) = \|x - y\| = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

Góc θ giữa một cặp residue được xác định bằng góc giữa hai mặt phẳng tạo nên từ ba nguyên tử N-C_α-C của mỗi residue.



Hình 3. Khoảng cách và góc giữa hai residue

Khoảng cách và góc là bất biến đối với phép dịch chuyển và xoay protein. Khoảng cách Euclide giữa hai nguyên tử C_α được tính trực tiếp từ các toạ độ trong không gian ba chiều của chúng. Góc giữa hai mặt phẳng tạo nên từ bộ ba nguyên tử N-C_α-C được tính toán dựa trên góc của cặp vector pháp tuyến có gốc xuất phát từ nguyên tử C_α của mỗi mặt phẳng. Vector pháp tuyến này được tính bởi công thức (2)

$$\vec{n} = \frac{\overrightarrow{NC_\alpha} \times \overrightarrow{C_\alpha C}}{\|\overrightarrow{NC_\alpha} \times \overrightarrow{C_\alpha C}\|} \quad (2)$$

Góc giữa hai vector pháp tuyến n1 và n2 được tính theo công thức (3)

$$\cos\theta = \frac{\|\vec{n1}\|^2 + \|\vec{n2}\|^2 - \|\vec{n2} - \vec{n1}\|^2}{2 * \|\vec{n1}\| * \|\vec{n2}\|} \quad (3)$$

Để mô tả các đặc trưng cục bộ từ một tập các residue, nhóm tác giả dùng một cửa sổ có kích thước w trượt qua trên chuỗi C_α xương sống của protein. Các khoảng cách và các góc giữa residue đầu tiên và các residue còn lại trong cửa sổ sẽ được tính toán và thêm vào vector đặc trưng, mỗi cửa sổ ứng với một vector đặc trưng.

Cho tập $P=\{p_1,p_2,..p_n\}$ đại diện cho một protein, trong đó p_i là residue thứ i trong cấu trúc xương sống của protein. Vector đặc trưng của protein được định nghĩa là $P_v=\{p_{v1}, p_{v2}, \ldots, p_{vn-w+1}\}$, trong đó w là độ rộng cửa sổ trượt và p_{vi} là vector đặc trưng có

$p_{vi}=(d(p_i,p_{i+1}), \cos\theta(p_i,p_{i+1}),..., d(p_i,p_{i+w-1}), \cos\theta(p_i,p_{i+w-1}))$

với $d(p_i, p_j)$ là khoảng cách giữa hai residue thứ i và j và $\cos\theta(p_i,p_j)$ cho bởi góc giữa hai residue. Với cửa sổ có kích thước w thì chiều của mỗi vector đặc trưng p_{vi} là 2(w-1).

b) **Chuẩn hoá vector đặc trưng**

Do các vector đặc trưng chứa các thông tin về khoảng cách và góc liên kết với đơn vị đo lường khác nhau nên cần phải được chuẩn hoá. Thêm nữa việc chuẩn hoá sẽ giúp hạn chế bớt miền giá trị của các thành phần trong vector đặc trưng. Góc θ thuộc phạm vi [0, π], vì vậy cos θ ∈ [-1, 1]. Để chuẩn hóa khoảng cách, chúng ta cần phải biết cận trên về khoảng cách giữa residue thứ i và residue thứ (i+w-1) trong protein.

Tất cả các khoảng cách và các góc đều được chuẩn hoá và đưa về một số nguyên trong khoảng [0, b-1] với b là một tham số cho trước.

Mỗi khoảng cách d trong vector đặc trưng sẽ được chuẩn hoá theo công thức (4)

$$d = \left\lfloor \frac{d * b}{4.025 * (w-1)} \right\rfloor \quad (4)$$

trong công thức (4) giá trị hằng số 4.025 là khoảng cách trung bình giữa hai nguyên tử C_α , và w là độ rộng cửa sổ trượt.

