proteinstrukturvorhersage im AlphaFold-zeitalter

Maschinelles Lernen ermöglicht eine „Strukturbiologie für alle“

von Filip König, Karel van der Weg und Holger Gohlke

Um biologische Prozesse zu verstehen und zu beeinflussen, ist die Kenntnis der detailierten räumlichen Struktur von Proteinen von großer Bedeutung. Über die letzten Jahrzehnte hat sich eine große Bandbreite an unterschiedlichsten Methoden zur Vorhersage von Proteinstrukturen etabliert. Aktuelle Entwicklungen in der Proteinstrukturvorhersage mithilfe von maschinellem Lernen ermöglichen akkurate Einblicke ohne aufwändige experimentelle Unterstützung. Dieser Artikel gibt einen Überblick über die verschiedenen Ansätze zur Vorhersage von Proteinstrukturen und die in unserer Gruppe entwickelten Programme zur komplementären Vorhersage von proteinbezogenen Eigenschaften.

Ein großer Anteil molekularer Prozesse in Zellen wird von Proteinen ausgeführt. Beispiele sind die Katalyse von Stoffwechselreaktionen durch Enzyme, die Signaltransduktion und Kommunikation durch Rezeptoren sowie die molekulare Infektionsabwehr durch Antikörper. Um derartig vielfältige Prozesse durch Proteine ablaufen zu lassen, haben sich im Verlauf der Evolution eine große Bandbreite an unterschiedlichen Proteinfaltungen und Strukturmotiven herausgebildet. Die Kenntnis der räumlichen Struktur eines Proteins ist notwendig, um seine Funktionsweise im biologischen Kontext im Detail verstehen zu können. Um beispielsweise die Aktivität eines viralen Enzymes durch rationales Design eines Inhibitors zu blockieren, ist eine genaue räumliche Beschreibung des aktiven Zentrums unerlässlich. Eine Vielzahl an experimentellen Methoden zur Strukturaufklärung von Proteinen ist verfügbar; hierbei stellen die Röntgenkristallographie, Kernresonanzspektroskopie und Kryoelektronenmikroskopie die Methoden mit den bestmöglichen Auflösungen dar. Diese Methoden sind jedoch aufwändig, teuer und ohne Garantie auf Erfolg. Daher hat sich als alternative Möglichkeit zur Erlangung einer Proteinstruktur in den letzten Jahren die computerbasierte Vorhersage etabliert.

Proteinstrukturvorhersage

Um Proteinstrukturen vorherzusagen, wurden in den letzten Jahrzehnten unterschiedlichste Ansätze entwickelt. Eine traditionsreiche Methode ist die Molekulardynamik (MD)- Simulation. Hierbei wird die Energielandschaft eines Proteins mit Hilfe von vereinfachten Potenzialen durchmustert. Während MD-Simulationen in der Beschreibung von Proteinstrukturdynamik eine wichtige Rolle spielen, eignen sie sich nur, um den Faltungsvorgang von kleinen Proteinen ohne weitere kontextuelle Information zu simulieren. Zum einen lassen sich die üblichen Zeitskalen, in denen sich Proteinfaltungen ausbilden, aufgrund der erforderlichen Rechenleistung nur schwer abdecken, zum anderen spielt die intrinsische Ungenauigkeit der Potenziale eine Rolle.

Als vielversprechende Alternative zur Proteinstrukturvorhersage auf Basis einer physikalischen Beschreibung und unter Ausnutzung struktureller, evolutionärer Information hat sich recht früh das Konzept der Homologie-Modellierung etabliert. Hierbei wird die evolutionäre Konservierung der Struktur genutzt: Haben zwei Proteine eine ähnliche Aminosäuresequenz, besitzen sie meistens auch eine ähnliche Faltung. Durch Vergleich mit den Sequenzen aller strukturaufgelösten Proteine in der Protein Data Bank (PDB) können dann akkurate Modelle erzeugt werden. Der Ansatz ist jedoch limitiert, falls keine ho-

