SO SÁNH CẤU TRÚC PROTEIN SỬ DỤNG MÔ HÌNH TỔNG QUÁT

Văn Đình Vỹ Phương[1], Phan Mạnh Thường[1] , Trần Văn Lăng[2]
[(1)] *Khoa Công nghệ thông tin, Trường Đại học Lạc Hồng*
[(2)] *Viện Cơ học và Tin học ứng dụng, VAST*
{phuong,thuong}@lhu.edu.vn, tvlang@vast-hcm.ac.vn

Tóm tắt. *Bài viết trình bày phương pháp so sánh hai cấu trúc protein. Thực hiện xếp chồng và rút ngắn khoảng cách giữa nguyên tử Carbon-α của các phần tử hai protein để tìm ra được mô hình tương đồng cao nhất của hai protein. Nguồn protein thực hiện trong phương pháp được lấy từ ngân hàng protein thế giới - Protein Data Bank (PDB). Mặc dù có nhiều phương pháp thực hiện so sánh cấu trúc, nhưng vẫn còn nhiều vấn đề cần nghiên cứu và mở rộng. Phương pháp được trình bày trong bài báo được mở rộng từ phương pháp Chimera. Phương pháp đưa ra được kết quả tối ưu hơn so với cách sắp xếp chồng đơn thuần. Tính toán sự trùng khớp từ việc xếp hàng cấu trúc, rút ngắn khoảng cách hai cấu trúc và tiến hành dịch chuyển, giúp cho việc thể hiện sự tương đồng của protein một cách chính xác hơn. Tuy nhiên, vẫn còn một số hạn chế gặp phải và chưa giải quyết được: xử lý định hướng chuỗi liên kết; so sánh nhiều cấu trúc protein tại một thời điểm*.

Từ khoá: *cấu trúc protein, so sánh cấu trúc*

1. Đặt vấn đề

Protein đóng vai trò chính trong quá trình sinh học của động, thực vật. Với chuỗi trình tự amino acid giống nhau, nhưng sự liên kết phần tử, nếp gấp khác nhau sẽ tạo ra cấu trúc protein khác nhau, dẫn đến chức năng và cách thức hoạt động của protein đó cũng khác nhau. Việc dự đoán cấu trúc bậc 3 của protein để biết quy trình hoạt động, chức năng của protein vẫn là một thách thức lớn trong lĩnh vực sinh học tính toán.

Có nhiều cách thức để tìm cấu trúc protein, bằng kỹ thuật thực nghiệm có phương pháp chụp x-quang tinh thể, cộng hưởng từ hạt nhân, hoặc bằng các phương pháp dự đoán như Ab-Initio, mô hình hóa tương đồng.

Phương pháp cộng hưởng từ hạt nhân (NMR) **[1]** được sử dụng để xác định cấu trúc và tính năng của các protein. Việc xác định cấu trúc của protein theo phương pháp này là một quá trình tốn thời gian và đòi hỏi phải phân tích tương tác của dữ liệu. Có rất nhiều giai đoạn liên quan đến việc thực hiện cộng hưởng từ hạt nhân; chẳng hạn như chuẩn bị mẫu, cộng hưởng, tạo ra bản trữ, tính toán và xác định cấu trúc.

Với phương pháp X-quang tinh thể **[3]** hay được gọi là nhiễu xạ đơn tinh thể qua tia X, là một kỹ thuật phân tích trong đó sử dụng các mô hình nhiễu xạ tạo ra bằng cách bắn phá một tinh thể duy nhất với tia X để xác định cấu trúc tinh thể. Các mô hình nhiễu xạ được ghi lại và sau đó phân tích để tìm ra bản chất của tinh thể. Phương pháp này được sử dụng trong sinh hóa để xác định cấu trúc của một loạt các phân tử bao gồm DNA và protein.

