Tạp chí Phát triển Khoa học và Công nghệ – Khoa học Tự nhiên 2024, 8(1):2828-2837

Open Access Full Text Article

Bài tổng quan

Mô hình hóa cấu trúc protein trên máy chủ đám mây

Lê Mạnh Liêm, Nguyễn Văn Minh Thường, Đinh Thuận Thiên, Trần Văn Hiếu*

TÓM TẮT

Cấu trúc không gian là yếu tố quyết định chức năng của protein. Vì thế, việc giải mã cách protein gấp cuộn đã trở thành bài toán nan giải trong gần nửa thế kỷ qua. Các phương pháp xác định cấu trúc protein có thể kể đến như tinh thể học tia X (X-ray crystallography), chụp cộng hưởng từ hạt nhân (NMR Spectroscopy) hay kính hiển vi điện tử (Electron Microscopy) lại vô cùng tốn kém và mất thời gian do khó khăn trong quá trình tinh thể hóa cấu trúc; trong khi đó, các phương pháp giải trình tự thế hệ mới hiện nay cho phép xác định trình tự của hàng chục ngàn phân tử trong vòng vài ngày. Để rút ngắn khoảng cách chênh lệch giữa trình tự và cấu trúc protein được xác định, các phương pháp và thuật toán dự đoán cấu trúc protein từ trình tự amino acid đã được phát triển. Những năm gần đây, hướng tiếp cận học máy (Machine Learning) đã được cộng đồng các nhà nghiên cứu quan tâm nhờ vào việc phân tích các đặc trưng nội tại của trình tự protein để dự đoán cấu trúc, nhờ đó giúp các công cụ dự đoán hiện nay có thể dự đoán cấu trúc protein với độ chính xác tiệm cận với phương pháp thực nghiệm. Cấu trúc được dự đoán chính xác được ứng dụng cho quá trình thiết kế thuốc, thiết kế kháng thể, tìm hiểu về tương tác protein-protein, và với các phân tử khác. Bài tổng quan cung cấp thông tin về các phương pháp dự đoán như mô hình hoá tương đồng, xâu chuỗi/nhận diện gấp cuộn, dự đoán *ab initio* và một số công cụ hỗ trợ trong việc dự đoán cấu trúc protein.

Từ khoá: công cụ dự đoán cấu trúc, dự đoán ab initio, dự đoán cấu trúc, đánh giá cấu trúc, mô hình hoá tương đồng, tinh chỉnh cấu trúc, xâu chuỗi/nhận diện gấp cuộn

Phòng thí nghiệm Cảm biến Sinh học; Bộ môn Công nghệ Sinh học phân tử - môi trường, Khoa Sinh học - Công nghệ Sinh học, Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM, Việt Nam

Liên hệ

Trần Văn Hiếu, Phòng thí nghiệm Cảm biến Sinh học; Bộ môn Công nghệ Sinh học phân tử - môi trường, Khoa Sinh học - Công nghệ Sinh học, Trường Đại học Khoa học Tự nhiên, ĐHQG-HCM, Việt Nam

Email: tvhieu@hcmus.edu.vn

Lịch sử

- Ngày nhận: 29-10-2022
- Ngày chấp nhận: 27-02-2024
- Ngày đăng: 31-3-2024

DOI :

https://doi.org/10.32508/stdjns.v8i1.1244

GIỚI THIỆU

Cấu trúc protein đầu tiên được xác định là cấu trúc myoglobin với phương pháp tinh thể hóa tia X. Từ đó đến nay đã có nhiều phương pháp khác hỗ trợ xác định cấu trúc protein như chụp cộng hưởng từ hạt nhân (NMR), kính hiển vi điện tử (EM); tuy nhiên, các phương pháp này rất phức tạp, tốn nhiều thời gian và sức lao động; cụ thể đến từ việc tinh thể hóa cấu trúc protein [1,2]. Các protein có cấu trúc bất định (structural disorder), tồn tại xoắn xuyên màng, hay có nhiều vùng uốn và motif coil-coil trong cấu trúc rất khó để tinh thể hóa [3]. Ngược lại, công nghệ giải trình tự thông lượng cao (high-throughput sequencing) ngày càng phát triển [4]; điều này đã tạo nên sự chênh lệch số lượng trình tự-cấu trúc trên cơ sở dữ liệu. Nhằm rút ngắn khoảng cách cấu trúc-trình tự, các phương pháp dự đoán cấu trúc protein đang được quan tâm.

Bài tổng quan này được thực hiện với mong muốn đưa ra cái nhìn tổng quát về những phương pháp dự đoán cấu trúc protein như mô phỏng tương đồng (homology modeling), xâu chuỗi/nhận diện kiểu gấp cuộn (threading/fold recognition), dự đoán *ab initio*, giới thiệu một số công cụ đã và đang được phát triển để phục vụ quá trình dự đoán cấu trúc protein (Hình 1).

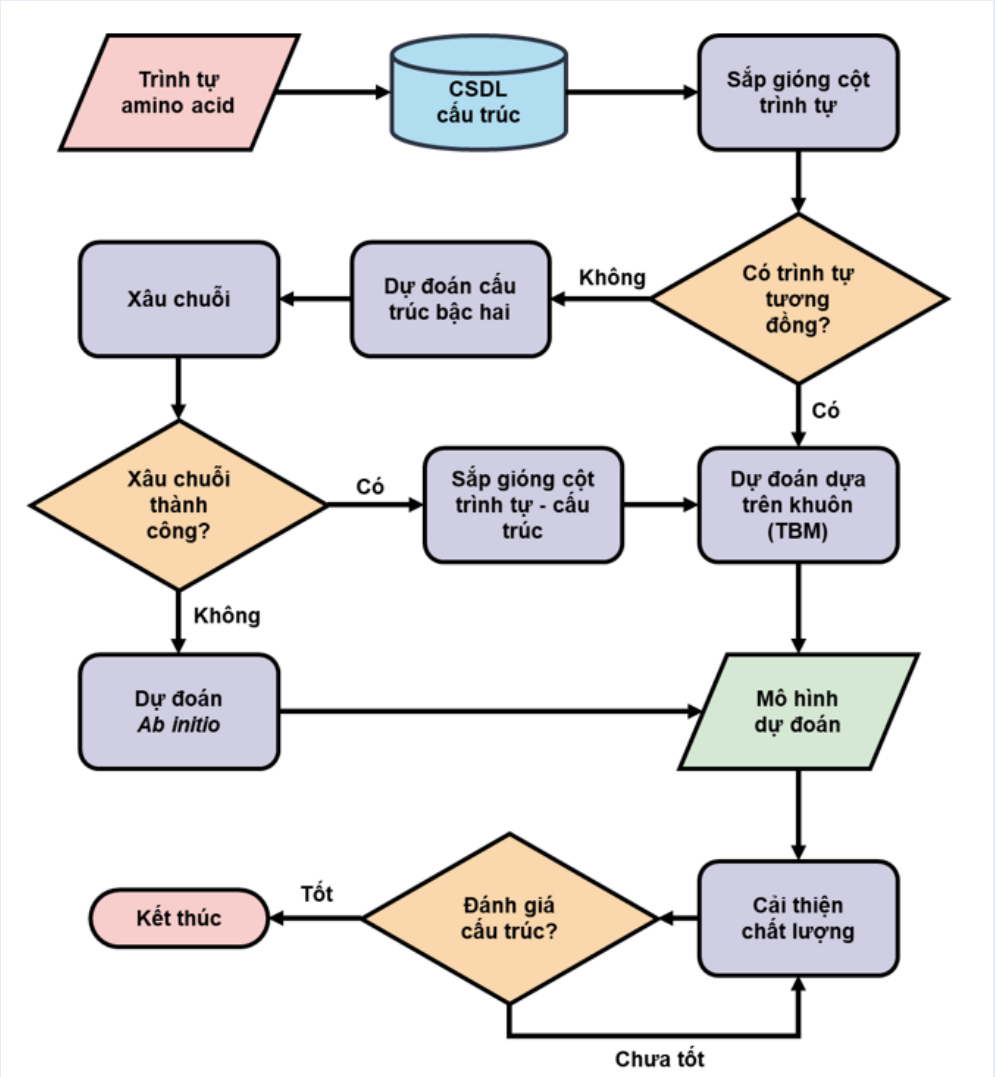
DỰ ĐOÁN CẤU TRÚC PROTEIN DỰA TRÊN KHUÔN (TEMPLATE-BASED METHOD – TBM)

