Diss. ETH Nr. 11995

Self-correcting distance geometry for the automatic assignment of NMR NOESY spectra and the prediction of protein tertiary structures

A dissertation submitted to the
Swiss Federal Institute of Technology Zürich
for the degree of
Doctor of Natural Sciences

presented by
Christian Mumenthaler
Dipl. Phys. ETH

born on June 21, 1969
citizen of Murgenthal (Aargau)

accepted on the recommendation of
Prof. Dr. Kurt Wüthrich, Referent
Prof. Dr. Werner Braun, Korreferent
Dr. Peter Güntert, Korreferent

1996

Summary

Distance geometry using the variable target function method was introduced a decade ago to calculate three-dimensional protein structures from data collected by NMR experiments (Braun & Gō, 1985). The method was further improved a few years later to account for the increasing size and complexity of the protein structures analyzed by NMR spectroscopy (Güntert et al., 1991). A prerequisite for these studies was that the experimental input data set was correct and self-consistent. In contrast to this, "self-correcting" distance geometry presented in this thesis is specialized to inconsistent input data sets which are iteratively improved by structure calculations combined with the filtering of violated constraints. This considerably broadens the scope of distance geometry, and makes it suitable for tasks that are as different as the protein tertiary structure prediction and the automated assignment of NOESY spectra.

In the first section, self-correcting distance geometry is introduced and employed for the automated assignment of two-dimensional NOESY spectra. The method assigns the NOESY peak list and calculates three-dimensional protein structures simultaneously, using a list of NOESY peak positions, a list of proton chemical shifts, and $^3J_{NH\alpha}$ coupling constants. Tests were performed with simulated peak lists of the proteins dendrotoxin K, α-amylase inhibitor tendamistat and the DNA-binding domain of the 434 repressor protein, and with an experimental NOESY peak list of dendrotoxin K. For an assumed tolerance of ±0.01 ppm in the chemical shifts of the peak positions, only about 10% of the NOESY cross peaks can be unambiguously assigned based on their chemical shifts alone. The automated method presented here assigned about 80% of all cross peaks with this chemical shift tolerance, and 95% to 99% of the assignments were correct.

In the second section, the previously described method is brought from an experimental stage to practical application through its integration into the distance geometry package DIANA and by numerous enhancements including an automatic calibration, extension to 3D NOESY spectra, and the simultaneous handling of different peak lists. Twelve experimental peak lists from six different proteins were re-assigned from scratch with 0.6% to 2.6% differences to the manual assignment. The final structures

had target function values that are comparable to those of high-quality NMR structures, and the average backbone RMSD to the mean structure ranged from 0.7 Å to 2.2 Å.

In the third section, a variation of the self-correcting distance geometry algorithm is applied to the prediction of protein tertiary structures, using a test set of eight α-helical proteins. With the knowledge of the amino acid sequence and the helical segments alone, the completely automated method calculated the correct backbone topology of six proteins. The accuracy of the predicted structures ranged from 2.3 Å to 3.1 Å for the helical segments compared to the experimentally determined structures. For two proteins, the predicted constraints were not restrictive enough to yield a conclusive prediction. The method can be applied to all small globular proteins, provided the secondary structure is known or can be predicted with high reliability.

Finally, distance geometry with an error-tolerant target function is used for the modeling of the two N-terminal domains of the cell receptor protein CD46. This glycoprotein located on the surface of human cells is exploited by the cell entry mechanism of the measles virus. 3D models of the individual domains have been built on the basis of known structures of a related protein. Possible relative orientations of the two domains were sampled by a Monte Carlo simulation, leading to four possible models. These models were tested by a mutagenesis study in a cooperation with the group of PD Dr. R. Cattaneo (University of Zürich). The results are consistent with one of the models and suggest that the viral attachment protein hemagglutinin does not bind at the membrane-distal tip of CD46, but near the concave interface region of both domains.