Các góc trong vector đặc trưng sẽ được chuẩn hoá theo công thức (5)

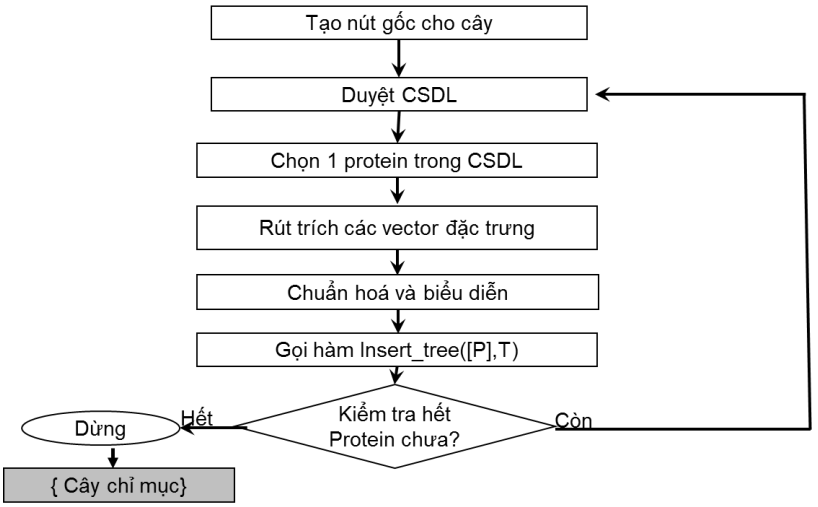
$$\cos\theta = \left\lfloor \frac{(\cos\theta + 1) * b}{2} \right\rfloor \quad (5)$$

Sau khi chuẩn hoá, cấu trúc protein sẽ được biểu diễn bằng một chuỗi "trình tự" các giá trị rời rạc theo các vector đặc trưng, trong đó vector thứ i biểu diễn đặc trưng của residue thứ i trong chuỗi xương sống của protein.

c) Xây dựng cây chỉ mục

Để tiến hành lập chỉ mục cho tập dữ liệu cấu trúc protein, bài báo đề xuất xây dựng một cấu trúc cây nhiều nhánh theo thuật toán như trong hình 4.

Đầu tiên, thuật toán sẽ đọc dữ liệu cấu trúc của từng protein trong cơ sở dữ liệu, sau đó tiến hành rút trích đặc trưng dựa theo thuật toán đã trình bày nhằm "trình tự" hoá cấu trúc ba chiều của mỗi protein bằng một tập các vector đặc trưng ứng với cấu trúc xương sống của nó. Sau khi chuẩn hoá các vector đặc trưng, mỗi "trình tự" cấu trúc protein sẽ được thêm vào trong cây chỉ mục để phục vụ cho việc tra cứu .



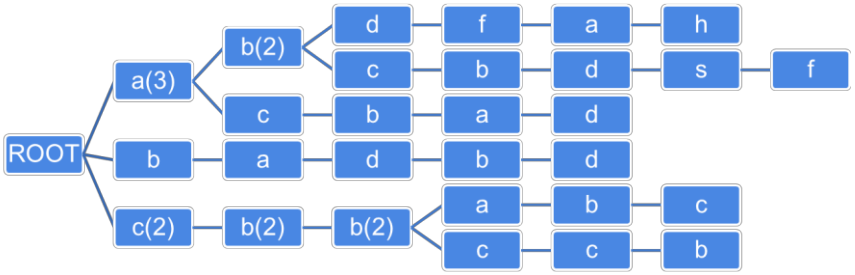
Hình 4. Thuật toán tạo cây chỉ mục dựa trên đặc trưng cấu trúc của protein.

Ví dụ: Xây dựng cây chỉ mục từ tập gồm sáu cấu trúc protein đã trình tự hoá, ở đây mỗi trình tự protein được biểu diễn bởi một tập các ký tự, mỗi ký tự ứng với một vector đặc trưng đã được chuẩn hoá.

P1={a,b,d,f,a,h}; P2={b,a,d,b,d}; P3={a,b,c,b,d,s,f};

P4={c,a,b,a,b,c}; P5={c,a,b,c,c,b}; P6={a,c,b,a,d};

Kết quả sẽ được cấu trúc cây như hình 5.