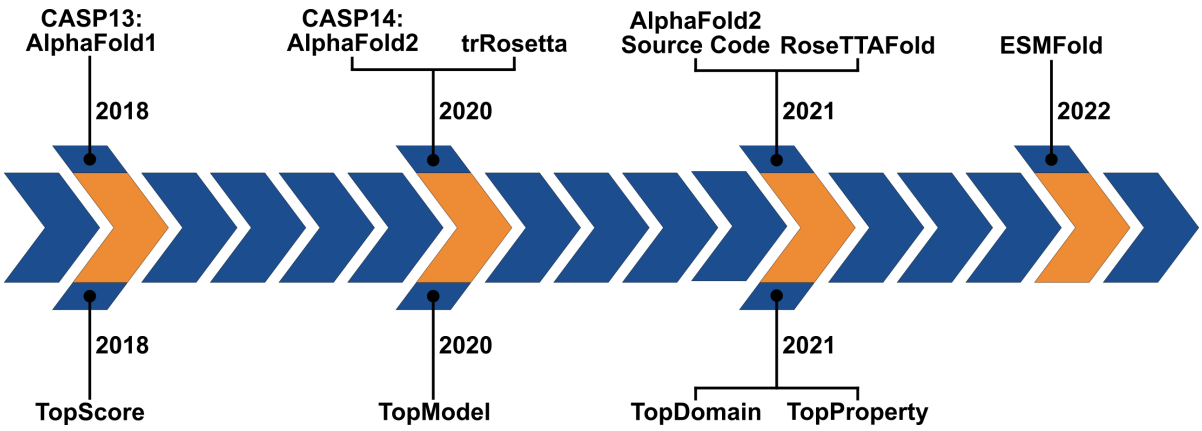

Abbildung 1: Zeitstrahl für die jüngste Entwicklung von Proteinstrukturvorhersageprogrammen (Grafik: Filip König).

mologen Strukturen in der Datenbank vorhanden sind. Das in unserer Gruppe entwickelte Programm TopModel basiert auf dem Homologie-Modellierungsansatz und nutzt obendrein Ansätze maschinellen Lernens.

Um Strukturen in Abwesenheit von aufgelösten, homologen Strukturen vorhersagen zu können, hat sich im letzten Jahrzehnt die Koevolutions-basierte Vorhersage entwickelt. Hierbei wird die Zielsequenz mit allen bekannten Proteinsequenzen aus Datenbanken verglichen. Da es deutlich mehr bekannte Sequenzen als Strukturen gibt, ist die Chance, homologe Sequenzen zu finden, stark erhöht. Zeigen sich für ein Aminosäurepaar in den homologen Sequenzen konzertierte Mutationen, kann dies als Hinweis auf die räumliche Nähe der beiden Aminosäuren im gefalteten Protein gewertet werden. Die Grundidee ist hierbei in der Evolution zu sehen: Mutiert eine Aminosäure sehr häufig zusammen mit einer anderen Aminosäure, kann davon ausgegangen werden, dass deren Interaktion untereinander unerlässlich für die strukturelle Stabilität des Proteins ist. Dies lässt sich als Indikator für eine Interaktion innerhalb des Proteins interpretieren. Lassen sich genug dieser paarweisen Interaktionen finden, können sie als Distanzbeschränkungen zwischen Aminosäuren genutzt und so akkurate Modelle erzeugt werden.

In der jüngsten Vergangenheit haben sich Vorhersagemodelle auf Basis von tiefen neuronalen Netzen (DNN) durchgesetzt. Hierbei werden meist explizit berechnete Kennzahlen wie die Bewertung von gekoppelt auftretenden Mutationen oder die Beibehaltung von Aminosäuren im Verlauf der Evolution als Eigenschaften genutzt und Architekturen wie faltende (konvolutive) neuronale Netzwerke angewendet, um räumliche Nähe von Aminosäuren innerhalb eines Proteins vorherzusagen. Diese Kontakte werden dann, ähnlich wie beim Koevolutions-basierten Vorgehen, in Form von Distanzbeschränkungen genutzt, um den damit eingeschränkten Suchraum nach nativen Konformationen abzusuchen. Beispiele für diese Art von Deep Learning basierten Programmen sind AlphaFold1 und trRosetta (Yang *et al.,* 2020).

Die AlphaFold-Revolution

Ein entscheidender Durchbruch in der Vorhersage von Proteinstrukturen gelang der Google-Tochter DeepMind mit dem Programm AlphaFold2 (Jumper *et al.,* 2021). Im Rahmen des alle zwei Jahre stattfindenden Wettbewerbs CASP, in dem verschiedene Forschungsgruppen versuchen, Strukturen vorherzusagen, die noch nicht veröffentlicht wurden, konnte AlphaFold2 für einen Großteil der Zielstrukturen Vorhersagen treffen, die im Genauigkeitsbereich experimenteller Methoden liegen. Im Vergleich zu früheren *deep learning*-basierten Methoden nutzt AlphaFold2 eine ganze Reihe an Innovationen, die in Kombination zu höherer Genauigkeit führen. Beispielsweise berechnet AlphaFold2 keine expliziten Koevolutionsbewertungen, sondern überlässt es der neuronalen Architektur, aus multiplen Sequenzalignierungen Informationen über die Interaktionen von Aminosäuren innerhalb des Proteins zu extrahieren. Eine weitere wesentliche Innovation ist das End-to-end Konzept: Statt Aminosäurekontakte vorherzusagen und diese anschließend als Distanzbeschränkungen zu nutzen, liefert AlphaFold2 direkt eine Proteinstruktur mit kartesischen Koordinaten. In Analogie zur physikalischen Simulation, nutzt AlphaFold2 vorhergesagte Strukturen erneut als Input, so dass eine iterative Verbesserung erfolgt. Gegenwärtig wurden mit AlphaFold2

bereits mehr als 200 Millionen Strukturen vorhergesagt und öffentlich zugänglich gemacht (https://alphafold.ebi.ac.uk/)

