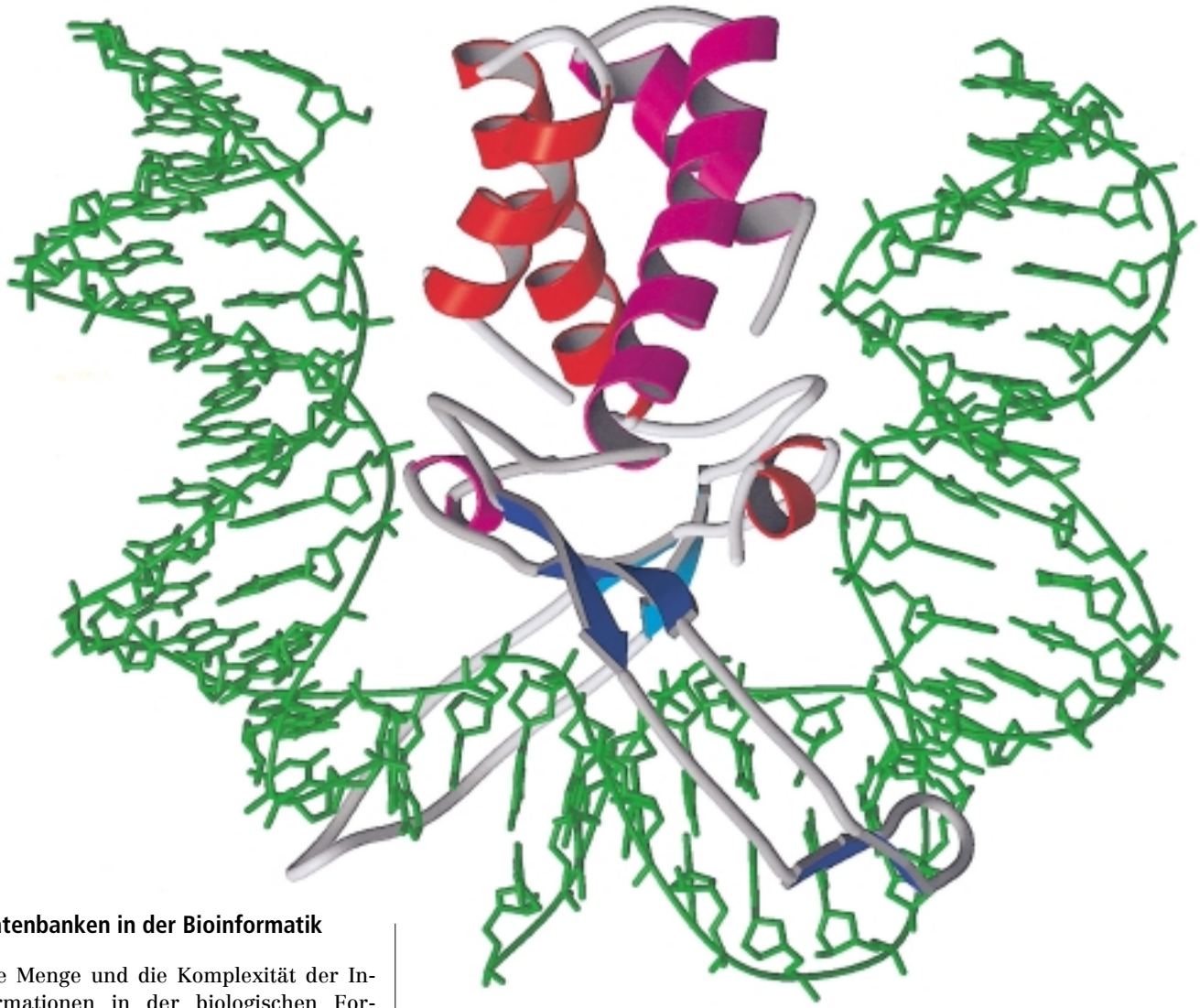
Die Strukturprognose von Proteinen

Herausforderung für die Molekulare Bioinformatik



Datenbanken in der Bioinformatik

Die Menge und die Komplexität der Informationen in der biologischen Forschung nehmen derzeit rasant zu. Viele der Informationen sind in traditionell strukturierten Datenbanken organisiert und stehen der Forschung über das Internet zur Verfügung. Zu den bekanntesten zählen Entrez/Genbank (http://www.ncbi.nlm.nih.gov), EMBL (http://www.ebi.ac.uk), und SwissProt (http://www.expasy.ch), in denen DNA- und Proteinsequenzen sowie Tertiärstrukturen allgemein verfügbar sind. Unter den neu gewonnenen Daten haben die Genomsequenzen zweifellos den bedeutendsten Anteil.

