CASP und CAMEO

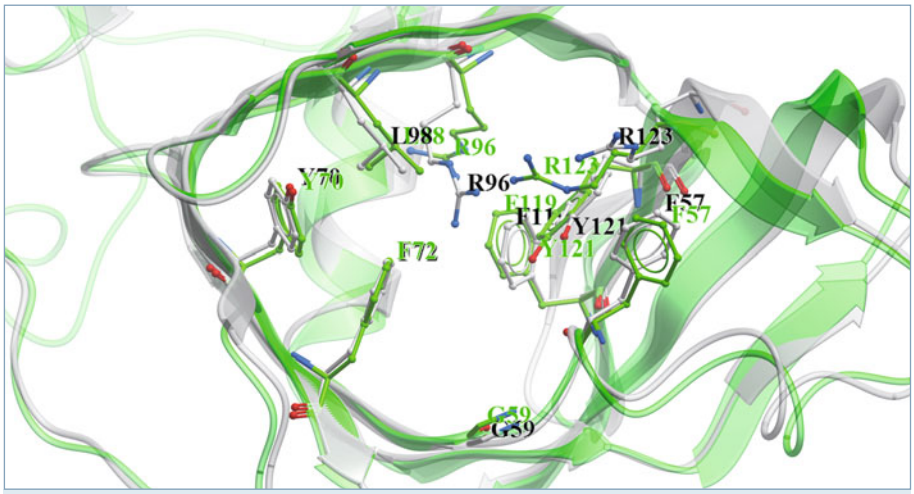
Unabhängige Validierung von Methoden zur Proteinstrukturvorhersage

JÜRGEN HAAS, TORSTEN SCHWEDE
SIB – SWISS INSTITUTE OF BIOINFORMATICS, BIOZENTRUM UNIVERSITÄT BASEL

10.1007/s12268-014-0432-3

Computergestützte Vorhersagen sind ein Teil unseres täglichen Lebens geworden – im Alltag wie in der Wissenschaft. Computersimulationen machen Vorhersagen über das Wetter und das Weltklima, Börsenkurse und die volkswirtschaftliche Entwicklung, die Crash-Stabilität einer neuen Autokarosserie oder die Bindung eines Wirkstoffs an einen Rezeptor bei der Entwicklung neuer Medikamente. Eine Frage bleibt dabei jedoch häufig offen: Wie verlässlich sind diese Vorhersagen? Aus unserer Alltagserfahrung wissen wir, dass die Wettervorhersage für morgen recht zuverlässig ist, während die Sechs-Wochen-Wettervorhersage ungefähr so nutzlos ist wie die Prognose unseres Bankers für Börsenkurse. Aber wie steht es mit der Genauigkeit von wissenschaftlichen Vorhersagen? Wenn wir der wissenschaftlichen Literatur Glauben schenken, dann sind die meisten veröffentlichten Methoden besser als alle anderen – was bereits aus statischen Gründen eher unwahrscheinlich erscheint und einen deutlichen Hinweis auf einen *reporting bias* liefert. Wie lässt sich also die Vorhersagegenauigkeit einer Methode unabhängig und objektiv evaluieren?

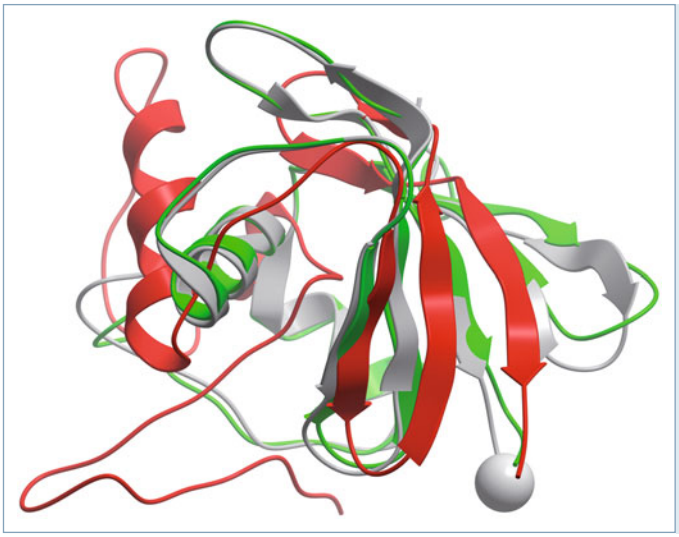
Die Vorhersage der dreidimensionalen (3D-) Struktur eines Proteins aus seiner Aminosäuresequenz wurde häufig als der „heilige Gral" der Strukturbioinformatik bezeichnet, und entsprechend viele unterschiedliche Ansätze wurden dazu entwickelt und publiziert. Im Rahmen von CASP (*Critical Assessment of Techniques for Protein Structure Prediction*, http://predictioncenter.org) wurden im Verlauf der letzten 20 Jahre Mechanismen entwickelt, um diese Methoden zur Proteinstrukturvorhersage objektiv und unabhängig zu evaluieren [1]. Diese Konzepte waren wegweisend für eine Reihe verwandter Initiativen, wie z. B. CAMEO zur vollautomatischen Evaluierung von Servern zur Strukturmodellierung (www.cameo3d.org [2]), CAPRI (Proteinkomplexe), CAFA (Proteinfunktion) oder BioCreative (Textmining). Im Folgenden wollen wir die Konzepte zur unabhängigen Validierung von Methoden zur Proteinstrukturvorhersage durch CASP und CAMEO vorstellen.



