GEFÖRDERT VOM

Zuwendungsempfänger

Thema der Förderung

Hybrid approach to predict super-tertiary and quaternary structure of proteins and protein complexes in cells (InCelluloProtStruct)

Verantwortlicher

Prof. Dr. Holger Gohlke

Förderkennzeichen

031L0182

Die Verantwortung für den Inhalt dieser Veröffentlichung liegt beim Autor.

Kurzbericht zum Projekt „Hybrid approach to predict super-tertiary and quaternary structure of proteins and protein complexes in cells“ (InCelluloProtStruct), Förderperiode 2019 – 2022 plus kostenneutraler Verlängerung bis 2024.

Das Verständnis der Struktur von Proteinen ist in den Biowissenschaften von entscheidender Bedeutung, da die dreidimensionale Form eines Proteins seine Funktion bestimmt. Proteine sind an praktisch jedem biologischen Prozess beteiligt, von der Katalyse von Reaktionen als Enzyme bis hin zur Regulierung von Zellsignalen und Immunreaktionen. Die genaue Kenntnis der Struktur eines Proteins ermöglicht es, Erkenntnisse darüber zu gewinnen, wie es mit anderen Molekülen interagiert, wie Mutationen zu Krankheiten führen können und wie Medikamente so gestaltet werden können, dass sie auf bestimmte Proteine abzielen. Dieses Wissen ist für den Fortschritt in Bereichen wie Biochemie, Molekularbiologie und Arzneimittelentdeckung unerlässlich. Die experimentelle Bestimmung von Proteinstrukturen ist jedoch oft ein zeitaufwändiger, arbeitsintensiver und kostspieliger Prozess. Techniken wie die Röntgenkristallographie, die Kernspinresonanzspektroskopie (NMR) und die Kryo-Elektronenmikroskopie (Kryo-EM) werden traditionell zur Lösung von Proteinstrukturen eingesetzt. Diese Methoden liefern zwar hochauflösende Details, erfordern aber auch eine komplexe Probenvorbereitung, spezifische Umgebungsbedingungen und teure Geräte.

Um diese Einschränkungen zu überwinden, haben rechnergestützte Ansätze zur Vorhersage von Proteinstrukturen zunehmend an Bedeutung gewonnen. Fortschritte bei Algorithmen und Modellen des maschinellen Lernens, haben gezeigt, dass es möglich ist, Proteinstrukturen mit hoher Genauigkeit vorherzusagen **(Mulnaes, D., Gohlke, H., TopScore: Using deep neural networks and large diverse datasets for accurate protein model quality assessment. J. Chem. Theory Comput. 2018, 14, 6117-6126.) (König, F., van der Weg, K., Gohlke, H., Proteinstrukturvorhersage im AlphaFold-Zeitalter: Maschinelles Lernen ermöglicht eine „Strukturbiologie für alle“. Gesundhyte 2023, 48-50.)**. Diese Methoden bieten eine schnellere und kostengünstigere Alternative zu experimentellen Ansätzen und liefern wertvolle strukturelle Erkenntnisse, ohne dass teure Laborarbeiten erforderlich sind.

Zu den *ungelösten Fragestellungen* bei Antragstellung gehörten

- Die akkurate Vorhersage von Proteinstrukturen bei geringer Sequenzidentität zu verfügbaren Templaten.
- Die Vorhersage von direkten Aminosäurekontakte in Protein-Protein-Interaktionen (PPI).
- Die automatischen Kuration von experimentell aufgeklärten Proteinstrukturen zur Datensatz- und Modellierungsvorbereitung.
- Die Vorhersage der Stöchiometrie von Protein-Protein-Komplexen.

Das *übergeordnete Ziel* des Projektes war es daher, die Entwicklung eines vollautomatischen computergestützten Arbeitsablaufs zur Vorhersage qualitativ hochwertiger Modelle der Struktur von Proteinen und Proteinkomplexen zu entwickeln, wobei Meta-Ansätze zum Einsatz kommen, die mehrere Methoden der rechnerischen Strukturvorhersage und Quellen verfügbarer biologischer Informationen kombinieren. Zusätzlich sollten experimentelle Informationen, z.B. aus SAXS-Experimenten, zur Validierung und Anleitung der Modellbildung verwendet werden.

Wesentliche Ergebnisse

Im *ersten Teilprojekt* wurde TopSuite als automatisierter Arbeitsablauf zur Proteinstruktur und –eigenschaftsvorhersage als Programmbundle entwickelt. Zunächst wurde TopModel als Metamethode zur templatbasierten Proteinstrukturvorhersage fertig entwickelt, wobei es zur prospektiven Vorhersage mit anschließender experimenteller Validierung von LipoP aus *C. difficile* unter Verwendung von SAXS-Daten verwendet wurde **(Mulnaes, D., Porta, N., Clemens, R., Apanasenko, I., Reiners, J., Gremer, L., Neudecker, P., Smits, S.H.J., Gohlke, H., TopModel: Template-based protein structure prediction at low sequence identity using top-down consensus and deep neural networks. J. Chem. Theory Comput. 2020, 16, 1953-1967.)**. TopModel-Vorhersagen wurden seitdem erfolgreich in Projekten in Zusammenarbeit mit experimentellen Partnern eingesetzt, wobei die Genauigkeit der vorhergesagten Modelle durch experimentelle Validierung bestätigt wurde **(Nutschel, C.,**