Hình 5. Cây chỉ mục dựa trên đặc trưng cấu trúc của các protein.

d) Truy vấn dữ liệu trên cây chỉ mục

Cho một truy vấn Q, trước tiên các vector đặc trưng của cấu trúc Q sẽ được trích xuất và chuyển đổi thành một chuỗi "trình tự" như mô tả trong mục 2a và 2b. Sau đó, việc tra cứu sẽ được thực hiện qua ba giai đoạn: tìm kiếm, xếp hạng và chọn tối ưu. Giai đoạn tìm kiếm thống kê các cấu trúc trong cơ sở dữ liệu phù hợp với Q theo một ngưỡng khoảng cách ε giữa các vector, giai đoạn thứ hai xếp hạng tất cả các protein chứa chuỗi phù hợp tìm thấy, và giai đoạn sau cùng sử dụng thuật toán Smith-Waterman[9] để tìm kiếm cấu trúc tương đồng cục bộ tốt nhất dựa trên truy vấn Q và tập gồm các protein được lựa chọn.

Thuật toán tìm kiếm mẫu truy vấn Q trên cấu trúc cây chỉ mục được trình bày như sau:

Input: đoạn cấu trúc protein Q, ngưỡng so khớp nhỏ nhất ε

Output: Tập các cấu trúc protein thoả điều kiện tìm kiếm được sắp xếp theo số lượng residue so khớp giảm dần

*Function Search(**tree Root, mức i, chuỗi truy vấn Q, ngưỡng ε**){*

While (i < (chiều cao cây - độ dài chuỗi Q)){

- *Gom nhánh theo mức i;*
- *For each node tại mức i*

- *Nếu (node N[j] trùng khớp với Q[0])*
  - *For each nhánh con của N[j]: Nếu so khớp với phần còn lại của chuỗi Q thoả ngưỡng ε thì:*
    - *Thêm nhánh vào tập kết quả*
    - *Loại nhánh khỏi cây*
  - *Return Search (Root, i+1, Q[0], ε);*
- *Ngược lại*
  - *Return Search (N[j], i+1, Q[i+1], ε);*

} end while

}end function

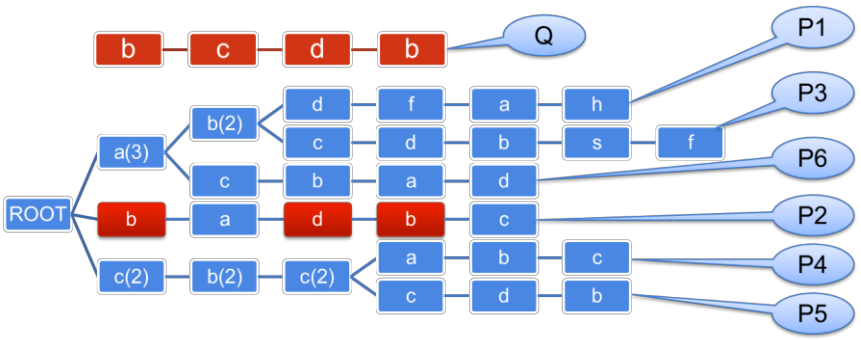
Function Query (tree Root, mẫu truy vấn Q, top k mẫu cần chọn, ngưỡng ε){

- *Khởi tạo tập kết quả rỗng;*
- *Rút trích đặc trưng và tạo chuỗi trình tự cấu trúc cho truy vấn Q*
- *Xây dựng cây chỉ mục;*
- *Search (Root, i=0, Q, ε);*
- *Sắp xếp tập kết quả giảm dần theo số lượng so khớp m;*
- *Chọn k mẫu tốt nhất trong tập kết quả và áp dụng thuật toán Smith-Waterman tìm sắp hàng cấu trúc cục bộ tốt nhất.*

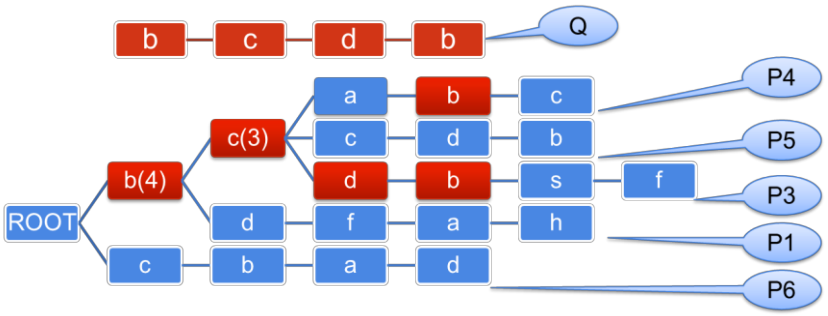
}end function

Ví dụ: Tìm kiếm mẫu truy vấn Q={b,c,d,b}trên cây chỉ mục từ tập các cấu trúc protein đã trình tự hoá với ngưỡng ε=3. Tập gồm P1={a,b,d,f,a,h}; P2={b,a,d,b,c}; P3={a,b,c,d,b,s,f}; P4={c,b,c,a,b,c}; P5={c,b,c,c,d,b}; P6={a,c,b,a,d}