Việc tìm kiếm cấu trúc protein bằng các phương pháp thực nghiệm rất khó khăn và tốn thời gian, các nhà nghiên cứu đã cố gắng để tự động hóa quá trình xác định cấu trúc ba chiều của protein bằng các phương pháp dự đoán.

Đối với các phương pháp dự đoán, trong đó phương pháp mô hình hóa tương đồng **[4]** là phương pháp liên quan đến việc xác định một cấu trúc protein được gọi là mẫu với các chuỗi truy vấn. Sau đó các nguyên tử trong chuỗi tìm kiếm sẽ được so khớp với bản đồ các nguyên tử có trong bản mẫu. Các chuỗi so khớp với các mẫu cấu trúc được sử dụng để tạo ra một mô hình cấu trúc kết quả. Phương pháp này dựa trên nguyên tắc là trong hầu hết các trường hợp tương đồng về trình tự thì cũng giống nhau về cấu trúc. Các bước chính liên quan đến việc mô hình hóa tương đồng được tóm tắt như sau: chọn mẫu, sắp hàng mẫu đích, xây dựng mô hình và đánh giá mô hình.

Phương pháp Ab-initio **[2]** xây dựng mô hình ba chiều của protein từ đầu dựa trên các nguyên lý vật lý và không đòi hỏi bất kỳ dữ liệu đầu vào như là một cấu trúc đã được biết đến hoặc một mô

hình cấu trúc. Dự đoán cấu trúc protein theo phương pháp Ab-Initio đòi hỏi các thuật toán mạnh mẽ và tài nguyên tính toán lớn.

Hiện nay số lượng các cấu trúc protein có trong PDB (Ngân hàng dữ liệu protein) **[5]** phát triển nhanh chóng với khoảng 73.153 (17/5/2011) cấu trúc đã biết. Tuy nhiên, đây cũng chỉ là một con số quá nhỏ so với những cơ thể sống đang có xung quanh con người chúng ta. Chính vì vậy, việc gom nhóm và tìm hiểu cấu trúc của protein để phát hiện các mối quan hệ tiến hóa, xác định các motif (đoạn lặp), phát hiện mối quan hệ giữa cấu trúc và chức năng của protein là một nhu cầu to lớn của khoa học về sự sống.

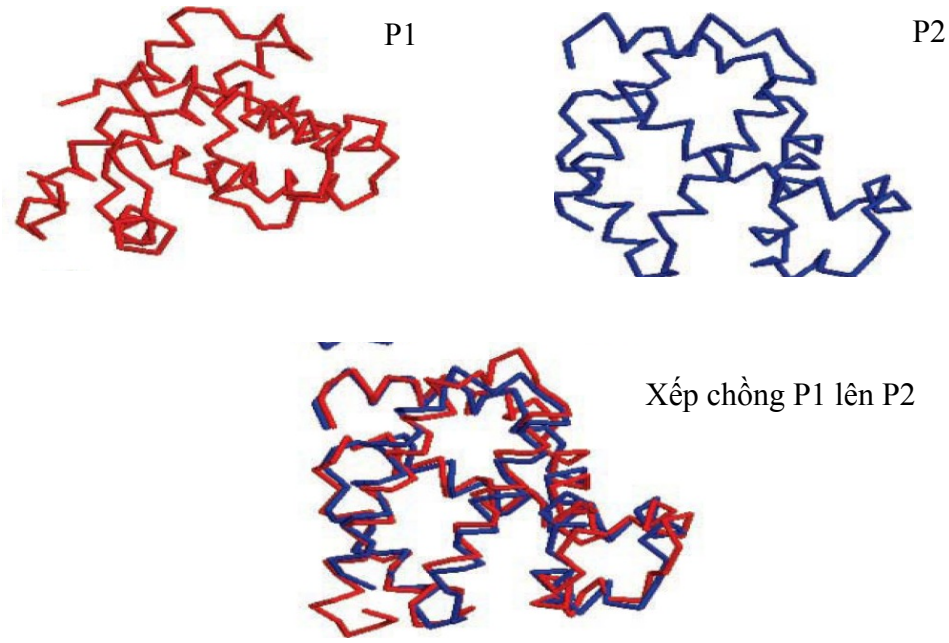
Bài viết được trình bày trong 4 phần; phần thứ nhất giới thiệu về vấn đề cần giải quyết, phần thứ hai trình bày phương pháp được đề xuất để xây dựng thuật toán tính toán; phần thứ ba giới thiệu mẫu dữ liệu để thử nghiệm và phần cuối cùng nêu lên một số kết luận và hạn chế.