Nach der Veröffentlichung des Quellcodes von AlphaFold2 und der gleichzeitigen Publikation von RoseTTAFold im Sommer 2021 wurden schnell adaptierte Versionen entwickelt mit speziellen Möglichkeiten, Proteinkomplexe vorherzusagen. Eine solche Entwicklung ist ESMFold, das eine ähnliche Grundstruktur wie AlphaFold2 nutzt, jedoch ein sogenanntes *protein language model* anstelle von multiplen Sequenzalignments und damit Strukturvorhersagen in kürzester Laufzeit ermöglicht.

TopSuite

Neben der Strukturvorhersage gibt es eine Reihe weiterer Fragestellungen in Bezug auf Proteine, die sich mit *deep learning* bearbeiten lassen. Hierzu gehören beispielsweise die Bewertung der Qualität von vorhergesagten Strukturmodellen, die Vorhersage von Domänengrenzen in Multidomänenproteinen und die Vorhersage von Transmembrantopologien. Zur Lösung dieser Fragestellungen wurde seit 2017 die Softwaresuite TopSuite in unserer Gruppe entwickelt. Sie enthält Programme zur homologiebasierten Strukturvorhersage (TopModel, Mulnaes *et al.,* 2020), Bewertung der Strukturmodellqualität (TopScore, Mulnaes *et al.,* 2018), Domänengrenzenvorhersage (TopDomain, Mulnaes *et al.,* 2021) und Transmembran-topologievorhersage (TopProperty). Diese Programme sind Metamethoden, d. h. sie nutzen externe Programme, um primäre Vorhersagen zu produzieren, kombinieren diese zu einem normalisierten Satz an Eigenschaften und nutzen dann DNNs um die finalen Vorhersagen zu treffen. Auf diese Weise können die Vorteile von verschiedenen primären Vorhersagemethoden optimal genutzt werden. Alle Programme der TopSuite sind als Webserver verfügbar: https://cpclab.uni-duesseldorf.de/topsuite/.

TopModel

TopModel ist ein Programm zur Vorhersage von Proteinstrukturen, basierend auf dem Konzept der Homologie-Modellierung. Da die Qualität des resultierenden Proteinmodells empfindlich von der tatsächlichen evolutionären Nähe des Templates zum Zielprotein abhängt, wird innerhalb von TopModel eine besondere Gewichtung auf die Erkennung von falsch-positiven Homologen gelegt. TopModel nutzt 12 verschiedene Programme zur Identifikation von homologen Strukturen (*„Templates“*). Anschließend werden vorläufige Modelle auf Basis der identifizierten Templates erzeugt und eine Bewertung mit TopScore durchgeführt. In einem weiteren Schritt werden mit einem DNN-basierten Verfahren falsch-positive Templates identifiziert und entfernt. TopModel nutzt eine Kombination aus Clustering, Bewertung mit TopScore und Strukturverfeinerung zur Erzeugung der finalen Modelle. Beim Vorhandensein geeigneter Homologe ermöglicht TopModel eine sehr akkurate Modellierung von Proteinstrukturen.

Um TopModel mit AlphaFold2 zu vergleichen, wurden beide Methoden auf Proteine aus einem diversen Enzymdatensatz angewendet und die Qualität der resultierenden Modelle in Bezug auf den TopScore verglichen (Abbildung 2). TopScore gibt einen Wert zwischen 0 und 1 zurück, wobei ein niedriger Wert auf eine höhere Modellqualität hindeutet. Dabei zeigt sich, dass AlphaFold2 besonders im niedrigeren TopScore-Bereich Modelle erzeugt, die von TopScore als genauer bewertet werden im Vergleich zu den Modellen die TopModel erzeugt. Auf der anderen Seite zeigt sich für den höheren TopScore-Bereich, dass TopModel dort Modelle erzeugt, die einen besseren TopScore zeigen als die von AlphaFold2.

Insgesamt haben auf maschinellem Lernen beruhende Verfahren einen beträchtlichen Fortschritt bei der Strukturvorhersage von Proteinen hervorgebracht, wobei je nach Fragestellung gezielt angewendete Programme noch einmal Genauigkeitsvorteile liefern können.