Mô phỏng tương đồng (Homology modeling)

Phương pháp mô phỏng tương đồng được sử dụng lần đầu bởi Browne dựa trên nguyên tắc những protein có độ tương đồng cao hơn 30% thì sẽ có cấu trúc tương tự nhau [5]. Mô hình hóa tương đồng thường được ưu tiên khi cần dự đoán cấu trúc cho một protein, với quá trình dự đoán các bước chính như sau:

- *Tìm kiếm và lựa chọn trình tự tham chiếu*: Trình tự tham chiếu có thể được tìm bằng cách sử dụng các công cụ như BLAST hoặc PSI-BLAST [6] để tìm các trình tự tương đồng trên PDB. Một số tiêu chí để lựa chọn trình tự tham chiếu là độ tương đồng cao, cùng họ protein, có quan hệ tiến hóa gần. Với độ tương đồng thấp hơn 30% thì kết quả sắp gióng cột tự động có thể không tối ưu do các đột biến điểm, đột biến thêm đoạn hoặc mất đoạn [7,8].
- *Mô hình hóa khung sườn*: Có nhiều phương án tiếp cận để mô hình hóa cấu trúc khung sườn

Trích dẫn bài báo này: Liêm L M, Thường N V M, Thiên D T, Hiếu T V. **Mô hình hóa cấu trúc protein trên máy chủ đám mây**. *Sci. Tech. Dev. J. - Nat. Sci.* 2024; 8(1):2828-2837.



Hình 1: Quy trình dự đoán cấu trúc protein

bao gồm tổ hợp khung cứng (rigid body assembly), chia thành phân đoạn ngắn theo tọa độ Cα, tích hợp các thông tin ràng buộc không gian về độ dài và góc liên kết (spatial restraint) và đột biến cấu trúc tham chiếu[9].

- *Mô hình hóa đoạn uốn (loop)*: Mô hình đoạn uốn có thể được mô phỏng dựa trên cấu trúc có sẵn hoặc dự đoán *ab initio* theo phương pháp Monte Carlo hoặc mô phỏng động lực học phân tử[10].
- *Mô hình hóa chuỗi bên*: Các chuỗi bên thường được dự đoán theo phương pháp trích xuất tham chiếu từ thư viện rotamer (rotamer library)[11]. Thư viện này được xây dựng từ các cấu trúc protein có sẵn với độ phân giải cao, với chuỗi bên phù hợp với cấu trúc mục tiêu được chọn dựa theo hàm năng lượng.

Với phương pháp mô phỏng tương đồng, chất lượng mô hình bị ảnh hưởng bởi độ tương đồng trình tự. Khi độ tương đồng giảm xuống khoảng 30-40% thì quá trình sắp gióng cột khó khăn hơn, dẫn đến hệ quả là 80% nguyên tử khung sườn có độ lệch RMSD khoảng 3.5 Å[12].

Hiện nay, một số công cụ cho phép dự đoán mô hình hóa tương đồng có thể kể đến như là SWISS-MODEL và MODELLER.

SWISS-MODEL (*https://swissmodel.expasy.org/*) là một server dự đoán cấu trúc protein dựa trên phương pháp mô phỏng tương đồng bao gồm ba chế độ dự

đoán: tự động, sắp gióng cột, và dự án[13]. Công cụ QMEAN cũng được tích hợp để đánh giá chất lượng mô hình dự đoán[14].

MODELLER (*https://salilab.org/modeller/*) là một phần mềm độc lập (stand-alone program) tương thích với nhiều hệ điều hành. MODELLER cho phép tích hợp các thông tin cưỡng chế không gian (spatial restraints) vào quá trình mô hình hóa cấu trúc và kết quả dự đoán được đánh giá bằng thuật toán DOPE[15]. SWISS-MODEL và MODELLER đều là những công cụ được sử dụng nhiều trong dự đoán cấu trúc protein. Tuy nhiên, SWISS-MODEL lại đem đến cho người dùng giao diện thân thiện và tự động hơn so với giao diện câu lệnh của MODELLER. Mặc dù vậy, MODELLER lại để người dùng theo dõi và điều khiển từng bước trong quá trình mô phỏng cấu trúc.

Phương pháp xâu chuỗi/Nhận diện kiểu gấp cuộn (Threading/Fold Recognition)

Phương pháp xâu chuỗi (threading) ra đời nhằm hỗ trợ dự đoán các protein có độ tương đồng thấp hơn 30%, dựa trên thực tế rằng các trình tự khác biệt hoàn toàn có thể có cấu trúc giống nhau[16]. Phương pháp này còn được biết đến như là phương pháp "so sánh trình tự với cấu trúc".