Coscolín, C., Mulnaes, D., David, B., Ferrer, M., Jaeger, K.-E., Gohlke, H., Promiscuous esterases counterintuitively are less flexible than specific ones. J. Chem. Inf. Model. 2021, 65, 5, 22383-2395.). Bemerkenswerterweise ist die Qualität der von TopModel vorhergesagten Strukturen nur um 4% schlechter als solcher von AlphaFold2 **(van der Weg, K., Gohlke, H., TopEnzyme: A framework and database for structural coverage of the functional enzyme space. Bioinformatics 2023, 39, btad116.)**. TopSuite wurde zudem um die DNN-basierten Metamethoden TopDomain **(Mulnaes, D., Golchin, P., Koenig, F., Gohlke, H., TopDomain: Exhaustive protein domain boundary meta-prediction combining multi-source information and deep learning. J. Chem. Theory Comput. 2021, 17, 4599–4613.)** und TopProperty **(Mulnaes, D., Schott-Verdugo, S., Koenig, F., Gohlke, H., TopProperty: Robust meta-prediction of transmembrane and globular protein features using deep neural networks. J. Chem. Theory Comput. 2021, 17, 7281–7289.)** zur Vorhersage von Proteindomänengrenzen sowie Sekundärstruktur und Membrantopologie erweitert. Um die Nutzung für eine breite wissenschaftliche Öffentlichkeit außerhalb der Arbeitsgruppe zu ermöglichen, wurde ein Webserver für die Programme von TopSuite implementiert **(Mulnaes, D., Koenig, F., Gohlke, H., TopSuite webserver: A meta-suite for deep learning-based protein structure and quality prediction. J. Chem. Inf. Model. 2021, 61, 548–553.)** (https://cpclab.uni-duesseldorf.de/topsuite/). Zudem wurden die Proteinstruktur- und eigenschaftsmodellierungskenntnisse dahingehend verwendet, um Vasor zu entwickeln, welches die Vorhersage von Mutationseffekten im Multidrug Resistance Protein 3 (MDR3) ermöglicht **(Behrendt, A., Golchin, P., König, F., Mulnaes, D., Stalke, A., Dröge, C., Keitel, V., Gohlke, H. Vasor: Accurate prediction of variant effects for amino acid substitutions in MDR3. Hepatol. Commun. 2022, 6, 3098–3111.)**. Auch hierzu wurde ein Webeserver entwickelt (https://cpclab.uni-duesseldorf.de/mdr3_predictor/), da eine einfache Zugänglichkeit insbesondere für Kliniker wichtig ist.

Im *zweiten Teilprojekt* wurde initial mit TopInterface ein DNN-basiertes Programm zu entwickeln, das direkte Aminosäurekontakte in Protein-Protein-Interaktionen (PPI) vorhersagen kann **(Koenig, F., Mulnaes, D., van der Weg, K., Gohlke, H., TopSuite: Collection of Deep Learning-based Metamethods to Predict and Evaluate Protein Structures and Properties. in: "Proceedings of the NIC Symposium 2022", Müller, M., Ch. Peter, Trautmann, A. (eds.), Jülich, 2022, 47-54.)**. Unser Ansatz belief sich darauf, analog zur Strukturvorhersage von Proteinmonomeren, koevolutionsbasierte Aminosäurekopplungen aus einem gepaarten multiplen Sequenzalignment (MSA) zu extrahieren und diese als Distanzbeschränkungen für physik-basiertes Docking zu nutzen. Um ein derartigen MSA zu konstruieren, müssen Sequenzen aus den MSA der Monomere gepaart werden. Hierfür entwickelten wir drei Verfahren. Zur Repräsentation der lokalen Nachbarschaft der Aminosäuren wurden 2D-Convolutional Neural Networks genutzt, wobei die Eigenschaften in den Kanaldimensionen organisiert und die Aminosäurepaare in Distanzklassen eingeteilt wurden. Durch einen vorläufigen Performancevergleich wurde ersichtlich, dass TopInterface mit der implementierten Architektur nicht die Genauigkeit von AlphaFold-Multimer erreicht. In der Folge wurde die ursprüngliche Entwicklung von TopInterface als PPI-Vorhersageprogramm abgebrochen. Ggw. verfolgen wir für TopInterface die Möglichkeit, durch Implementierung eines Verfahrens zur effizienten Paarung von Homologen für Proteine aus verschiedenen Spezies die Qualität der AlphaFold-Multimer Vorhersagen deutlich zu verbessern. Entsprechend haben wir die Zielsetzung für TopInterface angepasst.

Im *dritten Teilprojekt* entwickeln wir die Software InVivoPDB, die durch automatische Kuration von experimentell aufgeklärten Proteinstrukturen den Zugang zu hochwertigen Strukturdaten ermöglicht. Sie beinhaltet Programmteile zur Identifikation und Entfernung von Nicht-Protein-Strukturbestandteilen, wie z.B. Kristallisationsadditiven, Antikörpern oder Aufreinigungstags. Unvollständige Moleküle in der Kristallstruktur, wie beispielsweise organische Liganden, Nuklein- oder Aminosäuren, können automatisch „repariert“ und vervollständigt werden. Fehlende Proteinregionen, werden automatisch modelliert. Weiterhin steht ein umfangreicher Programmteil zur Strukturoptimierung mit Hilfe von MD-Software zur Verfügung, um energetisch ungünstige Konformationen zu identifizieren und zu beseitigen.

Im *vierten Teilprojekt* verfolgen wir die Frage, die korrekte Stöchiometrie von beteiligten Proteinkomplexen vorherzusagen, um den biologischen Kontext von Proteinen akkurater modellieren zu können. Hierfür wird mithilfe einer modernen Neural Network Architektur, wie zum Beispiel Transformern oder Graph Neural Networks, ausgetestet, ob eine akkurate Vorhersage basierend auf bekannten Komplexstrukturen möglich ist.