2. Phương pháp giải quyết

Xét hai protein P1 và P2. Trong Chimera trình tự đặt ra là sắp xếp cấu trúc (trình tự amino acid) hai protein, rồi sau đó xếp chồng hai protein; tiến hành thay đổi vị trí và thu nhỏ khoảng cách các phân tử để tìm sự tương đồng cấu trúc tốt nhất.

Cách tiếp cận trong bài viết thực hiện theo quy trình ngược lại, việc xếp chồng hai protein được thực hiện trước tiên. Sau đó, tính toán các khoảng cách của các nguyên tử α-carbon được sắp hàng trong hai cấu trúc protein bằng cách thực hiện việc chi tiết hóa về cấu trúc so khớp để giảm thiểu hơn nữa khoảng cách. Phương pháp tổng quát này cho một kết quả sắp hàng tối ưu, có thể tóm tắt như sau:

- Xây dựng một tập các vị trí chồng khớp ban đầu giữa hai cấu trúc cố định bằng cách giữ nguyên một cấu trúc, cấu trúc còn lại được dịch chuyển hoặc xoay để tìm vị trí so khớp tốt nhất.
- Sau khi xếp chồng, xác định các khoảng cách RMSD (*Root Mean Square Deviation*) tối thiểu.
- Tính toán lại khoảng cách giữa các nguyên tử α-carbon



Hình 1. Xếp chồng cấu trúc protein

Phương pháp này sử dụng các vị trí hình học của các nguyên tử α-carbon chính của cấu trúc protein làm dữ liệu đầu vào. Dữ liệu thử nghiệm bao gồm các protein có độ dài khác nhau và tỷ lệ nhận dạng khác nhau. Thuật toán chi tiết được cụ thể qua 2 giai đoạn:

Giai đoạn 1: Xếp chồng cấu trúc

- Giữ cố định P2 và xếp chồng P1 trên P2.
- Tiến hành dịch chuyển P1 để tìm được sự tương đồng cao nhất. Bài toán so sánh cấu trúc của các protein được chuyển thành bài toán so sánh các cấu trúc con giữa hai protein (hình 1).

Giai đoạn 2: Rút ngắn khoảng cách - cực tiểu hóa khoảng cách giữa các nguyên tử được sắp hàng trong protein

2.1 Xếp chồng cấu trúc protein

Gọi x_i là tọa độ ban đầu của nguyên tử thứ *i*, x'_i là tọa độ của nguyên tử thứ *i* sau khi được dịch chuyển và xoay, với *a* là vector tịnh tiến và *R* là ma trận xoay [7][8]:

$$x'_i = a + Rx_i \quad (1)$$

Phương pháp trong Chimera **[6]** được sử dụng để tìm so khớp của các nguyên tử $X_1, ..., X_n$ trong P1 với các nguyên tử $Y_1, ..., Y_n$ trong P2, với điều kiện là P2 được giữ cố định và P1 được chuyển dịch.

2.2 Cực tiểu hóa khoảng cách

Sau khi xếp chồng, việc cực tiểu hóa khoảng cách hai cấu trúc protein dựa trên việc tính toán khoảng cách giữa các nguyên tử α-carbon.

