Vorhersage der Tertiärstrukturen von Proteinen durch Deep-Learning Algorithmen

Name	Sergej Lamert
Studiengang	Angewandte Informatik
Fachsemester	7
Matrikelnummer	00727245
Kurs	Wissenschaftliches Arbeiten 2
Semester	WS 2020/2021
Kursleiterin	Prof. Dr. Christina Bauer

Zusammenfassung

Das Problem der Proteinfaltung ist mittlerweile mehr als fünf Jahrzehnte alt [1]. Die Problematik beinhaltet die Fragen, wie die Aminosäuresequenz, die sog. Primärstruktur, eines Proteins seine dreidimensionale Struktur, auch Tertiärstruktur genannt, vorgibt und wie ein Algorithmus auf die dreidimensionale Struktur schließen kann.
Bislang wurden Proteinstrukturen weitgehend durch praktische Methoden, wie z. B. Kernspinresonanz [2] oder Kristallographie [3] bestimmt. Diese Methoden sind jedoch sehr zeit- und kostenintensiv. Aus diesem Grund wurde der alle zwei Jahre stattfindende Wettbewerb "*Critical Assessment of Techniques for Protein Structure Prediction*" (CASP) im Jahr 1994 initiiert [4]. Bei diesem Wettbewerb müssen die Teilnehmer neben anderen Disziplinen, bisher noch nicht veröffentlichte dreidimensionale Strukturen von Proteinen aus ihren Aminosäuresequenzen vorhersagen.

Bei der Vorhersage von Tertiärstrukturen wird zwischen zwei Methoden unterschieden: dem „*homology modeling*", bei dem eine neue Struktur aus einer bereits bekannten, ähnlichen Proteinstruktur, abgeleitet wird und „*de novo*" Methoden, bei denen die Struktur lediglich aus der Primärstruktur hergeleitet werden soll [5][6]. Die Bestimmung der Vorhersagepräzision erfolgt per „*global distance test*" (GDT), welches ursprünglich im „*Local-Global Alignment*" Programm implementiert wurde [7]. Im Zusammenhang mit CASP werden in diesem Fall die experimentell bestimmten Strukturen und die durch den Algorithmus berechnete Struktur miteinander verglichen und bestimmt, wie sehr diese sich in ihrer dreidimensionalen Struktur ähneln.

Die bisher höchsten Werte, die im CASP per GDT ermittelt wurden, stammen vom AlphaFold [8] Algorithmus und seinem Nachfolger AlphaFold 2 [9]. AlphaFold erreichte beim CASP13 einer Genauigkeit von 68.3 [10]. AlphaFold 2 erreichte im Jahr 2020 mit einem Wert von 92.4 im GDT die Höchste Punktzahl unter den Teilnehmern [11]. Die Trainingsdaten, die AlphaFold 2 benutzt hat, stammen von der „Protein Data Bank" PDB. In dieser sind sowohl die Primärstrukturen als auch die dazugehörigen Tertiärstrukturen von mehr als 150.000 Proteinen in einer Datenbank hinterlegt [12]. Des Weiteren nutzt der Algorithmus Datenbanken wie UniRef90 [13] und BFD [14][15], welche Proteinsequenzen mit unbekannter dreidimensionaler Struktur beinhalten. Als Inputdaten verwendet AlphaFold 2 die Primärstruktur eins Proteins sowie „multiple sequence alignment (MSA) [9]. Der von AlphaFold 2 verwendete Deep-Learning Algorithmus interpretiert die Inputdaten und macht dabei Vorhersagen bezüglich Atomkoordinaten, Torsion sowie die Entfernung von Proteinresten, um somit Aussagen über die dreidimensionale Struktur des Proteins machen zu können [9].

Die Möglichkeit Proteinstrukturen in wesentlich geringerer Zeit zu bestimmen, eröffnet das Potential für eine schnellere Erforschung von Krankheiten, die durch falsch gefaltete Proteine bedingt sind, wie Alzheimer oder Parkinson [16] als mit bisherigen Methoden. Eine weitere, potentielle Anwendungsmöglichkeit besteht in der Möglichkeit Viren wie das SARS-CoV-2 schneller zu erforschen. So gelang es DeepMind das Protein ORF3a des Virus erfolgreich vorherzusagen, welches mit dem experimentellen Ergebnis des Brohawn lab der UC Berkeley übereinstimmt [17].