Eingehende Darstellung zum Projekt „Hybridansatz zur Vorhersage der Supertertiär- und Quartärstruktur von Proteinen und Proteinkomplexen in Zellen“ (InCelluloProtStruct), Förderperiode 2019 – 2022 plus kostenneutraler Verlängerung bis 2024.

Inhaltsverzeichnis

Inhalt

Inhaltsverzeichnis 3
Übergeordnetes Ziel des Projektes 4
Ausgangspunkt in Form eigener Entwicklungen 4
Projekt 1: Entwicklung eines automatisierten Arbeitsablaufs zur Proteinstrukturvorhersage in Form eines Webservers 4
Ziele des Projekts 4
Ergebnisse 4
Relevante Publikationen 5
Projekt 2: Entwicklung eines Metaansatzes zur Protein-Protein-Kontakt-Vorhersage 6
Ziele des Projekts 6
Ergebnisse 6
Relevante Publikationen 7
Projekt 3: Entwicklung einer Software zur automatischen Kuration von experimentell aufgeklärten Proteinstrukturen 7
Ziele des Projekts 7
Ergebnisse 7
Relevante Publikationen 8
Projekt 4: Entwicklung einer DNN-basierten Stöchiometrievorhersage von Proteinkomplexen basierend auf der Struktur 8
Ziele des Projekts 8
Ergebnisse 8
Relevante Publikationen 9
Wichtigste Positionen des zahlenmäßigen Nachweises 10
Zeitplan des Vorhabens 10
Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten 10
Voraussichtlicher Nutzen und Verwertbarkeit sowie kurzfristige Planung 10
Während der Durchführung des Vorhabens dem Zuwendungsempfänger bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen 10
Erfolgte oder geplante Veröffentlichungen des Ergebnisses nach Nr. 5 der NABF 11

Übergeordnetes Ziel des Projektes

Ziel dieses Projekts war die Entwicklung eines vollautomatischen computergestützten Arbeitsablaufs zur Vorhersage qualitativ hochwertiger Modelle der Struktur von Proteinen und Proteinkomplexen, wobei Meta-Ansätze zum Einsatz kommen, die mehrere Methoden der rechnerischen Strukturvorhersage und Quellen verfügbarer biologischer Informationen kombinieren. Zusätzlich sollten experimentelle Informationen, z.B. aus SAXS-Experimenten, zur Validierung und Anleitung der Modellbildung verwendet werden.

Ausgangspunkt in Form eigener Entwicklungen

Wir hatten TopScore, eine neue Metamethode für die Bewertung der Qualität von Proteinmodellen, entwickelt (Mulnaes et al., 2018), die basierend auf einer sehr großen (910 Ziele, ~$1{,}7\times10^5$ Modelle und ~$2{,}9\times10^7$ Reste) und vielfältigen Menge von Proteinstrukturen als Trainingsdaten eine konstant hohe Vorhersageleistung gezeigt hat. TopScore verwendet eine Reihe von *deep neural networks* (DNN), um die Ergebnisse von fünfzehn verschiedenen primären Prädiktoren zu vereinen und zeigte eine signifikante Verbesserung gegenüber jedem einzelnen. Zudem hatten wir einen Prototypen einer neuen Metamethode für die templatbasierte Proteinstrukturvorhersage entwickelt (TopModel), eine Pipeline, die aus fünf verschiedenen Modulen besteht, TopThreader, TopAligner, TopBuilder, TopScore, und TopRefiner. TopModel hat eine hohe Leistung bei der Templatauswahl, dem Alignment und der Modellverfeinerung gezeigt, als es mit den CASP10-, CASP11- und CASP12-Datensätzen verglichen wurde.

Projekt 1: Entwicklung eines automatisierten Arbeitsablaufs zur Proteinstrukturvorhersage in Form eines Webservers

Ziele des Projekts

Ziele des Projektes waren die Entwicklung eines automatisierten Arbeitsablaufs zur Proteinstrukturvorhersage durch Ergänzung der bisherigen Entwicklungen von TopScore und TopModel und die zur Verfügungstellung für die wissenschaftliche Öffentlichkeit als Webserver.

Ergebnisse

TopSuite wurde als automatisierter Arbeitsablauf zur Proteinstruktur und –eigenschaftsvorhersage zunächst als Programmbundle entwickelt. Zunächst wurde TopModel als Metamethode zur templatbasierten Proteinstrukturvorhersage fertig entwickelt (Abbildung 1A, B), wobei es zur prospektiven Vorhersage mit anschließender experimenteller Validierung von LipoP aus *C. difficile* unter Verwendung von SAXS-Daten verwendet wurde. TopModel-Vorhersagen wurden seitdem erfolgreich in Projekten in Zusammenarbeit mit experimentellen Partnern eingesetzt, wobei die Genauigkeit der vorhergesagten Modelle durch experimentelle Validierung bestätigt wurde (Nutschel et al., 2021, Twizerimana et al., 2023, Lin et al., 2023). Bemerkenswerterweise ist die Qualität der von TopModel vorhergesagten Strukturen nur um 4% schlechter als solcher von AlphaFold2 (Van der Weg et al., 2023). TopSuite wurde zudem um die DNN-basierten Metamethoden TopDomain und TopProperty zur Vorhersage von Proteindomänengrenzen sowie Sekundärstruktur und Membrantopologie erweitert. Die Programme der TopSuite sind einzelnen zum Download und zur lokalen Installation verfügbar. (https://cpclab.uni-duesseldorf.de/index.php/Software). Um die Nutzung für eine breite wissenschaftliche Öffentlichkeit außerhalb der Arbeitsgruppe zu ermöglichen, wurde ein Webserver für die Programme von TopSuite implementiert (https://cpclab.uni-duesseldorf.de/topsuite/) (Abbildung 1C). Zudem wurden die Proteinstruktur- und eigenschaftsmodellierungskenntnisse dahingehend verwendet um Vasor zu entwickeln, welches die Vorhersage von Mutationseffekten im Multidrug