Eine zentrale Aufgabe der Bioinformatik besteht in der Organisation dieser komplexen und großen Datenmengen, aber auch in der Aufdeckung neuartiger Informationszusammenhänge im Sinne eines *data mining*. Die durch Gensequenzierung erhaltenen Informationen sind in vielen Fällen nur dann praktisch verwertbar, wenn die funktionelle Bedeutung einer bestimmten (Gen-)Sequenz aufgeklärt wird. Die exprimierten Proteine des zellulären Proteoms bilden den zentralen Mittelpunkt biologischer Prozesse.

Tertiärstruktur von Proteinen

Der Schlüssel zum Verständnis der biologischen und funktionellen Eigenschaften von Proteinen liegt in ihrer Struktur begründet. *Life Sciences*-Unternehmen benötigen diese biologischen Eigenschaften beispielsweise zur Einschätzung von eigenen experimentellen Arbeiten und für Erfindungen und Patente. Die Struktur eines gegebenen Proteins wird meist experimentell bestimmt. Dies erfordert jedoch einen hohen Zeit- und Kostenaufwand,

Abb. 1: Modellstruktur des DNA-bindenden Proteins HU aus Thermotoga maritima, einem hyperthermophilen Eubakterium, mit daran gebundener doppelsträngiger DNA. Das Modell unterscheidet sich von der – später – experimentell bestimmten Kristallstruktur nur in der DNA-Bindungsstelle; die Kristallstruktur wurde ohne Zusatz von DNA bestimmt und gibt daher die für funktionelle Studien wichtigen DNA-Bindungseigenschaften nicht wieder.

Keywords

Strukturrechnung, Tertiärstrukturprognose, Genomanalyse, CASP-Wettbewerb

Gerald Böhm

und ein (schneller) Erfolg ist nicht garantiert. Ein Computerverfahren zur Modellierung der Struktur kann dagegen rasch und kostengünstig erfolgen und bereits vor der aufwändigen experimentellen Strukturaufklärung wesentliche Eigenschaften des untersuchten Zielproteins korrekt darstellen (Abb. 1). Die Erstellung solcher Strukturmodelle ist Teil der modernen molekularen Bioinformatik; das Faltungsproblem, also die Vorhersage der Tertiärstruktur von Proteinen aufgrund von Sequenzinformationen, gilt heute als die *Königsdisziplin* der Bioinformatik. Bis heute ist noch nicht verstanden, nach welchem Mechanismus sich eine gegebene Aminosäuresequenz zu einer nativen und funktionellen Protein-Tertiärstruktur faltet, somit existiert auch kein eindeutiger mathematischer Algorithmus zur Ableitung der Tertiärstruktur anhand von Sequenzinformationen.

Wissensbasierte Modellierung

Aus diesen Gründen werden bei Vorhersagen der Tertiärstruktur von Proteinen wissensbasierte Ansätze zugrunde gelegt, die derzeit als die zuverlässigsten Verfahren zur Strukturprognose angesehen werden [1]. Hierbei wird versucht, bei Kenntnis der Sequenz eines unbekannten Proteins und einer dazu „verwandten" Templatstruktur durch Vergleichende Modellierung *(homology modeling)* auf ein Tertiärstrukturmodell zu schließen. Eine bislang unbekannte Faltungstopologie kann daher nicht vorhergesagt werden. Es wird jedoch erwartet, dass binnen der nächsten fünf bis acht Jahre im Rahmen der Initiative *„structural genomics"* alle relevanten natürlichen Topologien bekannt sein werden; völlig neuartige Proteintopologien sind danach kaum mehr zu erwarten [2].

Die üblicherweise eingesetzten Verfahren der Vergleichenden Modellierung sind ab einem bestimmten Verwandtschaftsgrad (etwa 50 % Sequenzidentität von unbekanntem Protein und Templat [3]) relativ robust und zuverlässig, können aber auch dann Details wie beispielsweise Unterschiede in der Elektrodynamik im aktiven Zentrum eines Proteins nur mit begrenzter Auflösung darstellen. Es ist daher sehr wichtig, zu jedem Tertiärstrukturmodell auch dessen Zuverlässigkeit zu bestimmen, damit eine Überinterpretation der Modelle ausgeschlossen ist. Für Vergleichende Modellierung stehen heute verschiedene kommerzielle und nichtkommerzielle Verfahren und Algorithmen zur Verfügung [4].