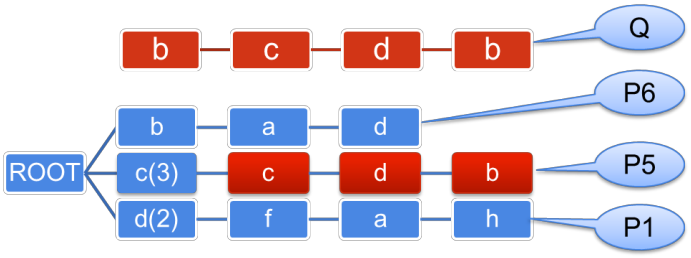
- Truy vấn tại mức gốc (mức 0)→ Tập kết quả={P2 (số so khớp m=3)}



- Truy vấn tại mức 1→ Tập kết quả={P4 (số so khớp m=3), P3 (m=4)}



- Truy vấn tại mức 2→ Tập kết quả={P5 (số so khớp m=3)}



3. Một số kết quả thử nghiệm

a) Các nguồn dữ liệu cấu trúc protein

Các cấu trúc protein bậc ba được lưu trữ nhiều tại ngân hàng dữ liệu Protein (PDB – Protein Data Bank) [1], đó là kho lưu trữ chính cho thực nghiệm xác định cấu trúc bậc ba của Protein. Ngân hàng PDB được tạo ra vào năm 1971 tại Phòng thí nghiệm quốc gia Brookhaven (BNL) ở Mỹ. Những cấu trúc được xác định nhờ sử dụng phương pháp tinh thể học. Hiện nay có hơn 73153 cấu trúc protein trong kho lưu trữ tại PDB và hàng năm có hơn 6000 công trình mới được lưu trữ.

Các protein trong cơ sở dữ liệu SCOP [2], được tổ chức tại Phòng thí nghiệm Sinh học Phân tử của Hội đồng Nghiên cứu Y khoa (MRC) ở Cambridge, Anh, mô tả các mối quan hệ cấu trúc và tiến hóa giữa các cấu trúc protein đã được biết đến. SCOP đã được chấp nhận là phù hợp nhất và phân loại các tập dữ liệu đáng tin cậy nhất, do thực tế rằng SCOP xây dựng quyết định phân loại của nó dựa trên những quan sát trực quan các yếu tố cấu trúc của protein do các chuyên gia thực hiện. Protein được phân loại một cách có thứ bậc, phản ánh mối quan hệ của chúng về cấu trúc và tiến hóa. Các cấp chính của hệ thống phân cấp là "họ gia đình" (family) (dựa trên các mối quan hệ tiến

hóa của các protein), "siêu họ" (superfamily) (dựa trên một số đặc điểm chung về cấu trúc), và "gấp cuộn" (fold) (dựa trên các yếu tố cấu trúc bậc hai).

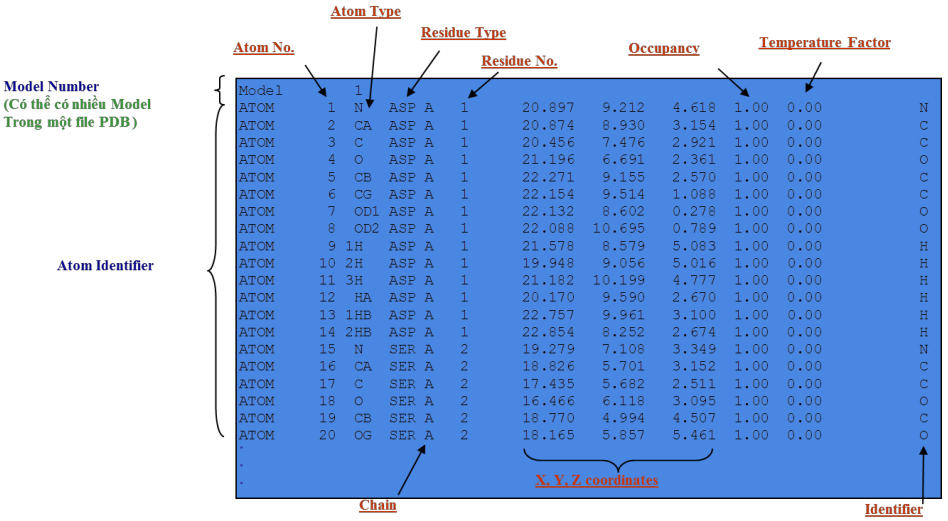
Cơ sở dữ liệu CATH [3], được tổ chức tại Đại học UCL London, hiện có 104238 cấu trúc, sử dụng phương pháp tự động để phân loại protein, và cũng có những đóng góp của các chuyên gia khi phương pháp tự động không cho kết quả đáng tin cậy. Cơ sở dữ liệu CATH được xây dựng bằng cách áp dụng công cụ so sánh cấu trúc bậc hai SSAP. SSAP sử dụng một kỹ thuật lập trình quy hoạch động hai lớp để so khớp hai protein và tìm ra cấu trúc liên kết tối ưu của hai protein.