Các bước chính của phương pháp xâu chuỗi bao gồm:

- *Xây dựng thư viện cấu trúc tham chiếu*: Các cấu trúc và kiểu gấp cuộn protein có thể được thu nhận từ các cơ sở dữ liệu như PDB, CATH, FSSP, và SCOP[17].
- *Thiết kế hàm chấm điểm*: Hàm chấm điểm được thiết kế tốt sẽ bao gồm các tiêu chí về đột biến, tiềm năng tương tác giữa các cặp amino acid, mức độ phù hợp của trình tự với cấu trúc bậc hai, các điểm phạt do mất đoạn/thêm đoạn,…[18]
- *Sắp gióng xâu chuỗi (threading alignment)*: Sắp gióng trình tự mục tiêu với các kiểu gấp cuộn tiềm năng được đánh giá bởi hàm chấm điểm.
- *Xây dựng mô hình 3D*: Cấu trúc được mô hình hóa bắt đầu từ việc chọn các kiểu sắp gióng xâu chuỗi có tiềm năng lớn nhất rồi mô hình hóa dựa theo cấu trúc tham chiếu đã chọn.

Tuy vậy, phương pháp threading vẫn tồn tại nhược điểm chủ yếu nằm ở bản chất cấu trúc protein, cụ thể là tính thoái hóa kiểu cấp cuộn và những thiếu sót về dữ liệu thông tin môi trường khi thiết kế hàm chấm điểm[19].

Một số công cụ cho phép dự đoán cấu trúc protein bằng phương pháp xâu chuỗi/nhận diện gấp cuộn có thể kể đến bao gồm I-TASSER và MUSTER.

I-TASSER (*https://zhanggroup.org/I-TASSER/*) là server dự đoán cấu trúc được phát triển bởi Yang Zhang Lab dựa trên thuật toán LOMETS để tìm kiếm cấu trúc tham chiếu, Monte Carlo để mô phỏng và phân nhóm cấu trạng bằng SPICKER. Các cấu trúc có ΔG thấp tiếp tục được cải thiện chất lượng mô phỏng động lực học phân tử với FG-MD và ModRefiner[20].

MUSTER (*https://zhanggroup.org/MUSTER/*) được xây dựng dựa trên thuật toán xâu chuỗi MUSTER (MUlti-Sources ThreadER) nhận diện tham chiếu từ PDB bằng quy hoạch động. Trong đó, các thông tin về trình tự và cấu trúc được tích hợp vào quá trình tìm kiếm bao gồm: hồ sơ trình tự, cấu trúc bậc hai, hồ sơ cấu trúc phụ thuộc độ sâu, xác suất tiếp xúc dung môi, góc nhị diện khung sườn, và ma trận điểm kị nước[21]. Cả hai công cụ trên đều được phát triển bởi nhóm nghiên cứu Yang Zhang. So với MUSTER dự đoán cấu trúc protein dựa trên việc xâu chuỗi tìm kiếm trình tự và mô phỏng bằng MODELLER, I-TASSER sử dụng thuật toán mô phỏng động lực học tích hợp với khả năng dự đoán tâm hoạt tính và bản thể học của gene (gene ontology).

DỰ ĐOÁN CẤU TRÚC PROTEIN KHÔNG DỰA TRÊN KHUÔN (TEMPLATE-FREE METHOD – TFM)

Thực tế cho thấy một số lượng lớn dữ liệu trình tự không hề tương đồng với những họ protein đã biết trước đó. Vì thế, phương pháp dự đoán không dựa trên khuôn mẫu đã ra đời (*ab initio*). Một thuật toán dự đoán *ab initio* tốt phải thỏa mãn 3 tiêu chí sau: tích hợp hàm năng lượng có độ chính xác cao; phương pháp lấy mẫu hiệu quả; và tiêu chí lựa chọn kết quả tốt nhất từ những cấu trúc dự đoán tiềm năng[22].

Hàm năng lượng

Hàm năng lượng biểu diễn cấu trúc protein dưới dạng các giá trị toán học, từ đó dẫn hướng cho thuật toán dự đoán tìm ra các cấu hình có ΔG thấp nhất. Hàm năng lượng có thể được phân thành 2 nhóm:

- *Hàm năng lượng dựa trên vật lý học*: Dự đoán cấu trúc protein được thực hiện bằng cách kết hợp hàm năng lượng vật lý cơ học lượng tử hay trường lực với các phương pháp lấy mẫu cấu hình nhanh[23–25]. Một số trường lực cơ lý thuyết bao gồm AMBER, CHARMM, OPLS, GROMOS[26].
- *Hàm năng lượng thiết kế từ kiến thức và các cấu trúc phân mảnh có sẵn*: Các hàm này được thiết kế bằng cách khai thác các đặc tính năng lượng không phụ thuộc trình tự và không thuộc trình tự, thế tương tác phụ thuộc khoảng cách nguyên

tử và xu hướng hình thành cấu trúc bậc hai từ các cấu trúc sẵn có trên PDB [27].

Phương pháp lấy mẫu (tìm kiếm) cấu hình

- *Mô phỏng Monte Carlo (MC)*: Mô phỏng MC cổ điển yêu cầu rất nhiều tài nguyên máy tính và dễ gặp sai sót do các trạng thái ổn định giả (metastable state). Một số phương pháp cải tiến để khắc phục hạn chế trên gồm MC trao đổi lặp (REM) và MC tối thiểu hóa (MCM) [28,29].
- *Mô phỏng động lực học phân tử (MDs)*: MDs phản ánh tương đối chính xác quá trình gấp cuộn của protein nhưng lại tốn nhiều tài nguyên và thời gian nên thường xuyên được sử dụng để nghiên cứu sự gấp cuộn của protein [30]. Mô phỏng động lực động lực học phân tử tăng tốc (accelerated molecular dynamics – aMD) là phương pháp lấy mẫu cải thiện từ MD cổ điển bằng cách giảm thiểu rào chắn năng lượng, phân cách các trạng thái khác nhau của hệ thống, giảm thời gian mô phỏng [31].

Phương pháp chọn mô hình kết quả

Có hai nhóm phương án tiếp cận chính để lựa chọn mô hình dự đoán bao gồm phương án dựa trên năng lượng (energy-based method) và phương án dựa trên năng lượng tự do (free-energy based method). Các phương pháp dựa trên năng lượng bao gồm:

- *Chấm điểm dựa trên hàm năng lượng vật lý*: Do năng lượng của cấu hình tự nhiên thấp hơn các cấu trúc mồi (structure decoy), các hàm năng lượng vật lý được sử dụng để hỗ trợ lựa chọn mô hình kết quả [32].
- *Chấm điểm bằng hàm năng lượng dựa trên kiến thức*: Hàm chấm điểm dựa trên kiến thức có thể là hàm năng lượng thô hóa (coarse-grained), trong đó các amino acid được biểu diễn dưới dạng một hoặc một vài nguyên tử đại diện. Cấu trúc dự đoán không thể hoàn toàn giống cấu trúc tự nhiên; vì thế yêu cầu về hàm chấm điểm cho phép phát hiện những cấu hình gần với tự nhiên (near-native) được đặt ra, tuy nhiên việc phát triển các hàm này vẫn còn gặp nhiều khó khăn [33].
- *Chấm điểm dựa trên độ tương hợp giữa trình tự và cấu trúc*: Phương pháp này không phụ thuộc vào năng lượng mà dựa vào độ tương hợp giữa trình tự với cấu trúc. Một trong các cách tiếp cận đầu tiên là ứng dụng điểm xâu chuỗi 3D-1D của Luthy dựa trên việc gán cho từng amino acid các điểm số được tính bởi các thông số môi trường, bao gồm: diện tích vùng kị nước, tỷ lệ chuỗi bên bao phủ bởi nguyên tử O và N, và cấu trúc bậc hai lân cận [34]. Ngoài ra, hàm sai số bình phương của Colovos để mô tả các tương tác không cộng hóa trị giữa các cặp nguyên tử CC, CN, CO, NN, NO và OO cũng được sử dụng [35]. Nhiều phương pháp và chương trình mới đã ra đời để cải thiện phương pháp của Luthy như VERIFY3D và GenTHREADER [22].
- *Phân nhóm các cấu trúc trung gian*: Các nhóm cấu trúc được phân cụm dựa trên ΔG có kích thước lớn thường giống với cấu trúc tự nhiên nhất [36]. Các thuật toán phân cụm phổ biến hiện nay là SPICKER và ROSETTA [37].

Một số công cụ cho phép dự đoán cấu trúc protein bằng phương pháp *ab initio* có thể kể đến bao gồm Phyre2, RoseTTAFold, MetaFold và nổi tiếng nhất hiện nay là AlphaFold2.

Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2) kết hợp nhiều thuật toán dự đoán với nhau, chẳng hạn như HHblits, mô hình hidden Markov và HHsearch để dự đoán cấu trúc protein dựa trên khuôn [38]. Đối với những vùng không có cấu trúc tham chiếu, thuật toán *poing* được sử dụng để mô hình hóa khung sườn của protein; *poing* tăng tốc quá trình gấp cuộn bằng cách sử dụng động lực học Langevin và mô hình dung môi đặc biệt để nhanh chóng đẩy các amino acid kị nước vào trong phần lõi; cuối cùng, các chuỗi bên amino acid sẽ được dựng lên trên phần khung sườn đã được dự đoán [39].

RoseTTAFold (https://github.com/RosettaCommons/RoseTTAFold), phát triển bởi David Baker và cộng sự, được xem như một công cụ tiên phong trong dự đoán cấu trúc protein *ab initio*. RoseTTAFold được xây dựng với kiến trúc mạng neuron 3-track (three-tracked neural network), bao gồm kiến trúc mạng 2-track kết hợp với một track cấu trúc làm việc trên tọa độ khung sườn 3D (Hình 2A). Trong kiến trúc mạng này, thông tin sẽ được di chuyển giữa thông tin trình tự amino acid (1D), bản đồ khoảng cách (2D), và tọa độ cấu trúc (3D) để tìm kiếm và liên kết mối quan hệ giữa trình tự, khoảng cách, và tọa độ. Mô hình 3D của protein được xây dựng bằng cách kết hợp và trung bình hóa các thông tin 1D, khoảng cách 2D, và dự đoán hướng xoay; sau đó, mô hình cuối được tạo dựng thông qua hai hướng tiếp cận: pyRosetta và mô hình SE(3)-Transformer [40].

AlphaFold 2 (https://github.com/google-deepmind/alphafold) là hệ thống trí tuệ nhân tạo phát triển bởi DeepMind được sử dụng trong dự đoán cấu trúc 3D của protein, và được xem là lời giải cho bài toán

gấp cuộn protein[41]. So sánh với RoseTTAFold, AlphaFold2 sử dụng kiến trúc ba module: Module đầu vào: AlphaFold2 tìm kiếm các trình tự tương đồng với trình tự truy vấn trong các cơ sở dữ liệu (UniRef90, UniClust30, BFD,…) và trích xuất các thông tin đồng tiến hóa (co-evolutionary) từ kết quả sắp gióng cột trình tự (multiple sequences alignment – MSA) cũng như ma trận khoảng cách cặp (pairwise distance matrix – PDM). Module Evoformer: Đây được xem như máy phiên dịch các thông tin có được từ module trước bao gồm kết quả MSA và PDM. Lợi ích quan trọng nhất của module này chính là việc nó có thể chuyển đổi thông tin qua lại giữa biểu diễn MSA và PDM; thông tin MSA có thể được giải thích dưới dạng thông tin PDM và ngược lại; từ đó, các thông tin có thể được cải thiện lẫn nhau và đem lại kết quả dự đoán tốt hơn. Module cấu trúc: bao gồm 8 block, module này thu nhận các thông tin có được từ module Evoformer cùng với khung sườn protein để dự đoán cấu trúc 3D của protein. Ngoài ra, AlphaFold2 còn sử dụng thêm cơ chế tái chế (recycling) để tinh chỉnh cấu trúc, tăng độ chính xác của mô hình (Hình 2B). Chính vì vậy, cấu trúc dự đoán từ AlphaFold2 cho ra kết quả có độ tin cậy cực kì cao. So sánh giữa kết quả xác định cấu trúc bằng các phương pháp thực nghiệm và cấu trúc dự đoán cho thấy độ lệch RMSD rất thấp, chỉ từ 1-2 Å (Hình 3).

ESMFold (https://esmatlas.com/resources?action=fold) được phát triển bởi Meta-AI và dựa trên mô hình ngôn ngữ lớn (Large Language Model). ESMFold dựa vào thông tin biểu diễn từ mô hình ngôn ngữ để trực tiếp dự đoán cấu trúc cho trình tự đầu vào, thay vì sử dụng thông tin MSA như AlphaFold2 hay RoseTTAFold. Sự khác biệt về kiến trúc kể trên giúp tiết kiệm tài nguyên CPU dùng cho MSA và giảm số chiều dữ liệu, giúp ESMFold dự đoán nhanh hơn AlphaFold2 từ vài lần đến vài chục lần[42].

ĐÁNH GIÁ VÀ TINH CHỈNH MÔ HÌNH DỰ ĐOÁN

Độ chính xác của mô hình dự đoán là một yếu tố quan trọng cho những nghiên cứu cơ chế phức tạp như thiết kế thuốc, khảo sát gắn protein (protein docking), và dự đoán chức năng protein. Ứng với từng phương pháp xác định hay dự đoán mà cấu trúc ba chiều protein có thể được chấp nhận ứng dụng cho từng nghiên cứu khác nhau (Bảng 1).Tuy nhiên, các khiếm khuyết vẫn có thể xuất hiện, đặc biệt với các cấu trúc dự đoán từ phương pháp không dựa trên khuôn[43].