▲ **Abb. 1:** Analyse der Seitenkettengenauigkeit eines Modells von SWISS-MODEL (grün, GDT = 95) für das CAMEO-Target 4NW4_A (grau). Zur Darstellung wurde das Modell mit der Referenzstruktur überlagert. Sauerstoffatome sind rot, Stickstoffatome blau, Kohlenstoffatome in der experimentellen Struktur grau und im Modell grün dargestellt. In diesem Beispiel repräsentiert das Modell die Koordinaten des Targets sehr gut.

Proteine sind für eine Vielzahl an molekularen Aufgaben in der lebenden Zelle verantwortlich, und jedes Protein besitzt seiner Funktion entsprechende spezifische strukturelle Eigenschaften. Mithilfe von experimentellen Methoden lassen sich die 3D-Strukturen von Proteinen aufklären. Im weltweiten Archiv für Proteinstrukturen, der *Protein Data Bank* (PDB; www.pdb.org), sind gegenwärtig etwa 100.000 Strukturen abgelegt. Allerdings sind wir heute weit davon entfernt, für alle Proteine eine experimentelle Struktur zu kennen: Die Anzahl der bekannten Proteinsequenzen ist mit etwa 53 Millionen Sequenzen (www.uniprot.org) um zwei Größenordnungen größer als die Zahl der bekannten Proteinstrukturen. Aus diesem Grund sind Methoden zur Vorhersage der Strukturen mithilfe von Computerprogrammen von großem Interesse. Dabei lassen sich grob zwei unterschiedliche Ansätze unterscheiden: *De novo*-Vorhersagemethoden versuchen die dreidimensionale Struktur eines Proteins aufgrund seiner Aminosäuresequenz, Vorhersagen der lokalen Strukturen und allgemeinem Wissen über Strukturprinzipien vorherzusagen. Bisher ist die Anwendbarkeit von *de novo*-Methoden jedoch auf kleinere Proteindomänen beschränkt und die Genauigkeit für praktische Anwendungen häufig nicht ausreichend. Im Gegensatz dazu versucht man bei Methoden zur vergleichenden Modellierung eine Homologiebeziehung zwischen dem Zielprotein (Target) und anderen Proteinen mit bekannten Strukturen (Templates) herzustellen und diese Information zum Erstellen eines 3D-Modells zu verwenden. Homologiemodellierung wird heute in der Forschung

als Routinemethode verwendet, um Strukturmodelle zu generieren [3]. Meist kommen dabei vollautomatische Server zum Einsatz, die es auch Nicht-Bioinformatikern erlauben, verlässliche Modelle zu erhalten, wie z. B. der SWISS-MODEL server (http://swissmodel.expasy.org), HHpred (http://toolkit.lmb.uni-muenchen.de/hhpred) oder Phyre (www.sbg.bio.ic.ac.uk/phyre2). Dabei spielt die Genauigkeit von Modellen eine große Rolle und wird häufig numerisch als Bruchteil von richtig vorhergesagten Aminosäuren gemessen, z. B. durch den GDT (*global distance test*) oder lDDT (*local distance difference test*). Modelle mit GDT-Werten nahe bei 100 sind sehr gut (**Abb. 1**), solche unter 50 haben nur begrenzt Ähnlichkeit mit der tatsächlichen Struktur des Zielproteins (**Abb. 2**, rotes Modell).



◀ **Abb. 2:** Bänderdarstellung zweier Modelle für ein CASP-Target, die CorC/HlyC-Transporter-assoziierte Domäne (grau, 4GH0, Kette A). Strukturell überlagert sind ein akkurates Modell der Methode HHPred-thread (grün, GDT = 92) und ein weitgehend falsch gefaltetes Modell (rot, GDT = 37).