Die Modellierung lässt sich grundsätzlich nach folgenden Arbeitsschritten durchführen:

1. Identifizierung verwandter Proteine durch Vergleich auf Sequenzbasis (Sequenzhomologien) oder mit anderen Verfahren (z. B. *threading*)
2. Alignment der Sequenzen von unbekanntem Protein und Elterstrukturen; es sollten möglichst viele (verschiedene) Elterstrukturen einer gemeinsamen Faltungstopologie mit eingebunden werden
3. Identifizierung strukturell konservierter und variabler Regionen (Proteinkern und Loops)
4. Ableitung der Koordinaten des Proteinkerns (strukturell konservierte Bereiche, insbesondere in den Regionen periodischer Sekundärstruktur)
5. Vorhersage der Konformation der Loops (strukturell variable Bereiche) einschließlich der Modellierung von Insertionen und Deletionen in diesen Segmenten
6. Validierung der Modellstruktur und Qualitätsanalyse, ggf. noch geometrische Verfeinerung der Modellstruktur.

Der Wettbewerb CASP

Für jeden der erläuterten Schritte existieren eine Reihe von kommerziell oder frei verfügbaren Werkzeugen, jedoch erfordern die einzelnen Prozesse auch viel Erfahrung und handwerkliches Geschick. Die Entwicklung neuer, zuverlässiger und automatisierbarer Algorithmen zur Tertiärstrukturvorhersage ist daher eine vordringliche Herausforderung. Der Erfolg solcher neuer Verfahren wird in einem internationalen und öffentlichen Wettbewerb im Abstand von jeweils zwei Jahren bewertet [5]. Bei diesem CASP-Wettbewerb (Critical Assessment of Techniques for Protein Structure Prediction, vgl. http://predictioncenter. llnl.gov/) können Forschergruppen ihre Vorschläge für bislang unbekannte Proteinstrukturen einreichen, deren experimentelle Darstellung jeweils kurz vor der Aufklärung steht. Nach erfolgreicher experimenteller Strukturaufklärung werden die bis dato eingereichten Modelle mit der realen Struktur verglichen; dadurch werden erfolgreiche Verfahren objektiv bewertet. Der CASP-Wettbewerb ist heute anerkannter Standard bei der Beurteilung neuer Modellierungsverfahren.

Ausblick

Nach der Bereitstellung einer qualitativ hochwertigen Datenbank der humanen Genomsequenz und anderer wichtiger Genome wird der nächste große Schritt der Forschung in der Bestimmung wichtiger biologischer Funktionen der Bestandteile des zellulären Proteoms liegen. Hierzu kann die molekulare Bioinformatik mit ihrer anspruchsvollsten Disziplin, der Tertiärstrukturvorhersage von Proteinen anhand der Sequenz, wertvolle Hilfestellungen leisten. Die Entwicklung neuartiger und automatisierbarer Verfahren zur Tertiärstrukturprognose im Rahmen der Vergleichenden Modellierung wird daher in Zukunft enorme Bedeutung gewinnen.

Literatur

[1] Böhm, G.: Biophys. Chem. 59, 1-32 (1996)

[2] Berman, H.M.; Bhat, T.N.; Bourne, P.E.; Feng, Z.; Gilliland, G.; Weissig, H.; Westbrook, J.: Nature Struct. Biol. 7, 957-959 (2000)

[3] Hilbert, M.; Böhm, G.; Jaenicke, R.: Proteins: Struct. Funct. Genet. 7, 138-151 (1993)

[4] Eisenhaber, F.; Persson, B.; Argos, P.: Crit. Rev. Biochem. Mol. Biol. 30, 1-94 (1995)

[5] Moult, J.; Hubbard, T.; Fidelis, K.: Pedersen, J.T.: Proteins: Struct. Funct. Genet. Suppl 3, 2-6 (1999)

Dr. Gerald Böhm

Studium der Biologie und Physik an der Universität Regensburg und Promotion dort bei Prof. Dr. Rainer Jaenicke über die Strukturvorhersage von extremophilen Proteinen. Nach Anstellungen am Universitätsklinikum Regensburg (Prof. Dr. Hans Wolf) und am Institut für Biotechnologie der Martin-Luther-Universität Halle-Wittenberg (Prof. Dr. Rainer Rudolph) gründete er im Jahr 2000 das Unternehmen ACGT ProGenomics AG und ist dort als Vorstand tätig.

ACGT ProGenomics AG
Biozentrum Halle
Weinbergweg 22
06120 Halle (Saale)
www.acgt-progenomics.de

Easy Info 000