Forschungsprojekt in Kurzform

Name:
InCelluloProtStruct – Hybridansatz zur Vorhersage der Supertertiär- und Quartärstruktur von Proteinen und Proteinkomplexen in Zellen.
Fördermaßnahme:
„Computational Life Sciences“ des BMBF
Beteiligte Partner:
AG Prof. Dr. Holger Gohlke, Institut für Pharmazeutische und Medizinische Chemie, Heinrich-Heine-Universität Düsseldorf

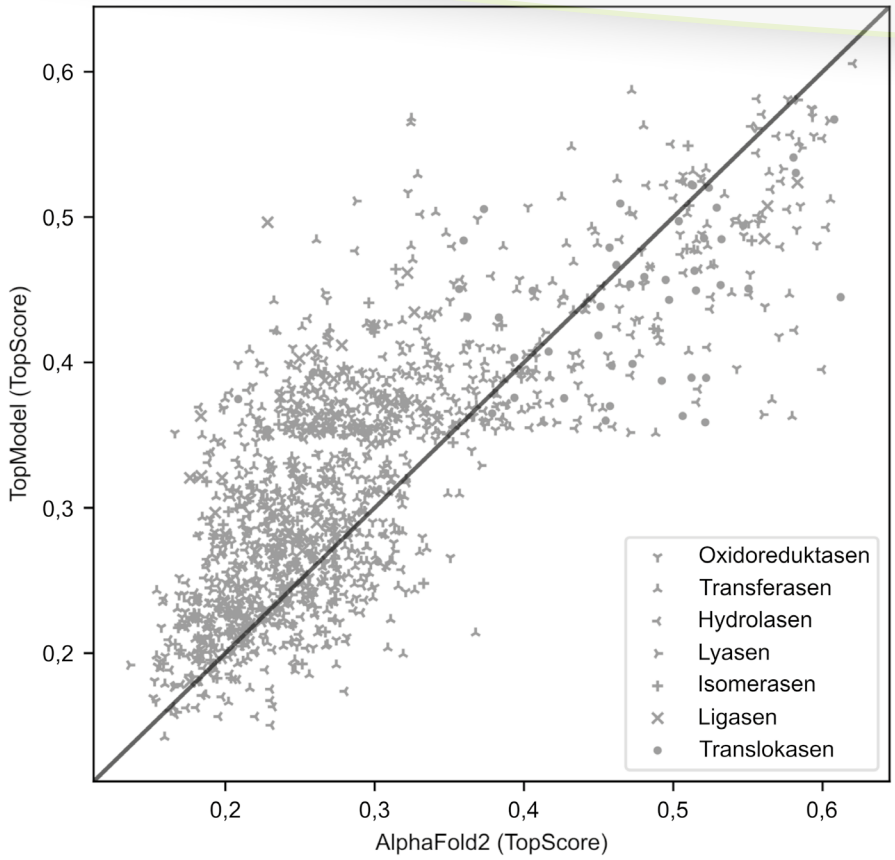

Abbildung 2: Vergleich der TopScore-Werte für Strukturen die mit AlphaFold2 und TopModel vorhergesagt wurden. Abbildung adaptiert auf Basis von: https://www.biorxiv.org/content/10.1101/2022.06.13.495871v1 (Grafik: Karel van der Weg).

Referenzen:

Jumper, J. *et al.,* (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596, 583-589.

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S. und Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proc. Natl. Acad. Sci. 117, 1496-1503.

Mulnaes, D., Porta, N., Clemens, R., Apanasenko, I., Reiners, J., Gremer, L., Neudecker, P., Smits, H.J.S. und Gohlke, H.. (2020). TopModel: template-based protein structure prediction at low sequence identity using top-down consensus and deep neural networks. J. Chem. Theory Comput. 16, 1953-1967.

Mulnaes, D. und Gohlke, H. (2018). TopScore: using deep neural networks and large diverse data sets for accurate protein model quality assessment. J. Chem. Theory. Comput. 14, 6117-6126.

Mulnaes, D., Golchin, P., Koenig, F. und Gohlke, H. (2021). Top-domain: Exhaustive protein domain boundary metaprediction combining multisource information and deep learning. J. Chem. Theory Comput. 17, 4599-4613.

Kontakt:

Prof. Dr. Holger Gohlke
Heinrich-Heine-Universität Düsseldorf
Institut für Pharmazeutische und Medizinische Chemie
Düsseldorf
gohlke@uni-duesseldorf.de

https://cpclab.uni-duesseldorf.de

Foto: Steffen Köhler, HHU Düsseldorf