Zusammenfassung

Distanzgeometrie im Torsionswinkelraum wurde vor etwas mehr als zehn Jahren eingeführt, um aufgrund der aus NMR Experimenten gewonnenen Information in Form von Distanzeinschränkungen 3D Proteinstrukturen zu berechnen (Braun & Gō, 1985). Die zunehmende Grösse der mit NMR untersuchten Proteine führte im Laufe der Jahre zu wesentlichen Verbesserungen der Methode (Güntert et al., 1991). Die Hauptvoraussetzung der Methode war die Richtigkeit und Selbstkonsistenz des Datensatzes. Im Gegensatz dazu ist die "selbst-korrigierende" Distanzgeometrie, welche in dieser Dissertation behandelt wird, auf inkonsistente Datensätze spezialisiert, welche iterativ mit Strukturrechnungen und dem Filtrieren von verletzten Distanzeinschränkungen verbessert werden. Dank dieser neuen Algorithmen kommt die Distanzgeometrie in völlig neuen Bereichen wie der Voraussage von Proteinstrukturen sowie der Zuordnung von NOESY Spektren zum Einsatz.

Im ersten Kaptitel wird die selbst-korrigierende Distanzgeomtrie eingeführt und für die automatische Zuordnung von zweidimensionalen NOESY Spektren verwendet. Die Zuordnung wird dabei in einem iterativen Verfahren mit der Bestimmung der 3D Struktur kombiniert. Die Eingabedaten bestehen aus drei Listen, welche Kreuzsignal-Positionen, chemische Verschiebungen der einzelnen Protonen und $^3J_{HN\alpha}$ Kopplungskonstanten enthalten. Getestet wurde die Methode mit drei simulierten und einer experimentellen NOESY Kreuzsignal-Liste. Bei einer Toleranz von ±0.01 ppm für die Kreuzsignal-Positionen können nur ungefähr 10% der Kreuzsignale aufgrund ihrer Lage eindeutig einem Protonenpaar zugewiesen werden. Die hier vorgestellte Methode ordnet bei derselben Toleranz über 80% der Kreuzsignale zu, wovon 95% bis 99% korrekt.

Im zweiten Kapitel wird die zuvor behandelte Methode von einem experimentellen in ein praktisches Stadium gebracht durch die Integrierung der Algorithmen in das Distanzgeometrieprogramm DIANA gekoppelt mit zahlreichen Verbesserungen, welche unter anderem eine automatische Kalibrierung der Kreuzsignal-Volumina, die Erweiterung der Methode auf 3D Spektren und die parallele Verarbeitung mehrerer Kreuzsignal-Listen beinhalten. Die automatische Zuordnung von zwölf experimentel-

len Kreuzsignal-Listen von sechs unterschiedlichen Proteinen weist bloss 0.6% bis 2.6% Unterschiede zur manuellen Zuordnung auf. Die Zielfunktions-Werte der automatisch berechneten Strukturen sind mit denjenigen von hochaufgelösten NMR Strukturen vergleichbar, und die RMSD-Werte für die Rückgratatome zur mittleren Struktur liegen zwischen 0.7 Å und 2.2 Å.

Im dritten Kapitel wird eine Variante des selbst-korrigierenden Algorithmus auf die Strukturvorhersage von acht kleinen, α-helikalen Proteinen angewandt. Bei bekannter Aminosäure-Sequenz und dem Wissen über die helikalen Segmente konnte die Methode die richtige Topologie von sechs Proteinen berechnen. Die Unterschiede der vorausgesagten Strukturen zu den experimentell bestimmten Strukturen liegen im Bereich von 2.3 Å bis 3.1 Å für die Rückgratatome der helikalen Segmente. Für zwei Proteine waren die vorausgesagten Distanzeinschränkungen nicht restriktiv genug um eine Strukturvoraussage machen zu können. Die Methode kann auf alle kleinen α-helikalen Proteine angewandt werden falls die Sekundärstruktur bekannt ist oder mit hoher Sicherheit vorausgesagt werden kann.

Im letzten Kapitel wird Distanzgeometrie mit einer Fehler-toleranten Zielfunktion für die Modellierung der zwei N-terminalen Domänen des Zellrezeptor-Proteins CD46 eingesetzt. Dieses Glycoprotein ist in der Zellmembran verankert und wird vom Masern-Virus im ersten Schritt der Zellinfektion benützt. 3D Modelle wurden aufgrund einer bekannten Struktur eines homologen Proteins konstruiert. Vier mögliche relative Orientierungen der beiden Domänen wurden mit einer Monte-Carlo-Simulation identifiziert und im Rahmen einer Zusammenarbeit mit der Gruppe von PD Dr. R. Cattaneo (Universität Zürich) durch Mutationsstudien getestet. Die Resultate sind mit einem der Modelle konsistent und deuten darauf hin, dass das virale Bindungsprotein Hämagglutinin in der konkaven Verbindungsstelle zwischen der beiden Domänen bindet.