Resistance Protein 3 (MDR3) ermöglicht. Auch hierzu wurde ein Webeserver entwickelt (https://cpclab.uni-duesseldorf.de/mdr3_predictor/), da eine einfache Zugänglichkeit insbesondere für Kliniker wichtig ist.

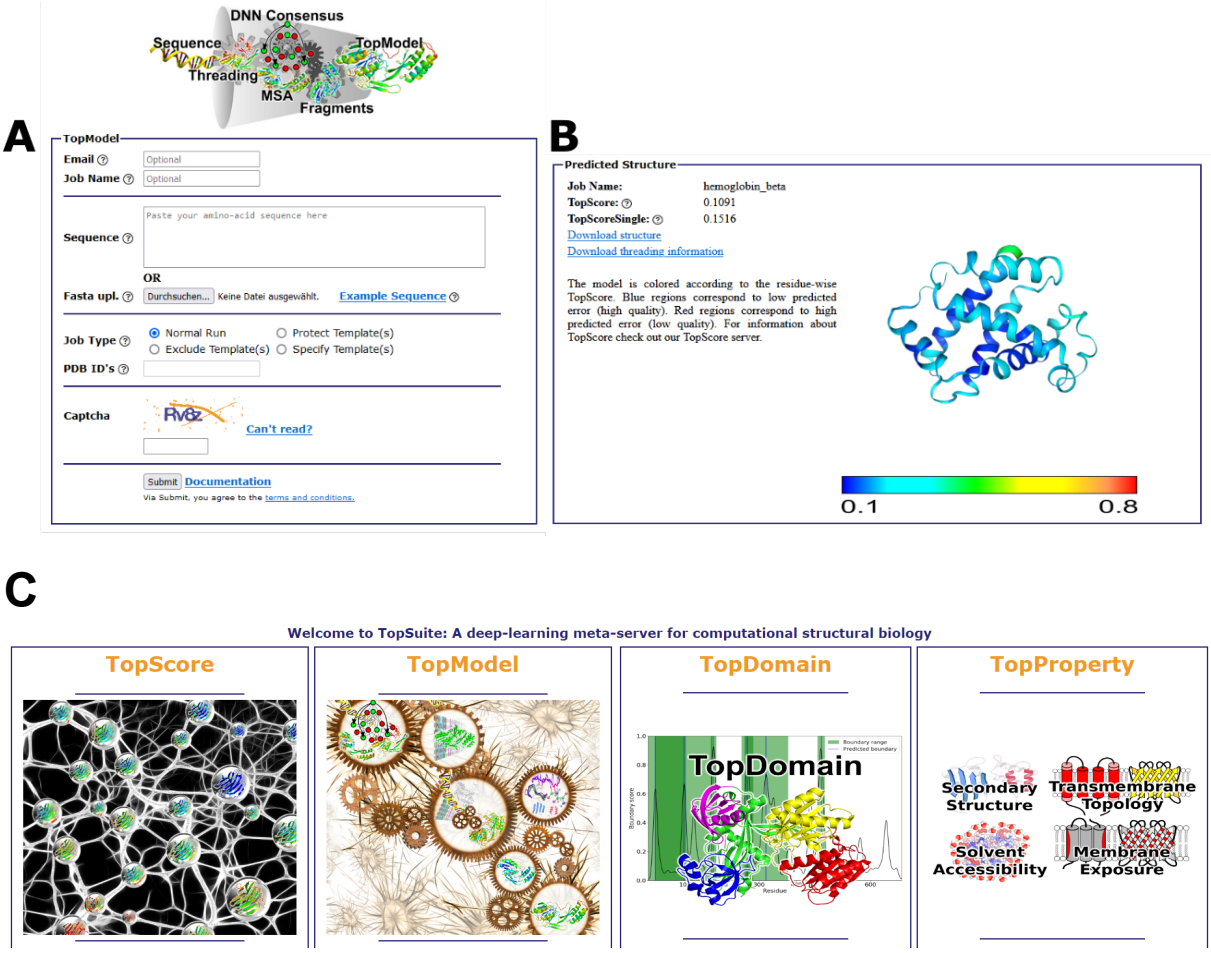

Abbildung 1: Userinterface des TopSuite-Servers. (A) Eingabeformular für die Vorhersage einer Proteinstruktur mittels des TopModel-Servers. Es besteht die Option, konkrete Strukturtemplate auszuwählen, die für die Vorhersage genutzt werden sollen. (B) Ergebnisvisualisierung der vorhergesagten Proteinstruktur für die ß-Kette des Hämoglobins. Die Struktur wird farblich gemäß des TopScores codiert, wobei blaue Bereiche auf eine hohe Vorhersagegenauigkeit und rote Bereiche auf eine niedrigere Genauigkeit hinweisen. (C) Gesamtheit aller im Rahmen der TopSuite verfügbaren Webserver (https://cpclab.uni-duesseldorf.de/topsuite/).