Phương pháp sắp hàng tổng quát là một quá trình ba bước:

Bước 1: Cho D_j là khoảng cách nguyên tử Y_j, $1 \le j \le N$. Việc tính toán D_j là một quá trình bao gồm hai bước:

- Bắt đầu với cấu trúc chồng như mô tả ở trên.
- Tiến hành so khớp nguyên tử Y_j với nguyên tử V_j, trong đó V_j được chọn từ tập (X_{j-1}, X_j, X_{j+1}) để cực tiểu D_j trong công thức (3) và Dst(A, B) là khoảng cách Euclide giữa hai điểm A và B được tính toán theo công thức (2).

$$\varepsilon = \sqrt{\frac{1}{N} \sum_{i=1}^{N} Dst(a + Rx_i, y_i)^2} \quad (2)$$

$$D_j = \min\left\{Dst(X_{j-1}, Y_j),\ Dst(X_j, Y_j),\ Dst(X_{j+1}, Y_j)\right\} \quad (3)$$

Bước 2: Tính toán các khoảng cách giữa mỗi cặp nguyên tử Y_j và V_j theo công thức (4).

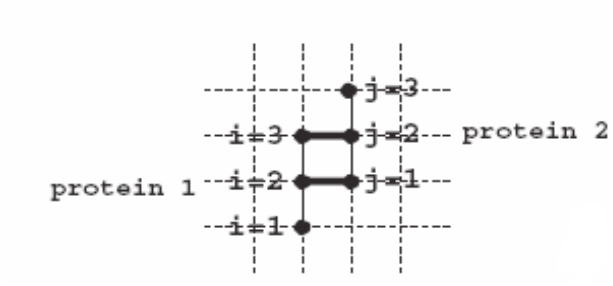
Giả sử V_j có tọa độ (v_j^1, v_j^2, v_j^3) và Y_j có tọa độ (y_j^1, y_j^2, y_j^3). Đối với giá trị T cố định (T là tham số nhiệt độ với giá trị T = 10 để các nguyên tử được ổn định), chúng ta tính toán tất cả các giá trị như sau:

$$d(1,j) = \left|v_j^1 - y_j^1\right|; v_j^{1'} = \frac{e^{-d(1,j)}}{\sum_{i=1}^{N} e^{-d(1,i)T}}$$

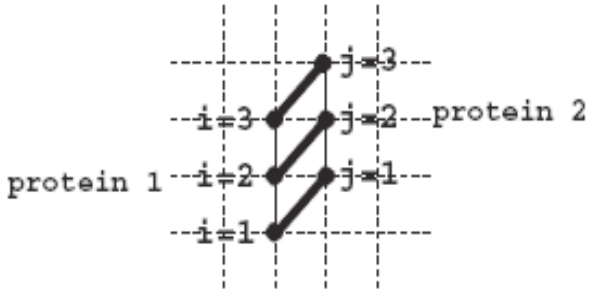
$$d(2,j) = \left|v_j^2 - y_j^2\right|; v_j^{2'} = \frac{e^{-d(2,j)}}{\sum_{i=1}^{N} e^{-d(2,i)T}} \quad (4)$$

$$d(3,j) = \left|v_j^3 - y_j^3\right|; v_j^{3'} = \frac{e^{-d(3,j)}}{\sum_{i=1}^{N} e^{-d(3,i)T}}$$

Trong hình 3 là cấu trúc sắp hàng mới, tốt hơn việc xếp chồng đơn thuần trong hình 2.



Hình 2. Sắp hàng protein thông thường



Hình 3. Sắp hàng protein sau khi tính giá trị

Bước 3: Tính khoảng cách giữa các nguyên tử carbon-α được sắp hàng.

Cho $(Y_1,V_{1'}),(Y_2,V_{2'}),\ldots,(Y_N,V_{N'})$ biểu thị các cặp của các nguyên tử được so khớp.

