Protein structure prediction using de novo methods

|  |  |
| --- | --- |
| **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. Das Problem beinhaltet die Fragen, wie die Aminosäuresequenz, die sog. Primärstruktur, eines Proteins seine dreidimensionale Struktur vorgibt und wie ein Algorithmus auf die dreidimensionale Struktur schließen kann. Bislang wurden Proteinstrukturen weitgehend experimentell durch Kernspinresonanz oder Kristallographie bestimmt. Diese Methoden sind jedoch sehr zeit- und kostenintensiv. Aus diesem Grund wurde der alle zwei Jahre stattfindende Wettbewerb “CASP”, “Critical Assessment of Techniques for Protein Structure Prediction”,im Jahr 1994 begonnen.Bei diesem Wettbewerb müssen die Teilnehmer bisher noch nicht veröffentlichte Proteinstrukturen aus ihren Aminosäuresequenzen vorhersagen. Bei der Vorhersage von Proteinstrukturen wird zwischen zwei Methoden unterschieden. Zwischen 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. Die Bestimmung der Vorhersagepräzision erfolgt per „Global distance test“, GDT. 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. Der momentan präziseste de novo Vorhersagealgorithmus ist die zweite Iteration von AlphaFold, welcher von DeepMind entwickelt wurde. Alphafold 2 konnte beim CASP14 im Jahr 2020 mit einer Punktezahl von 92.4 die Höchste Punktzahl unter den Teilnehmern erreichen. Die Trainingsdaten, die AlphaFold 2 benutzt hat, stammen von der „Protein Data Bank“ PDB. In dieser sind die Strukturen von ca. 170000 Proteinen in einer Datenbank hinterlegt. Des Weiteren nutzt der Algorithmus Datenbanken wie UniRef90 und BFD, welche Proteinsequenzen mit unbekannter Struktur beinhalten, zur Mustererkennung im unsupervised learning. Zur Modellierung der gefalteten Proteine nutz AlphaFold 2 räumlicher Graphen, bei denen die Knoten die Proteinresten darstellen und die Kanten die Reste miteinander verbinden, welche nahe beieinander liegen. Das von AlphaFold 2 verwendete Neuronale Netz, interpretiert die Struktur des Graphen und nutzt dabei unter anderem multiple sequence alignment (MSA) um den Graphen zu verfeinern. Die Möglichkeit Proteinstrukturen in wesentlich geringerer Zeit zu bestimmen, eröffnet Möglichkeiten für eine schnellere Erforschung von Krankheiten, die durch falsch gefaltete Protein bedingt sind, wie Alzheimer oder Parkinson als mit bisherigen Methoden?.

**Abstract**

The problem of protein folding remained unsolved for the last five centuries. 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 standard in determining the shape was either X-ray crystallography or Nuclear magnetic resonance. Since both methods mentioned before are quite costly in terms of time and effort, CASP was invented in 1994, an experiment where multiple participant predict the structure of unpublished protein sequences. 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. The precision of each test gets determined by the “global distance test”, 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, which is considered to be as good as experimental results. AlphaFold 2 uses labelled data from the protein data bank and unlabelled data from databases like UniRef90 and BFD. Furthermore it uses spatial graphs to model the protein structure and uses neural networks to interpret the structure. Being able to predict protein structures via computation has potential to accelerate the research concerning diseases based on protein misfolding like Alzheimer and Parkinson.

**Literaturverzeichnis**

(Andrew W. Senior et al., 2020; Baker & Sali, 2001; Dill & MacCallum, 2012; Hartl, 2017; Ilari & Savino, 2008; 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.; K Wüthrich, 1990; Kanelis, Forman-Kay, & Kay, 2001; Keith, 2008; Moult, Pedersen, Judson, & Fidelis, 1995; Steinegger, Mirdita, & Söding, 2019; Steinegger & Söding, 2018; The UniProt Consortium, 2010; University of California, 2020; Zemla, 2003; Zhang, 2008)

References

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

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

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

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

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

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

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

Kanelis, V., Forman-Kay, J. D., & Kay, L. E. (2001). Multidimensional NMR methods for protein structure determination. *IUBMB Life*, *52*(6), 291–302. https://doi.org/10.1080/152165401317291147

Keith, J. M. (Ed.) (2008). *Methods in molecular biology: Vol. 452*. *Bioinformatics*. Totowa, NJ: Humana Press.

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

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

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

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

University of California (2020). TS Analysis: Group performance based on combined z-scores. Retrieved from https://www.predictioncenter.org/casp14/zscores\_final.cgi

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

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