Relevante Publikationen

Mulnaes, D., Porta, N., Clemens, R., Apanasenko, I., Reiners, J., Gremer, L., Neudecker, P., Smits, S.H. and Gohlke, H., 2020. TopModel: template-based protein structure prediction at low sequence identity using top-down consensus and deep neural networks. *Journal of Chemical Theory and Computation*, *16*(3), pp.1953-1967.

Nutschel, C., Coscolín, C., David, B., Mulnaes, D., Ferrer, M., Jaeger, K.E. and Gohlke, H., 2021. Promiscuous esterases counterintuitively are less flexible than specific ones. *Journal of Chemical Information and Modeling*, *61*(5), pp.2383-2395.

Mulnaes, D., Golchin, P., Koenig, F. and Gohlke, H., 2021. TopDomain: exhaustive protein domain boundary metaprediction combining multisource information and deep learning. *Journal of Chemical Theory and Computation*, *17*(7), pp.4599-4613.

Mulnaes, D., Schott-Verdugo, S., Koenig, F. and Gohlke, H., 2021. TopProperty: Robust metaprediction of transmembrane and globular protein features using deep neural networks. *Journal of Chemical Theory and Computation*, *17*(11), pp.7281-7289.

Mulnaes, D., Koenig, F. and Gohlke, H., 2021. TopSuite web server: a meta-suite for deep-learning-based protein structure and quality prediction. *Journal of Chemical Information and Modeling*, *61*(2), pp.548-553.

Behrendt, A., Golchin, P., König, F., Mulnaes, D., Stalke, A., Dröge, C., Keitel, V. and Gohlke, H., 2022. Vasor: Accurate prediction of variant effects for amino acid substitutions in multidrug resistance protein 3. *Hepatology Communications*, *6*(11), pp.3098-3111.

König, F., van der Weg, K., Gohlke, H. 2023. Proteinstrukturvorhersage im AlphaFold-Zeitalter: Maschinelles Lernen ermöglicht eine „Strukturbiologie für alle“. *Gesundhyte* 2023, pp.48-50.

TopSuite webserver: https://cpclab.uni-duesseldorf.de/topsuite/

Vasor webserver: https://cpclab.uni-duesseldorf.de/mdr3_predictor/

Projekt 2: Entwicklung eines Metaansatzes zur Protein-Protein-Kontakt-Vorhersage

Ziele des Projekts

Ziele des Projektes waren, mit TopInterface ein DNN-basiertes Programm zu entwickeln, das direkte Aminosäurekontakte in Protein-Protein-Interaktionen (PPI) vorhersagen kann. Hierzu sollten mit 2D-Convolutional Neural Networks Aminosäurekontakte aus multiplen Sequenzalignments (MSA) identifiziert und in Distanzklassen eingeteilt werden, die anschließend in einem Protein-Protein-Docking als Distanzbeschränkungen genutzt werden können, um ein 3D-Struktumodell des Protein-Protein-Komplexes zu erhalten.

Ergebnisse

Unser Ansatz belief sich darauf, analog zur Strukturvorhersage von Proteinmonomeren, koevolutionsbasierte Aminosäurekopplungen aus einem gepaarten multiplen Sequenzalignment (MSA) zu extrahieren und diese als Distanzbeschränkungen für physik-basiertes Docking zu nutzen (Abbildung 2). Um ein derartigen MSA zu konstruieren, müssen Sequenzen aus den MSA der Monomere gepaart werden. Hierfür entwickelten wir drei Verfahren unter Berücksichtigung der Distanz der Gene im Genom, der Modellierung der Phylogenie basierend auf dem bit-score und der Identifikation von Genfusion/Genfission. Verschiedene Programme zur Bestimmung koevolutionärer Kopplungen wurden auf das kombinierte MSA angewendet. Zur Repräsentation der lokalen Nachbarschaft der Aminosäuren wurden 2D-Convolutional Neural Networks genutzt, wobei die Eigenschaften in den Kanaldimensionen organisiert und die Aminosäurepaare in Distanzklassen eingeteilt wurden.

Während der Entwicklung von TopInterface gelang DeepMind mit Alphafold2 und Alphafold-Multimer ein entscheidender Sprung bei der Lösung des Proteinstrukturvorhersage-Problems. Sowohl bei Alphafold2 als auch Alphafold-Multimer handelt es um sogenannte End-to-End-Methoden, bei der die Faltung des Proteins implizit innerhalb des Programms simuliert wird, anstatt der bis dato üblichen Extraktion von Distanzbeschränkungen und einer anschließenden expliziten physik-basierten Faltungssimulation. Durch einen vorläufigen Performancevergleich wurde ersichtlich, dass TopInterface mit der implementierten Architektur nicht die Genauigkeit von AlphaFold-Multimer erreicht. In der Folge wurde die Entwicklung von TopInterface als PPI-Vorhersageprogramm abgebrochen.

Insbesondere für die Vorhersage von Komplexstrukturen bestehend aus zwei Proteinen aus derselben Spezies zeigt AlphaFold-Multimer vielversprechende Ergebnisse. Anders sieht es aus für die Vorhersage von PPI's von Proteinen, die aus verschiedenen Spezies stammen. Hierfür sind besondere Routinen erforderlich, um die Homologe in einer biologisch sinnvollen Weise zu paaren. Da diese Routinen in der aktuellen Version von AlphaFold-Multimer nicht implementiert sind, ergibt sich für TopInterface die Möglichkeit, durch Implementierung eines Verfahrens zur effizienten Paarung von Homologen für Proteine aus verschiedenen Spezies die Qualität der AlphaFold-Multimer Vorhersagen deutlich zu verbessern. Entsprechend haben wir die Zielsetzung für TopInterface angepasst.