Biểu đồ Ramachandran biểu diễn các góc quay ψ (psi) và φ (phi) của chuỗi bên protein được sử dụng để đánh giá chất lượng hóa học lập thể của protein. Mỗi tổ hợp quay phi-psi sẽ thuộc một vùng cấu trúc bậc hai nhất định, nếu cấu trúc protein dự đoán là đáng tin cậy thì hầu hết các cặp góc sẽ rơi vào những vùng cho phép trên biểu đồ[44].

SAVES (*https://saves.mbi.ucla.edu/*) là một gói đánh giá chất lượng của mô hình dự đoán được tích hợp nhiều công cụ khác nhau như VERIFY3D, ERRAT, và PROCHECK[45,46].

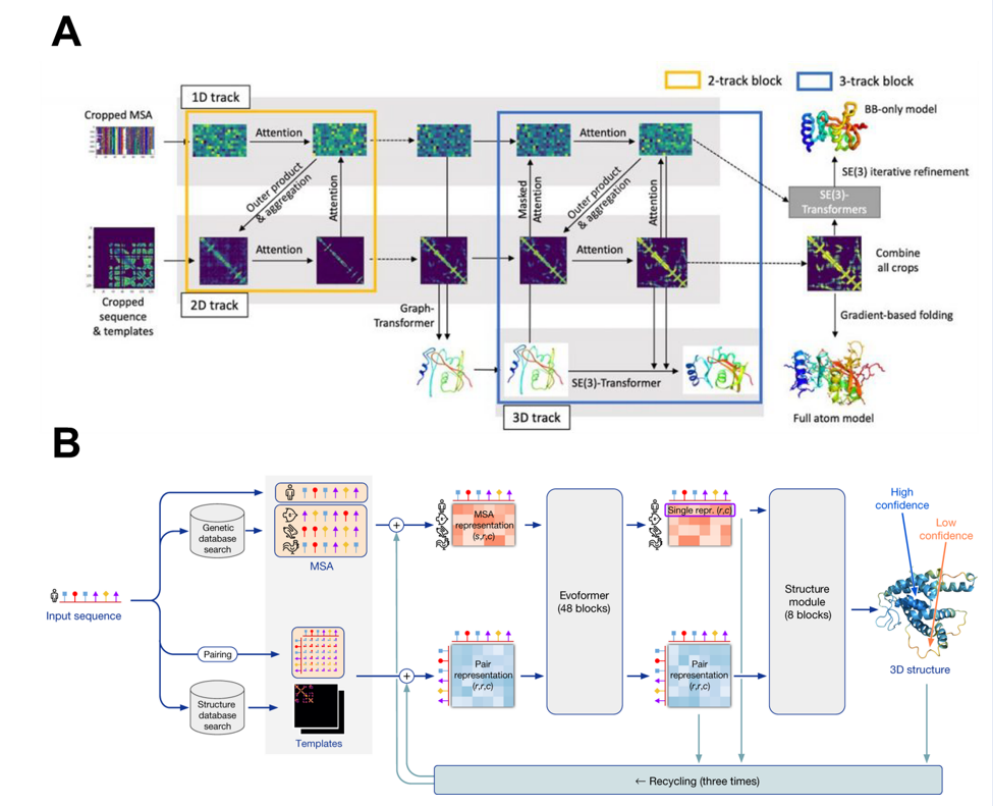
Tinh chỉnh cấu trúc (structure refinement) là bước quan trọng trong quá trình dự đoán để mô hình tiệm cận độ chính xác với cấu trúc thực tế thông qua quá trình điều chỉnh cấu trúc bậc hai và tái đóng gói chuỗi bên[47]. Quá trình tinh chỉnh bao gồm hai bước chính: lấy mẫu và chấm điểm tương tự như mô phỏng cấu trúc.

ModRefiner (*https://zhanggroup.org/ModRefiner/*) cho phép tinh chỉnh protein với độ phân giải cao. Quá trình tinh chỉnh bắt đầu từ Cα đến mô hình hóa mạch chính và toàn bộ nguyên tử. Cả chuỗi bên và mạch chính đều linh động trong mô phỏng tinh chỉnh cấu trúc, trong đó việc tìm kiếm cấu trúc được định hướng bằng trường lực[48].

MỘT SỐ ỨNG DỤNG MÔ HÌNH HÓA CẤU TRÚC

Đại dịch Covid-19 đã mở ra nhiều chiến lược nghiên cứu nhằm ức chế, điều trị, và phòng ngừa sự xâm nhiễm của SARS-CoV-2. Các nhà khoa học đã tìm kiếm và đánh giá khả năng ức chế các protein quan trọng trong việc nhân lên của virus như 3Clpro, Mpro, nsp12, nsp15 và nsp16. Kết quả cho thấy, nhiều hợp chất tổng hợp như saquinavir, aclarubicin, ZG-7 hay từ tự nhiên như absinthin, curcumin đều có tiềm năng ức chế cao với những protein nêu trên[49–52].