Trong đó, $V_j = v_j^{1'} + v_j^{2'} + v_j^{3'}$, $1 \le j \le N$ biểu thị khoảng cách tối thiểu tại bước lặp như mô tả ở trên. Khoảng cách dựa trên sắp hàng cấu trúc tổng quát cuối cùng ε_f được tính bởi công thức (5).

$$\varepsilon_f = \sqrt{\frac{1}{N}(v_1 + v_2 + \ldots + v_N)} \qquad (5)$$

3. Dữ liệu mẫu

Dữ liệu mẫu dùng để kiểm tra và mô phỏng được lấy từ ngân hàng protein PDB **[5]**. Mỗi cấu trúc có một số nhận dạng bốn ký tự được gọi là PDB ID hoặc số nhận biết PDB, ví dụ: 2RZS, 1GWB, và được lưu trữ trong một tập tin định dạng *.pdb hoặc *.ent.

Tập tin chứa thông tin về trình tự amino acid, tọa độ của phần tử trong không gian ba chiều v.v… Tọa độ của amino acid và nucleotide trong các protein và acid nucleic được liệt kê thành từng dòng (ATOM). Bài viết tập trung chủ yếu vào tọa độ không gian x, y, z để xác định tọa độ nguyên tử trong không gian - cột (G), (H), (I) của Bảng 1.

Bảng 1: Ví dụ mẫu về mục trong PDB

S.No.	(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)	(I)	(J)	(K)
(1)	ATOM	1	N	MET	A	1	40.184	17.101	24.260	1.00	50.62
(2)	ATOM	2	CA	MET	A	1	38.989	16.442	23.757	1.00	49.62

4. Kết luận

Mặc dù có nhiều phương pháp thực hiện so sánh cấu trúc, nhưng vẫn còn nhiều vấn đề cần nghiên cứu và mở rộng. Phương pháp được trình bày trong bài báo được mở rộng từ phương pháp Chimera. Phương pháp đưa ra được kết quả tối ưu hơn so với cách sắp xếp chồng đơn thuần. Tính toán sự trùng khớp từ việc xếp hàng cấu trúc, rút ngắn khoảng cách hai cấu trúc và tiến hành dịch chuyển, giúp cho việc thể hiện sự tương đồng của protein một cách chính xác hơn.

Tuy nhiên, vẫn còn một số hạn chế gặp phải và chưa giải quyết được như: xử lý định hướng chuỗi liên kết; so sánh nhiều cấu trúc protein tại một thời điểm.

Tài liệu tham khảo

[1] Hashim M., Hashimi A.L., Gorin A., Majumdar A., Gosser Y., Patel D.J. (2002). "Towards structural genomics of RNA: Rapid NMR resonance assignment and simultaneous RNA tertiary structure determination using residual dipolar coupling." J.Mol.Biol, Vol.318, pp. 637-649.

[2] Wikipedia – wikipedia, the free encyclopedia, 2010. [Online]. Available from: http://en.wikipedia.org/wiki/De_novo_protein_structure_prediction

[3] Lonsdale K. (1960). "International tables for X-ray crystallography errata." Acta Cryst, Vol.13, p. 49.

[4] Reddy C.S., Vijayasarathy K., Srinivas E., Sastry G.M., Sastry G.N. (2006). "Homology modeling for membrane proteins: A critical assessment." Computational Biology and Chemistry, Vol.30, pp. 120-126.

[5] Protein Data Bank. http://www.pdb.org/pdb/home/home.do

[6] Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C., Ferrin T.E. (2004). "UCSF Chimera – A visualization system for exploratory research and analysis." J.Comput.Chem, Vol 25, pp.1605-161.

[7] Berthold K. P. Horn. Closed-form solution of absolute orientation us-ing unit quaternions. In Journal of the Optical Society of America, volume 4, pages 629–642, 1986.

[8] Eric W. Weisstein. Rotationmatrix. MathWorld–A Wol-fram Web Resource, 2007. [Online]. Available from: http://mathworld.wolfram.com/RotationMatrix.html [cited 28. 11. 2007]