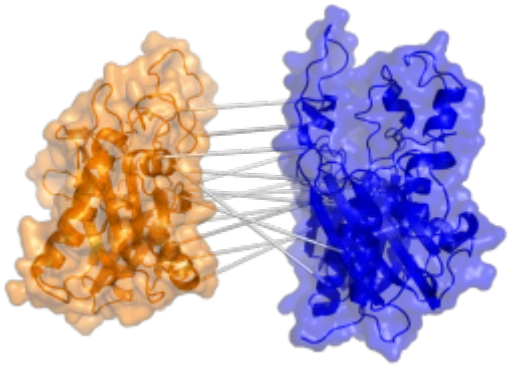

Abbildung 2:Interaktionsmodell der Proteine PqsC (orange) und PqsB (blau).Durch TopInterface vorhergesagte Kontakte zwischen den Proteinen sind durch graue Striche representiert. Die Strukturen sind dem PDB-Eintrag 5DWZ entnommen.

Relevante Publikationen

Koenig, F., Mulnaes, D., van der Weg, K., Gohlke, H. 2022. TopSuite: Collection of Deep Learning-based Metamethods to Predict and Evaluate Protein Structures and Properties. *Proceedings of the NIC Symposium 2022*, pp.47-54.

Projekt 3: Entwicklung einer Software zur automatischen Kuration von experimentell aufgeklärten Proteinstrukturen

Ziele des Projekts

Für die Simulation von Biomolekülen und die Entwicklung von struktur-basierten deep learning Programmen stellt die Protein Data Bank (PDB) die größte Quelle an hochwertigen (Protein-)Strukturdaten dar. Um diese nutzen zu können, ist in vielen Fällen eine Kuration der Einträge notwendig, da, bedingt durch die Strukturaufklärungsmethoden, häufig Bedingungen herrschen, die weit vom biologischen Kontext entfernt sind. Ziel des Projekts ist es, die Software InVivoPDB zu entwickeln, die durch automatische Kuration von experimentell aufgeklärten Proteinstrukturen den Zugang zu hochwertigen Strukturdaten ermöglicht.

Ergebnisse

InVivoPDB erlaubt es, fehlende Loop-Regionen in Kristallstrukturen zu modellieren. Es beinhaltet Programmteile zur Identifikation und Entfernung von Nicht-Protein-Strukturbestandteilen wie z.B. Kristallisationsadditiven, Antikörpern oder Aufreinigungstags, die als Teil des exprimierten Proteins ebenfalls in der Kristallstruktur vorhanden sind (Abbildung 3A). Unvollständige Moleküle in der Kristallstruktur, wie beispielsweise organische Liganden, Nuklein- oder Aminosäuren, können automatisch „repariert" und vervollständigt werden. Fehlende Proteinregionen, sogenannte Loops, werden automatisch modelliert: Loopmodelle werden mit ColabFold und ESMFold sowohl isoliert als

auch im biologischen Kontext des Gesamtproteins modelliert (Abbildung 3B). Eine Routine zur Bewertung der Loop-Qualität ermöglicht, die optimale Information für das nachfolgende Modellieren des biologischen Komplexes mit der Software MODELLER zu extrahieren. Weiterhin steht ein umfangreicher Programmteil zur Strukturoptimierung mit Hilfe von MD-Software zur Verfügung, um energetisch ungünstige Konformationen zu identifizieren und zu beseitigen.

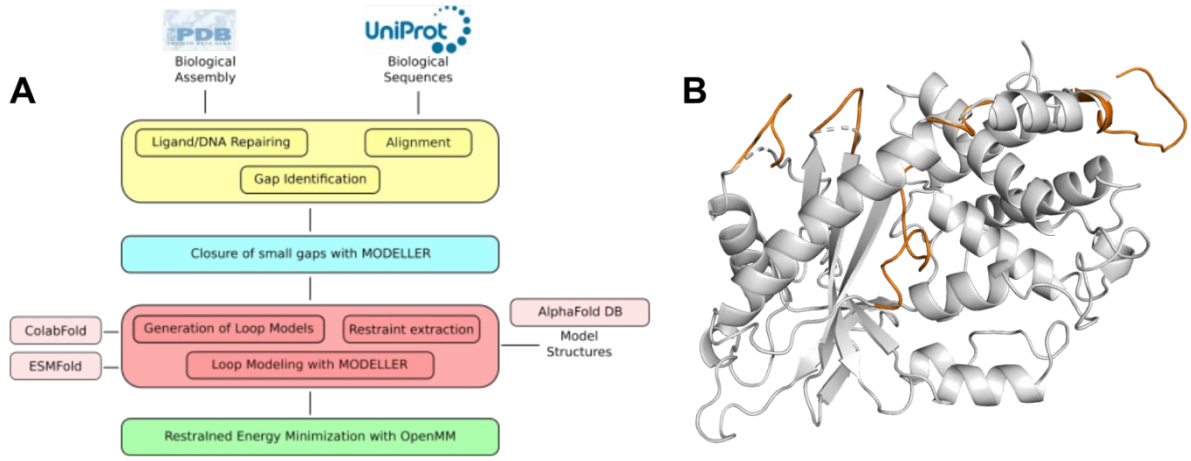

Abbildung 3: InVivoPDB Workflow zur Kuration von Kristallstrutkturen. (A) Schema von InVivoPDB. InVivoPDB ist modular aufgebaut und nutzt externe Tools wie AlphaFold, ColabFold, ESMFold und Modeller, um Strukturmodelle zu erzeugen und zu modellieren. (B) Kuration der Kristallstruktur (grau) durch Modellierung der fehlenden Regionen (orange) mit InVivoPDB. Die Struktur basiert auf dem PDB-Eintrag 6MJX.