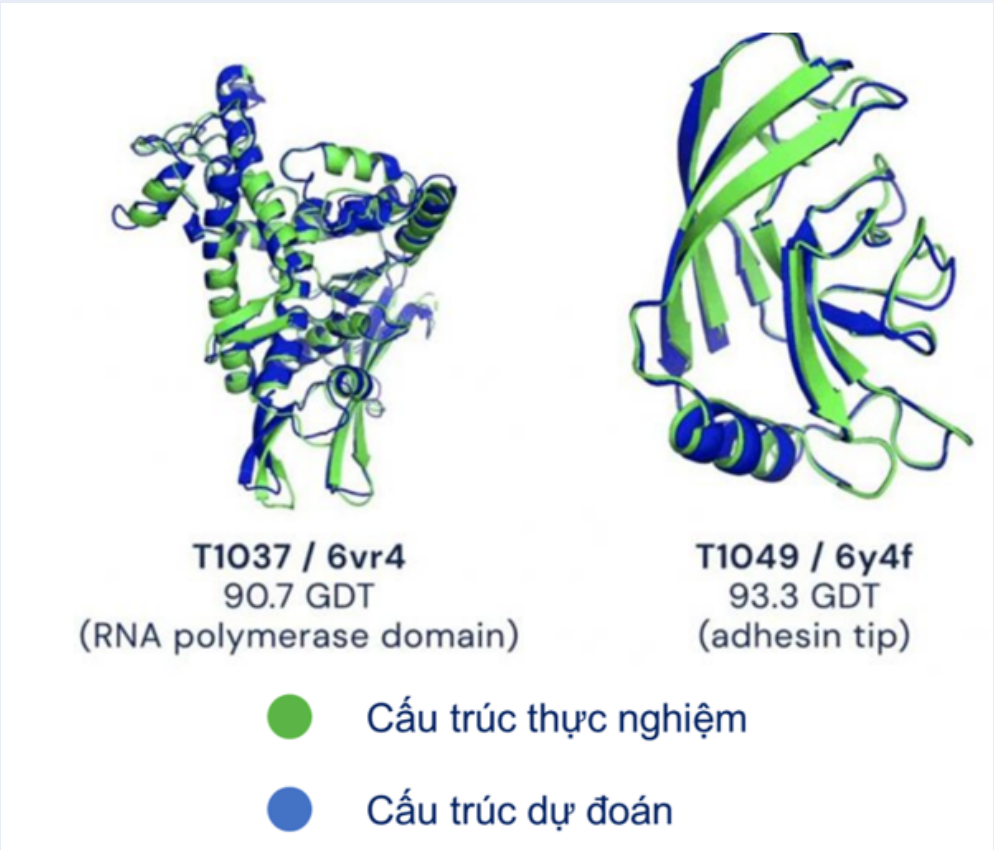
Trong nghiên cứu năm 2020, Kar và cộng sự đã ứng dụng phương pháp mô hình hóa cấu trúc protein để thiết kế và đánh giá vaccine đa epitope ngừa bệnh COVID-19 do virus SARS-CoV-2. Sau khi dự đoán các epitope tiềm năng, nhóm tác giả đã sử dụng webserver trRosetta để dự đoán cấu trúc vaccine và sau đó nghiên cứu tương tác của vaccine với các thụ thể của hệ miễn dịch (Hình 4)[53]. Tương tự trên nghiên cứu của Ariz và cộng sự nhằm thiết kế một loại vaccine cho bệnh đậu mùa khỉ, nhóm tác giả đã sử dụng tổ hợp các phương pháp tin sinh học cấu trúc, bao gồm mô hình hóa cấu trúc protein bằng I-TASSER và RoseTTAFold để mô hình hóa cấu trúc của vaccine[54]. Việc thiết kế kháng thể trung hòa trong bối cảnh đại dịch cũng là một chiến lược bên cạnh công cuộc phát triển vaccine. Bằng việc sử dụng phương pháp mô hình hóa tương đồng và machine learning, sử dụng cấu trúc protein gai của SARS-CoV-1 và hệ thống



Hình 2: Kiến trúc mô hình dự đoán cấu trúc protein của RoseTTAFold (**A**) và AlphaFold2 (**B**)

Bảng 1: Độ chính xác và tiềm năng ứng dụng của cấu trúc protein xác định theo phương pháp thực nghiệm, mô hình hoá tương đồng, xâu chuỗi và dự đoán *ab initio*

Độ tương đồng giữa cấu trúc tham chiếu và mục tiêu	Độ phân giải và độ chính xác của mô hình dự đoán	Phương pháp	Ứng dụng của mô hình dự đoán
100%	< 1 Å (100%)	Tinh thể hóa tia X Cộng hưởng từ hạt nhân (NMR)	Nghiên cứu cơ chế xúc tác Thiết kế và cải thiện phối tử
50%	1.5 Å (95%)	Mô hình hóa tương đồng	Dự đoán phối tử/thụ thể Xác định epitope kháng thể
30%	3.5 Å (80%)	Xâu chuỗi	Nghiên cứu tác động của đột biến điểm Cải thiện chất lượng cho cấu trúc NMR
Không đáng kể	4-8 Å (~80 amino acid được dự đoán chính xác)	Dự đoán ab initio	Kết hợp với bản đồ điện tử có độ phân giải thấp Nhận diện vùng bảo tồn của các amino acid trên bề mặt protein



Hình 3: So sánh độ tương đồng giữa cấu trúc thực nghiệm và cấu trúc dự đoán từ AlphaFold2

AS2TS, năm 2020, Thomas Desautels và cộng sự đã cho ra 20 mô hình kháng thể trung hòa cho SARS-CoV-2, có khả năng liên kết tốt với thụ thể ACE2 từ 89.263 mô hình[55].

KẾT LUẬN

Mối quan tâm về quá trình gấp cuộn và tương tác protein-protein được xem như là vấn đề cốt lõi trong lĩnh vực sinh học cấu trúc. Dự đoán cấu trúc protein bằng điện toán đám mây đã ra đời như một chìa khóa mở ra cơ hội nghiên cứu sâu và rộng hơn cho giới khoa học cũng như rút ngắn chênh lệch giữa số lượng trình tự và cấu trúc xác định. Những thuật toán, hàm số vật lý và công nghệ dữ liệu mới ngày càng phát triển đã góp phần đẩy mạnh công cuộc giải mã cấu trúc protein. Năm 2021, AlphaFold 2 ra đời và được xem như là lời giải cho bài toán gấp cuộn protein đã tồn tại gần 50 năm. Những hiểu biết về cấu trúc đem lại cho các nhà nghiên cứu cái nhìn sâu hơn về cơ chế hoạt động của những đại phân tử sinh học này và ứng dụng cho việc phát triển vaccine, kháng thể, và thiết kế thuốc trong tương lai.