Cơ sở dữ liệu FSSP [4] đã được tạo ra theo phương pháp phân loại DALI và được tổ chức tại Viện Tin sinh học châu Âu (EBI). Nó cung cấp một phân loại phức tạp của các cấu trúc protein. Sự tương tự giữa hai protein được xác định dựa trên cấu trúc bậc hai của chúng. Việc đánh giá từng cặp protein là một công việc tốn thời gian, vì vậy việc so sánh giữa một đại phân tử và tất cả các đại phân tử của các cơ sở dữ liệu có thể mất cả ngày. Do đó, một protein đại diện cho mỗi lớp được xác định và mỗi protein mới chỉ phải so khớp với protein đại diện của từng loại.

b) **Tổ chức lưu trữ**

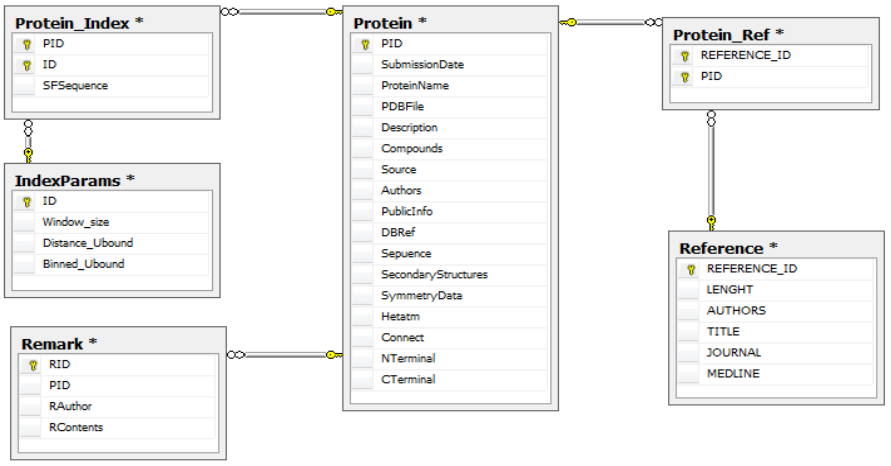
Các cấu trúc bậc ba của protein thông thường được lưu trữ theo các định dạng như: MMDB “Molecular Modeling DataBank” (định dạng chuẩn mô tả thông tin các liên kết peptide), mmCIF “Chemical Interchange Format” (dạng cơ sở dữ liệu quan hệ) và PDB “Protein DataBank” (dạng cột văn bản với nhiều mục thông tin tích hợp).

Trong số các định dạng nêu trên thì định dạng PDB là phổ biến hơn cả, trong tập tin PDB lưu trữ các thông tin về toạ độ của các nguyên tử trong không gian ba chiều theo hệ quy chiếu Euclide, ngoài ra còn có các thông tin về tác giả, các tham chiếu, và các kết quả thực nghiệm xác định cấu trúc protein.