Relevante Publikationen

InVivoPDB soll bis 01/2025 in einem peer-reviewed Journal publiziert werden. Es soll als Webserver für die wissenschaftliche Öffentlichkeit nutzbar gemacht werden.

Projekt 4: Entwicklung einer DNN-basierten Stöchiometrievorhersage von Proteinkomplexen basierend auf der Struktur

Ziele des Projekts

Um den biologischen Kontext von Proteinen akkurate zu modellieren, ist es von großer Bedeutung, die korrekte Stöchiometrie der beteiligten Proteinkomplexe zu kennen, d.h. zu wissen, wie viele Kopien eines Proteins den Komplex ausmachen. Weiterhin stellt sich die Frage, ob die Information über die Stöchiometrie eines Proteinkomplexes in der Sequenz bzw. der Struktur codiert ist. Hierfür soll mithilfe einer modernen Neural Network Architektur, wie zum Beispiel Transformern oder Graph Neural Networks, ausgetestet werden, ob eine akkurate Vorhersage basierend auf der Struktur möglich ist.

Ergebnisse

Um eine akkurate Vorhersage mit einem Neural Network zu bewerkstelligen, ist ein hochwertiger Datensatz notwendig. Der erstellte Datensatz basiert auf dem QSBIO Datensatz und Annotationen aus der UniProt Datenbank. Während sich die QSBIO Annotationen auf PDB Einträge beziehen, sind die UniProt Annotationen auch für Proteine ohne experimentell aufgeklärte Struktur verfügbar. Um ein strukturbasiertes Training zu ermöglichen, wurden mit AlphaFold-Strukturmodelle ausgewählt, die für

einen Großteil der Struktur eine hohen pLDDT-Score haben und damit eine hohe Wahrscheinlichkeit für ein akkurates Strukturmodell. Um die strukturelle Information möglichst effizient in die Vorhersage miteinzubringen, wurden „structure-aware“ Embeddings mit SaProt (Su et al., 2023) generiert. Die Struktur wird als Graph Neural Network repräsentiert, wobei jede Aminosäure durch die Koordinaten des C_α-Atoms repräsentiert wird und ein Embeddingvektor pro Aminosäure genutzt wird. Ziel ist es, die Struktur als *N*-Oligomer (*N* = 1-12) zu klassifizieren (Abbildung 4).

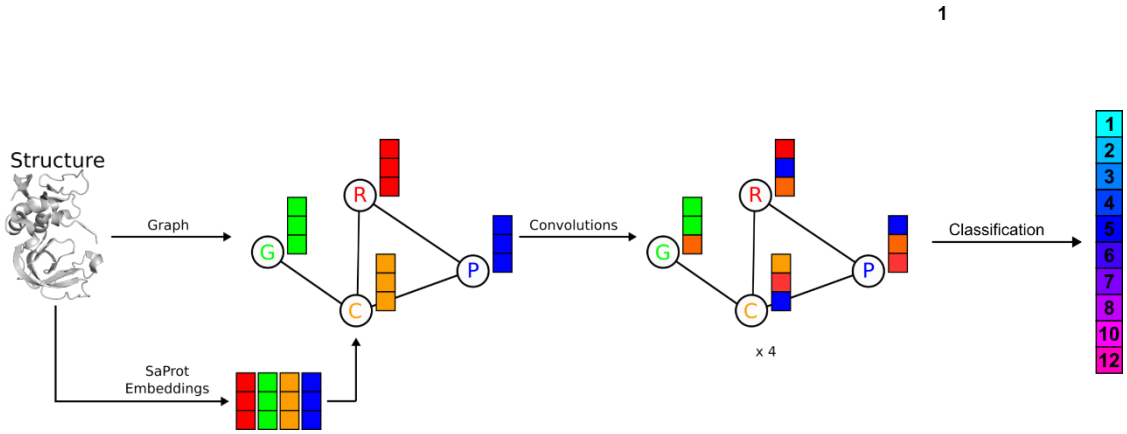

Abbildung 4: Workflow für Stöchiometrievorhersage. Die Proteinstruktur wird als Graph repräsentiert und die C_α-Atome werden mit den Embeddings annotiert. Verbundene Aminosäuren tauschen Nachrichten aus und verarbeiten so die Embeddings. Die Struktur wird im letzten Schritt in eine Oligomerklasse eingeordnet.

Relevante Publikationen

Das Tool soll bis 03/2025 in einem peer-reviewed Journal publiziert werden. Es soll als Webserver für die wissenschaftliche Öffentlichkeit nutzbar gemacht werden.

Wichtigste Positionen des zahlenmäßigen Nachweises

In den Jahren 2019, 2022 und 2023 wurden für die Modellierung von Proteinstrukturen und die Vorbereitung von Datensätzen für Lizenzen der Schrödinger Software nach Beantragung zur Mittelentsperrung insgesamt 5.946,49 € verausgabt. In den Jahren 2020 und 2023 wurden Datenspeicher zur Sicherung und Archivierung von im Projekt erzeugten Daten (Datensätze, aufbereitete Proteinstrukturen, Berechnungsergebnisse) in Höhe von insgesamt 6.559,23 € angeschafft.