DANH MỤC TỪ VIẾT TẮT

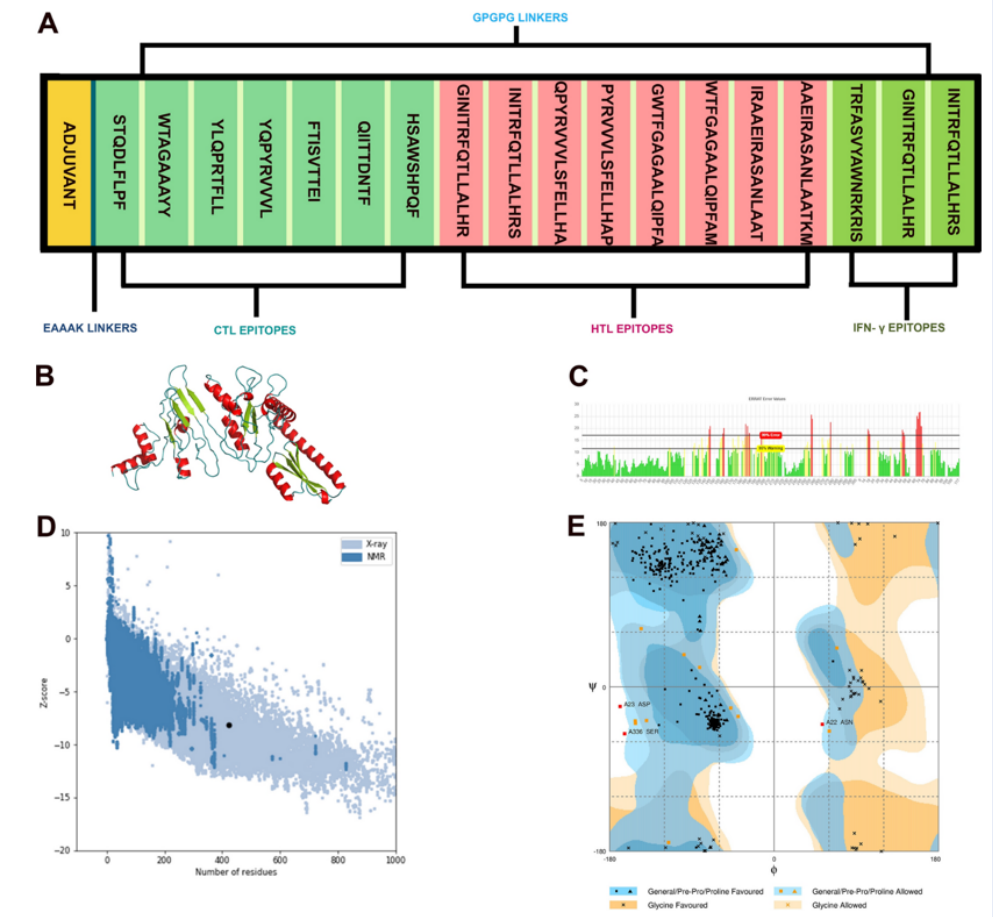
MR : Nuclear Magnetic Resonance
EM: Electron Microscope
CSDL: Cơ Sở Dữ Liệu
PDB: Protein Data Bank
BLAST: Basic Local Alignment Search Tool
PSI-BLAST: Position-Specific Iterative Basic Local Alignment Search Tool
FPPS: Families of Structurally Similar Proteins
SCOP: Structural Classification of Proteins
DOPE: Discrete Optimized Protein Energy
MC: Monte Carlo
REM: Replica Exchange Monte Carlo
MCM: Monte Carlo-minimization
MDs: Molecular Dynamics Simulation
aMD: accelerated Molecular Dynamics

XUNG ĐỘT LỢI ÍCH

Các tác giả cam kết không có xung đột lợi ích.

ĐÓNG GÓP CỦA CÁC TÁC GIẢ

Lê Mạnh Liêm: viết, tổng hợp và chỉnh sửa bản thảo



Hình 4: Kết quả mô phỏng cấu trúc vaccine đa epitope phòng SARS-CoV-2 của Kar và cộng sự. A) Thiết kế vaccine từ các epitope liên tục từ CTL, HTL và IFNγ; B) mô hình 3D của vaccine; C) đánh giá chất lượng cấu trúc thông qua ERRAT; D) đánh giá chất lượng cấu trúc bằng ProSA; E) đánh giá chất lượng lập thể cấu trúc bằng biểu đồ Ramachandran

Nguyễn Văn Minh Thường: viết, tổng hợp và chỉnh sửa bản thảo

Đinh Thuận Thiên: lên ý tưởng, viết, tổng hợp và chỉnh sửa bản thảo

Trần Văn Hiếu: lên ý tưởng, tham gia chỉnh sửa bản thảo và chấp thuận bản thảo.

Tất cả các tác giả đồng ý với bản thảo cuối cùng.

TÀI LIỆU THAM KHẢO

1. Brünger AT. X-ray crystallography and NMR reveal complementary views of structure and dynamics. Nature structural biology. 1997 Oct;4 Suppl:862-5;.
2. Chatham JC, Blackband SJ. Nuclear magnetic resonance spectroscopy and imaging in animal research. ILAR J. 2001;42(3):189-208;.
3. Slabinski L, et al. The challenge of protein structure determination–lessons from structural genomics. Protein Sci. 2007 Nov;16(11):2472-82;.
4. Reuter JA, Spacek DV, Snyder MP. High-Throughput Sequencing Technologies. 2015 May;58:586;.
5. Browne WJ, North AC, Phillips DC, Brew K, Vanaman TC, Hill RL. A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. J Mol Biol. 1969 May 28;42(1):65-86;.
6. Altschul SF, et al. Basic local alignment search tool. 1990;215:403-410;.
7. Yona G, Levitt M. Within the twilight zone: A sensitive profile-profile comparison tool based on information theory. 2002;315:1257-1275;.
8. Peng J, Xu J. A multiple-template approach to protein threading. 2011 Jun;79:1930-1939;.
9. Wallner B, Elofsson A. All are not equal: a benchmark of different homology modeling programs. Protein Sci. 2005 May;14(5):1315-27;.
10. Barozet A, Chacon P, Cortes J. Current approaches to flexible loop modeling. Curr Res Struct Biol. 2021;3:187-191;.
11. Johnson MS, et al. Knowledge-Based Protein Modeling. 2008;29:1-68;.
12. Xiang Z. Advances in Homology Protein Structure Modeling. 2006 Jun;7:217;.
13. Waterhouse A, et al. SWISS-MODEL: homology modelling of protein structures and complexes. 2018 Jul;46:W296;.
14. Bordoli L, et al. Protein structure homology modeling using SWISS-MODEL workspace. 2009;4:1-13;.