Hình 6. Một phần cấu trúc tập tin PDB

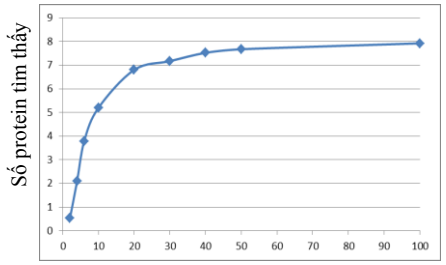
Nhóm tác giả bài báo đã thực hiện thu thập các cấu trúc đã được công bố từ các nguồn [1, 2, 3, 4] dưới định dạng PDB và tổ chức lưu trữ trong một cơ sở dữ liệu quan hệ để thuận tiện cho việc lập chỉ mục và tra cứu. Mô hình cơ sở dữ liệu quan hệ được đề xuất như trong hình 7.



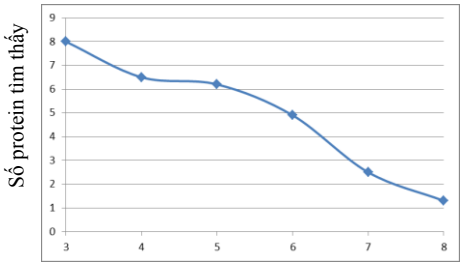
Hình 7. Mô hình cơ sở dữ liệu quan hệ lưu trữ thông tin cấu trúc protein

c) Một số kết quả thử nghiệm

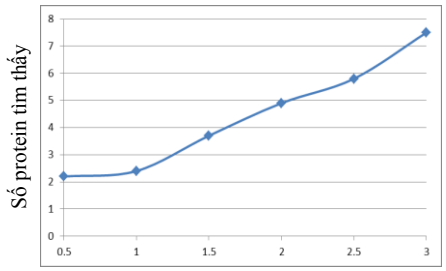
Dưới đây là một số kết quả thử nghiệm: Tập dữ liệu D dùng cho thử nghiệm được rút trích từ cơ sở dữ liệu SCOP [2] gồm các protein thuộc cả bốn lớp: cuộn α, phiến β, α+β và α/β. Tập dữ liệu bao gồm 10 protein thuộc mỗi "siêu họ" (superfamily) trong tổng số 181 "siêu họ" của SCOP, như vậy có tổng cộng 1810 protein. Mẫu truy vấn sẽ được lấy ngẫu nhiên từ tập dữ liệu D trong các thử nghiệm. Có 5 tham số trong các thử nghiệm gồm: w là độ rộng cửa sổ, b là giá trị chuẩn hoá, ε ngưỡng khoảng cách tối thiểu giữa hai vector, l là độ dài tối thiểu phải đạt của chuỗi so khớp lớn nhất và k là số lượng protein được lấy từ trên xuống theo điểm số. Thuật toán được cài đặt bằng C++ và chạy thử nghiệm trên môi trường Windows với cấu hình máy CPU Dual 1.6GHz, RAM 2GB. Số protein thể hiện trong đồ thị là số trung bình các protein tìm thấy trong 181 "siêu họ" qua các thử nghiệm.



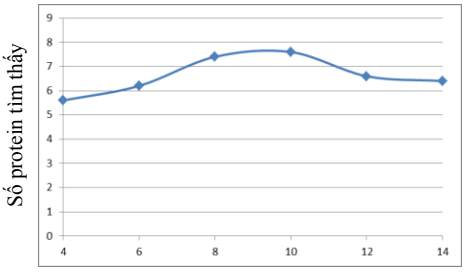
Hình 8. Số protein tìm thấy trong cùng superfamily theo số lượng k cutoff
(w=3, b=10, ε=3 và l=10)



Hình 9. Số protein tìm thấy trong cùng superfamily theo kích thước cửa sổ w
(b=10, ε=3 và l=15)



Hình 10. Số protein tìm thấy trong cùng superfamily theo ngưỡng khoảng cách ε
(w=3, b=10, và l=15)



Hình 11. Số protein tìm thấy trong cùng superfamily theo giá trị chuẩn hoá b
(w=3, ε=2.5, và l=15)

d) **Đánh giá và nhận xét**

Trong hình 8 cho thấy số protein tìm được trong cùng superfamily đặt được mức trung bình khoảng 7.36 với số cutoff từ 20 đến 100, kết quả này cho thấy hiệu quả tìm kiếm gần tương đương với PSIST. Kết quả ở hình 9 cho thấy thuật toán hoạt động ổn định với kích thước cửa sổ khoảng từ 3 đến 5, nếu vượt qua khoảng này thì hiệu quả giảm thấy rõ do các sai số phát sinh trong quá trình rút đặc trưng và chuẩn hoá vector. Có thể cải thiện vấn đề này bằng cách gia tăng giá trị chuẩn hoá như kết quả thể hiện trong hình 11, tuy nhiên việc này sẽ dẫn đến tăng thời gian xử lý và không gian lưu trữ các vector đặc trưng.