Zeitplan des Vorhabens

Der Zeitplan hatte sich aufgrund der Corona-Pandemie und der damit verbundenen Einschränkungen in den Arbeitsmöglichkeiten und der Kommunikation verschoben. Durch die genehmigte kostenneutrale Verlängerung konnten die ursprünglichen Ziele, unter Berücksichtigung von unten begründeten Refokussierungen, erfolgreich erreicht werden.

Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten

Viele biologische Prozesse, vor allem in eukaryontischen Organismen, werden von Multidomänen-Proteinen gesteuert, die auf komplexe Weise miteinander interagieren. Die Kenntnis dieser Strukturen ist der Schlüssel zum Verständnis und zur Modulation der Funktionen und um Funktionen und Wege zu verstehen und zu modulieren, die von solchen Proteinen gesteuert werden, was von hoher Relevanz für die medizinische und biotechnologische Forschung ist. Zu Beginn des Projektes war die DNN-basierte Proteinstruktur- und eigenschaftsvorhersage nur mäßig entwickelt. Unsere Metaansätze TopScore und der TopModel-Prototyp gehörten zu den führenden Methoden für Proteinstrukturvorhersage.
Konsequentes Ziel des Projekts war die Entwicklung eines vollautomatischen computergestützten Arbeitsablaufs zur Vorhersage qualitativ hochwertiger Modelle der Struktur von Proteinen und Proteinkomplexen, wobei Meta-Ansätze zum Einsatz kommen, die mehrere Methoden der rechnerischen Strukturvorhersage und Quellen verfügbarer biologischer Informationen kombinieren.
Unseres Wissens nach existierte kein derart veranlagtes Projekt in Deutschland zu dieser Zeit. Die Projektarbeit war zur Bearbeitung dieser Fragestellung somit notwendig und angemessen.

Voraussichtlicher Nutzen und Verwertbarkeit sowie kurzfristige Planung

Die im Rahmen des Projektes bereits entwickelten Programme und erzeugten Datensätze wurden gemäß guter wissenschaftlicher Praxis durch Publikationen und die Möglichkeit des Downloads für die wissenschaftliche Öffentlichkeit zugänglich gemacht. Sie können als Stand-Alone-Programmversionen zur lokalen Berechnung von Proteinstruktur- und eigenschaftsvorhersagen bzw. zum Trainieren eigener Modelle verwendet werden.
Wir erwarten, dass das Projekt 2 (TopInterface interlog alignment generation) bis 11/2024 abgeschlossen und veröffentlicht sein wird. Für InVivoPDB (Projekt 3) erwarten wir den Abschluss des Projekts bis 01/2025. Wir planen, den Oligomeric State Predictor (Projekt 4) bis 03/2025 fertig zu stellen. Wir bauen die TopSuite-Webserver kontinuierlich aus, um der wissenschaftlichen Öffentlichkeit einen einfachen Zugang zu unseren Methoden zu ermöglichen.

Während der Durchführung des Vorhabens dem Zuwendungsempfänger bekannt gewordener Fortschritt auf dem Gebiet des Vorhabens bei anderen Stellen

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 von Meta, das eine ähnliche Grundstruktur wie AlphaFold2 nutzt, jedoch ein sogenanntes protein language model anstelle von multiplen Sequenzalignments und damit Strukturvorhersagen in kürzester Laufzeit ermöglicht.
Seit AlphaFold2 haben sich zudem mehrere weitere wichtige Entwicklungen ergeben: AlphaFold-Multimer, eine Erweiterung von AlphaFold2, dient der Vorhersage von Proteinkomplexen, insbesondere von multimeren Proteinen, die durch Wechselwirkung miteinander funktionelle Einheiten bilden. Sie befasst sich mit der Herausforderung, die Strukturen von Protein-Protein-Interaktionen genau vorherzusagen, was für das Verständnis zellulärer Prozesse und die Entwicklung von Therapeutika, die auf Proteinkomplexe abzielen, entscheidend ist. AlphaFold3 liefert weitere Verbesserungen in Bezug auf Genauigkeit, Geschwindigkeit und die Fähigkeit, dynamische Aspekte von Proteinen vorherzusagen. Es ist zu erwarten, dass AlphaFold3 noch komplexere Proteindynamik, Membranproteine und eine verbesserte Vorhersage von Proteininteraktionen in biologischen Kontexten behandeln wird. AlphaFlow konzentriert sich auf die Integration von Proteinmodellierung mit Arbeitsabläufen und Pipelines in der Bioinformatik. Diese Entwicklung zielt darauf ab, AlphaFold-Vorhersagen für groß angelegte Anwendungen zugänglicher und skalierbarer zu machen, um eine systematischere Integration von Strukturvorhersagen in funktionelle Studien, Medikamentendesign und personalisierte Medizin zu ermöglichen.
Diese Fortschritte markieren insgesamt eine Verlagerung hin zu einer umfassenderen Proteinmodellierung, die über statische Strukturvorhersagen hinausgeht und dynamische Wechselwirkungen, Proteinkomplexe und funktionelle Erkenntnisse einschließt. Die Ressourcen, die z.B. von DeepMind und Meta in die Entwicklung dieser Ansätze fließen, sind umfassend. Die Fortschritte auf diesem Gebiet haben z.T. eine Refokussierung unseres Projektes InCelluloProtStruct notwendig gemacht.

Erfolgte oder geplante Veröffentlichungen des Ergebnisses nach Nr. 5 der NABF

Siehe oben zu den einzelnen Projekten zugeordnet.