Abstract

The problem of protein folding remained unsolved for the last five centuries [1]. The problem deals with the questions on how the primary structure of a protein determines its three-dimensional shape and how it can be predicted via computation. The de facto standards in determining the shape was either X-ray crystallography or nuclear magnetic resonance [2][3]. Since both methods mentioned before are quite costly in terms of time and effort, CASP was invented in 1994, an experiment where multiple participants predict the structure of unpublished protein sequences [4]. There are two methods to predict a proteins structure. "Homology modelling" uses known structures to derivate the model of a similar protein. "de novo" methods, use only the primary structure to predict the three-dimensional protein structure [5][6]. The precision of each test gets determined by the "global distance test" (GDT) [7], where the computed structure is compared to the reference structure. At the time of writing, the most precise de novo algorithm is called "AlphaFold 2", which was created by DeepMind. It achieved a score of 92.4 in the GDT at CASP14 [11]. AlphaFold 2 uses labelled data from the protein data bank and unlabelled data from databases like UniRef90 [13] and BFD [14][15]. It feeds all data mentioned into, its neural network to interpret the structure and predict attributes like atom coordinates and torsion of the protein. Being able to predict protein structures via computation has potential to accelerate the research concerning diseases based on protein misfolding like Alzheimer and Parkinson [16].

Literaturverzeichnis

[1] Dill, K. A., & MacCallum, J. L. (2012). The protein-folding problem, 50 years on. Science (New York, N.Y.), 338(6110), 1042–1046. https://doi.org/10.1126/science.1219021

[2] K Wüthrich (1990). Protein structure determination in solution by NMR spectroscopy. Journal of Biological Chemistry, 265(36), 22059–22062.

[3] Ilari, A., & Savino, C. (2008). Protein Structure Determination by X-Ray Crystallography. In J. M. Keith (Ed.), Methods in molecular biology: Vol. 452. Bioinformatics (pp. 63–87). Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-60327-159-2_3

[4] Moult, J., Pedersen, J. T., Judson, R., & Fidelis, K. (1995). A large-scale experiment to assess protein structure prediction methods (Vol. 23). Retrieved from https://zenodo.org/record/1229334 https://doi.org/10.1002/prot.340230303

[5] Zhang, Y. (2008). Progress and challenges in protein structure prediction. Current Opinion in Structural Biology, 18(3), 342–348. https://doi.org/10.1016/j.sbi.2008.02.004

[6] Baker, D., & Sali, A. (2001). Protein structure prediction and structural genomics. Science (New York, N.Y.), 294(5540), 93–96. https://doi.org/10.1126/science.1065659

[7] Zemla, A. (2003). Lga: A method for finding 3D similarities in protein structures. Nucleic Acids Research, 31(13), 3370–3374. https://doi.org/10.1093/nar/gkg571

[8] Andrew W. Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, . . . Demis Hassabis (2020). Improved protein structure prediction using potentials from deep learning. Nature, 577(7792), 706–710. https://doi.org/10.1038/s41586-019-1923-7

[9] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Kathryn Tunyasuvunakool, Olaf Ronneberger, Russ Bates, Augustin Žídek, Alex Bridgland, Clemens Meyer, Simon A A Kohl, Anna Potapenko, Andrew J Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Martin Steinegger, Michalina Pacholska, David Silver, Oriol Vinyals, Andrew W Senior, Koray Kavukcuoglu, Pushmeet Kohli, Demis Hassabis. High Accuracy Protein Structure Prediction Using Deep Learning. In Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book), 30 November - 4 December 2020. Retrieved from https://predictioncenter.org/casp14/doc/CASP14_Abstracts.pdf

[10] Senior, Andrew W.; Evans, Richard; Jumper, John; Kirkpatrick, James; Sifre, Laurent; Green, Tim et al. (2019): Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). In: *Proteins: Structure, Function, and Bioinformatics* 87 (12), S. 1141–1148. DOI: 10.1002/prot.25834.

[11] University of California (2020). TS Analysis: Group performance based on combined z-scores. Retrieved from https://www.predictioncenter.org/casp14/zscores_final.cgi

[12] Goodsell, David S.; Zardecki, Christine; Di Costanzo, Luigi; Duarte, Jose M.; Hudson, Brian P.; Persikova, Irina et al. (2020): RCSB Protein Data Bank: Enabling biomedical research and drug discovery. In: *Protein Science* 29 (1), S. 52–65. DOI: 10.1002/pro.3730.

[13] The UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Research, 38(Database issue), D142-8. https://doi.org/10.1093/nar/gkp846

[14] Steinegger, M., Mirdita, M., & Söding, J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nature Methods, 16(7), 603–606. https://doi.org/10.1038/s41592-019-0437-4

[15] Steinegger, M., & Söding, J. (2018). Clustering huge protein sequence sets in linear time. Nature Communications, 9(1), 2542. https://doi.org/10.1038/s41467-018-04964-5

[16] Hartl, F. U. (2017). Protein Misfolding Diseases. Annual Review of Biochemistry, 86, 21–26. https://doi.org/10.1146/annurev-biochem-061516-044518

[17] John Jumper, Kathryn Tunyasuvunakool, Pushmeet Kohli, Demis Hassabis, and the AlphaFold Team: Computational predictions of protein structures associated with COVID-19. Version 3. Hg. v. DeepMind. Online verfügbar unter https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19.