Kết quả cho thấy hiệu suất của thuật toán gần tương đương với PSIST và có phần tốt hơn ProGreSS, tuy nhiên nếu xét về mặt lưu trữ thì thuật toán PSIST cần nhiều không gian hơn cho cây hậu tố nếu phải chạy trên tập dữ liệu lớn và thao tác tìm kiếm cũng phức tạp hơn nhưng có độ chính xác cao hơn thuật toán bài báo đề xuất.

Thuật toán đề xuất có những điểm tốt

- Cây chỉ mục được xây dựng một lần và hiệu chỉnh nhiều lần trong quá trình tìm kiếm. Độ phức tạp tìm kiếm chuỗi Q độ dài l trên cây chỉ mục chiều cao h là O(k*(h-l)*b), k là số trung bình các nhánh có trùng giá trị ở mức i, b là số nhánh tại gốc.
- Việc gộp nhánh khi hiệu chỉnh cây sẽ cho phép tìm thấy cùng lúc nhiều cấu trúc thoả truy vấn, nhánh sau khi tìm thấy được loại bỏ khỏi cây để giảm không gian tìm kiếm trên các mức cao hơn.
- Thuật toán cho phép tìm trên toàn bộ không gian dữ liệu cấu trúc.

4. Kết luận

Trong bài báo này trình bày một hướng tiếp cận trong việc lập chỉ mục cho cơ sở dữ liệu cấu trúc bậc ba của protein dựa trên rút trích đặc trưng của protein theo thuật toán PSIST và đề xuất thuật toán tìm kiếm trên cấu trúc cây chỉ mục. Bài báo cũng trình bày về các nguồn dữ liệu cấu trúc bậc ba của protein, đề xuất mô hình cơ sở dữ liệu cho việc lưu trữ phục vụ thao tác lập chỉ mục và tra cứu thông tin các cấu trúc protein này. Dữ liệu dùng cho các thử nghiệm được rút trích từ 181 "siêu họ" của SCOP và các kết quả cho thấy độ chính xác tương đối cao và hiệu quả khi áp dụng các thuật toán đề xuất trên dữ liệu thử nghiệm.

Tài liệu tham khảo

[1] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, "The Protein Data Bank", Nucleic Acids Research, vol. 28, 2000, pp. 235-242.

[2] A.G. Murzin, S.E. Brenner, T. Hubbard, and C. Chothia, "Scop: A Structural Classification of Proteins Database for the Investigation of Sequences and Structures", J. Mol. Biol. 247, 1995, pp. 536-540.

[3] C.A. Orengo, A.D. Michie, D.T. Jones, M.B. Swindells, and J.M. Thornton, "CATH - A Hierarchic Classification of Protein Domain Structures", Structure, vol. 5, no. 8, 1997, pp. 1093-1108.

[4] L. Holm, and C. Sander, "The FSSP Database: Fold Classification Based on Structure - Structure Alignment of Proteins", Nucleic Acids Research, vol. 24, 1996, pp. 206-210.

[5] Can T. Kahveci T. Singh A.K. , A. and Y.F Wang, "Progress: Simultaneous searching of protein databases by sequence and structure", Pacific Symp. Bioinformatics, pages 264–275, 2004.

[6] T. Can and Y.Wang, "CTSS: a robust and efficient method for protein structure alignment based on local geometrical and biological features", IEEE Computer Society Bioinformatics Conference (CSB), pages 169–179, 2003.

[7] Mohammed J. Zaki Feng Gao, "PSIST: Indexing Protein Structures using Suffix Trees", in IEEE Computational Systems Bioinformatics Conference, Palo Alto, CA, August 2005.

[8] A. Salah Tarek F. Gharib and Abdel-Badeeh M.Salem, "PSISA: an Algorithm for Indexing and Searching Protein Structure using Suffix Arrays", In The WSEAS International Conference on Computers, pages 775–780, 2008.

[9] F. Smith and M. Waterman, "Identification of common molecular subsequences", J. Mol. Biol., (147):195–197, 1981.