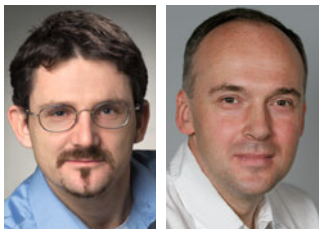
Um die Vorhersagegenauigkeit verschiedenster Techniken vergleichen zu können, wird seit 20 Jahren alle zwei Jahre ein CASP-Wettbewerb organisiert [1]. Die Organisatoren stellen dabei sicher, dass folgende Prinzipien zur Anwendung kommen: (1) Die Analyse basiert auf Vorhersagen (nicht *postdictions*), das heißt zum Zeitpunkt der Vorhersage ist die experimentelle Struktur des Proteins noch nicht bekannt. (2) Um die Vergleichbarkeit der Resultate sicherzustellen, werden die Targets allen beteiligten Gruppen zur selben Zeit zur Verfügung gestellt, und die Resultate müssen zur selben Deadline abgeliefert werden. Somit steht allen Gruppen auch dieselbe Hintergrundinformation zur Verfügung. (3) Die Modelle werden durch die Organisatoren vor der Auswertung vollständig anonymisiert, das heißt die Gutachter kennen die Identität der Methoden nicht und beurteilen die Qualität der Endergebnisse vorurteilsfrei. (4) Die Evaluierung der Modelle erfolgt durch einen unabhängigen Experten, der objektive Kriterien für die Auswertung festlegt und selbst nicht am CASP teilnehmen darf. Erst nach Abschluss der Analyse geben die Organisatoren die Identität der Gruppen bekannt. Während einer CASP-Runde werden Modelle für ca. 100 Proteine ausgewertet, und die besten Methoden an einem Meeting präsentiert.

Alle erfolgreichen Ansätze zur Strukturvorhersage stützen sich heute auf vollautomatische Computerprogramme. Um der stürmischen Entwicklung dieser Methoden gerecht zu werden, haben wir vor etwa zwei Jahren das CAMEO-Projekt (*Continuous Automated Model EvaluatiOn*, www.cameo3d.org) etabliert, um eine kontinuierliche differenzierte Evaluierung auf einer breiten Datenbasis zu ermöglichen. CAMEO basiert auf vorab veröffentlichten Aminosäuresequenzen von Strukturen, die in der folgenden Woche von der PDB publiziert werden. Diese Targetsequenzen werden automatisch an die teilnehmenden Server geschickt, die dann bis zum folgenden Mittwoch Zeit haben, um ihre Vorhersagen an CAMEO zur Evaluierung zu schicken. Im Verlauf von 112 Wochen wurden auf CAMEO 47.575 Vorhersagen für 2.006 Targets ausgewertet. Neben Methoden zur Vorhersage von 3D-Strukturen evaluiert CAMEO auch Techniken zur Vorhersage von Aminosäuren, die an der Bindung von Liganden beteiligt sind (*ligand binding site residue prediction*), und Programme zur Abschätzung der Modellgenauigkeit (*model quality estimation tools*).

Auf der Internetseite von CAMEO (www.cameo3d.org) lassen sich die evaluierten Methoden, basierend auf verschiedenen numerischen Kriterien, über verschiedene Zeiträume vergleichen. Mit CAMEO steht der *computational structural biology community* eine Plattform zur Verfügung, die unabhängige zeitnahe Benchmarks auf großen Datensätzen ermöglicht. CAMEO unterstützt die Entwickler von Vorhersagemethoden und fordert sie zugleich, da neue Methoden im direkten Vergleich mit konkurrierenden Ansätzen auf verschiedensten Aspekten der Strukturvorhersage gemessen werden. Dies kommt direkt den Nutzern von Strukturvorhersagemethoden zugute, die so die für ihre Anwendung am besten geeigneten Methoden aussuchen können. Als Community-Projekt ist CAMEO offen für Anregungen und Vorschläge, um die Entwicklung der nächsten Generation von Strukturvorhersagemethoden optimal zu fördern. ■

Literatur

[1] Moult J, Fidelis K, Kryshtafovych A et al. (2014) Critical assessment of methods of protein structure prediction (CASP) – round X. Proteins 82:1–6
[2] Haas J, Roth S, Arnold K et al. (2013) The Protein Model Portal: a comprehensive resource for protein structure and model information. Database 2013:bat031, doi: 10.1093/database/bat031
[3] Schwede T (2013) Protein modeling: what happened to the „protein structure gap“? Structure 21:1531–1540



Jürgen Haas (links) und Torsten Schwede

Korrespondenzadresse:
Prof. Dr. Torsten Schwede
Swiss Institute of Bioinformatics – SIB
Biozentrum Universität Basel
Klingelbergstraße 50–70
CH-4056 Basel
Tel.: +41-(0)61-267-1586
Fax: +41-(0)61-267-1584
Torsten.Schwede@unibas.ch
www.biozentrum.unibas.ch/schwede