15. Shen MY, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2006 Nov;15(11):2507-24;.
16. Floudas CA. Computational methods in protein structure prediction. Biotechnol Bioeng. 2007 Jun 1;97(2):207-13;.
17. Fox NK, Brenner SE, Chandonia JM. The value of protein structure classification information-Surveying the scientific literature. Proteins. 2015 Nov;83(11):2025-38;.
18. Rykunov D, Steinberger E, Madrid-Aliste CJ, Fiser A. Improved scoring function for comparative modeling using the M4T method. J Struct Funct Genomics. 2009 Mar;10(1):95-9;.
19. Skolnick J, Zhou H. Why Is There a Glass Ceiling for Threading Based Protein Structure Prediction Methods? J Phys Chem B. 2017 Apr 20;121(15):3546-3554;.
20. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc. 2010 Apr;5(4):725-38;.
21. Wu S, Zhang Y. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins. 2008 Aug;72(2):547-56;.
22. Lee J, Wu S, Zhang Y. Ab Initio Protein Structure Prediction. In: Rigden DJ, editor. From Protein Structure to Function with Bioinformatics. Springer Netherlands; 2009. pp. 3-25;.
23. Liwo A, Lee J, Ripoll DR, Pillardy J, Scheraga HA. Protein structure prediction by global optimization of a potential energy function. Proc Natl Acad Sci U S A. 1999 May 11;96(10):5482-5;.
24. Liwo A, Khalili M, Scheraga HA. Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. Proc Natl Acad Sci U S A. 2005 Feb 15;102(7):2362-7;.
25. Oldziej S, et al. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. Proc Natl Acad Sci U S A. 2005 May 24;102(21):7547-52;.
26. Lopes PE, Guvench O, MacKerell AD Jr. Current status of protein force fields for molecular dynamics simulations. Methods Mol Biol. 2015;1215:47-71;.
27. Li X, Liang J. Knowledge-Based Energy Functions for Computational Studies of Proteins. In: Xu Y, Xu D, Liang J, editors. Computational Methods for Protein Structure Prediction and Modeling: Volume 1: Basic Characterization. Springer New York; 2007. pp. 71-123;.
28. Kihara D, Lu H, Kolinski A, Skolnick J. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci U S A. 2001 Aug 28;98(18):10125-30;.
29. Li Z, Scheraga HA. Monte Carlo-minimization approach to the multiple-minima problem in protein folding. Proc Natl Acad Sci U S A. 1987 Oct;84(19):6611-5;.
30. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science. 1998 Oct 23;282(5389):740-4;.
31. Hamelberg D, Mongan J, McCammon JA. Accelerated molecular dynamics: a promising and efficient simulation method for biomolecules. J Chem Phys. 2004 Jun 22;120(24):11919-29;.
32. Lazaridis T, Karplus M. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J Mol Biol. 1999 May 7;288(3):477-87;.
33. Skolnick J. In quest of an empirical potential for protein structure prediction. Curr Opin Struct Biol. 2006 Apr;16(2):166-71;.
34. Luthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature. 1992 Mar 5;356(6364):83-5;.
35. Colovos C, Yeates TO. Verification of protein structures: patterns of nonbonded atomic interactions. Protein Sci. 1993 Sep;2(9):1511-9;.
36. Shortle D, Simons KT, Baker D. Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci U S A. 1998 Sep 15;95(19):11158-62;.
37. Bradley P, Misura KM, Baker D. Toward high-resolution de novo structure prediction for small proteins. Science. 2005 Sep 16;309(5742):1868-71;.
38. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols. 2015 Jun;10(6):845-858;.
39. Jefferys BR, Kelley LA, Sternberg MJ. Protein folding requires crowd control in a simulated cell. J Mol Biol. 2010 Apr 16;397(5):1329-38;.
40. Baek M, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021 Aug;373(6557):871-876;.
41. Jumper J, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug;596(7873):583-589;.
42. Lin Z, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023 Mar 17;379(6637):1123-1130;.
43. Adiyaman R, McGuffin LJ. Methods for the Refinement of Protein Structure 3D Models. Int J Mol Sci. 2019 May 9;20(9);.
44. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. Journal of Molecular Biology. 1963;7(1):95-99;.
45. Eisenberg D, Luthy R, Bowie JU. VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol. 1997;277:396-404;.
46. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. Journal of Applied Crystallography. 1993;26(2):283-291;.
47. Heo L, Park H, Seok C. GalaxyRefine: Protein structure refinement driven by side-chain repacking. Nucleic Acids Res. 2013 Jul;41(W1):W384-8;.
48. Xu D, Zhang Y. Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophysical journal. 2011 Nov 16;101(10):2525-34;.
49. Keretsu S, Bhujbal SP, Cho SJ. Rational approach toward COVID-19 main protease inhibitors via molecular docking, molecular dynamics simulation and free energy calculation. Scientific Reports. 2020 Oct 19;10(1):17716;.
50. Gerçek Z, Ceyhan D, Erçağ E. Synthesis and molecular docking study of novel COVID-19 inhibitors. Turkish journal of chemistry. 2021;45(3):704-718;.
51. Joshi T, et al. In silico screening of natural compounds against COVID-19 by targeting Mpro and ACE2 using molecular docking. European review for medical and pharmacological sciences. 2020 Apr;24(8):4529-4536;.
52. Yueniwati Y, et al. Molecular Docking Approach of Natural Compound from Herbal Medicine in Java against Severe Acute Respiratory Syndrome Coronavirus-2 Receptor. Open Access Macedonian Journal of Medical Sciences. 2021;9(A):1181-1186;.
53. Kar T, et al. A candidate multi-epitope vaccine against SARS-CoV-2. Scientific Reports. 2020 Jul 2;10(1):10895;.
54. Aziz S, et al. Contriving multi-epitope vaccine ensemble for monkeypox disease using an immunoinformatics approach. Original Research. 2022-Oct-13;.
55. Desautels T, Zemla A, Lau E, Franco M, Faissol D. Rapid in silico design of antibodies targeting SARS-CoV-2 using machine learning and supercomputing. bioRxiv. 2020;.

Science & Technology Development Journal – Natural Sciences 2024, 8(1):2828-2837

Open Access Full Text Article

Protein structure modeling using cloud-based server

Liem Manh Le, Thuong Minh Van Nguyen, Thien Thuan Dinh, Hieu Van Tran*

ABSTRACT

The structure of a protein plays an important role in determining its function. Deciphering how proteins fold has thus been a puzzle for nearly a half-century. Structure determination methods such as X-ray crystallography, uclear magnetic resonance spectroscopy or electron microscopy are tedious and time-consuming; meanwhile, new-generation sequencing methods are producing massive amounts of protein data in a matter of days. To bridge the protein sequence-structure gap, methods and algorithms for predicting protein 3D structure starting from only its amino acid sequence have been developed. In recent years, machine learning has been of interest to the research community thanks to its work in analyzing intrinsic features of proteins to predict their structure, novel computational methods capable of modeling protein structures to near experimental accuracy. Accurately predicted structure are used for drug and antibody design, understanding protein-protein interactions, and with other molecules. This review provides information about multiple prediction methods and tools for protein structure prediction.

Key words: ab initio modeling, homology modeling, protein modeling tools, structural evaluation, structural prediction, structural refinement, threading/fold recognition

Laboratory of Biosensors; Department of Molecular and Environmental Biotechnology, Faculty of Biology and Biotechnology, University of Science, Vietnam National University – Ho Chi Minh City, Vietnam

Correspondence

Hieu Van Tran, Laboratory of Biosensors; Department of Molecular and Environmental Biotechnology, Faculty of Biology and Biotechnology, University of Science, Vietnam National University – Ho Chi Minh City, Vietnam

Email: tvhieu@hcmus.edu.vn

History

- Received: 29-10-2022
- Accepted: 27-02-2024
- Published Online: 31-3-2024

DOI : https://doi.org/10.32508/stdjns.v8i1.1244

VNUHCM PRESS

Cite this article : Le L M, Nguyen T M V, Dinh T T, Tran H V. **Protein structure modeling using cloud-based server** . *Sci. Tech. Dev. J. - Nat. Sci.* 2024, 8(1):2828-2837.