

Diss. ETH Nr. 10135

**Neue Rechenverfahren für die
Proteinstrukturbestimmung mit Hilfe der
magnetischen Kernspinresonanz**

**Abhandlung
zur Erlangung des Titels
Doktor der Naturwissenschaften
der
Eidgenössischen Technischen Hochschule Zürich**

**vorgelegt von
Peter Güntert
Dipl. Phys. ETH**

**geboren am 29. März 1964
von Schleithem (Schaffhausen)**

**Angenommen auf Antrag von
Prof. Dr. Kurt Wüthrich, Referent
Prof. Dr. Wilfred van Gunsteren, Korreferent
1993**

Kurzfassung

Diese Abhandlung führt neue Rechenmethoden ein zur Ermittlung der dreidimensionalen Struktur von Proteinen in Lösung aufgrund von Daten, wie sie die magnetische Kernspinresonanzspektroskopie (Ernst *et al.*, 1987) liefert. Um das vor einigen Jahren dafür eingeführte Verfahren (Wüthrich, 1986) auf kleine Proteine effizient oder auf größere Proteine überhaupt anwenden zu können, waren und sind auf verschiedenen Stufen — bei den Kernresonanzexperimenten, der Herstellung und Auswertung der Spektren und den Strukturrechnungen — Neuerungen und Verbesserungen nötig. Die im Rahmen dieser Dissertation entwickelten Verfahren werden hier in der Reihenfolge, in der sie während einer Proteinstrukturbestimmung zum Einsatz kommen, vorgestellt und an Hand von praktischen Anwendungen erläutert. Der Schwerpunkt liegt bei der Berechnung dreidimensionaler Proteinstrukturen mittels Distanzgeometrierechnungen.

Dem Entwurf neuer Experimente in der hochauflösenden Kernspinresonanzspektroskopie dient der Produktoperatorformalismus, dessen erstmalige vollständige Implementation in *Mathematica*, einer Programmiersprache für symbolisches Rechnen, die Simulation von Kernspinresonanzexperimenten mit Spin-1/2-Systemen wesentlich erleichtert (Güntert *et al.*, 1993).

Das neue Programm PROSA zur Verarbeitung mehrdimensionaler Kernresonanzdaten wird vorgestellt, in dem die gebräuchlichen Verarbeitungsschritte wie lineare Vorhersage, digitale Filterung, Fourier-Transformation und automatische Phasen- und Basislinienkorrektur effizient und flexibel implementiert sind (Güntert *et al.*, 1992). Dank einer zuverlässigen Routine zur automatischen Phasenkorrektur wird keine Grafik benötigt, so daß das Programm ohne weiteres auch auf Hochleistungsrechnern, für die es besonders optimiert ist, laufen kann. Ebenfalls neu ist das FLATT Verfahren zur Basislinienkorrektur (Güntert & Wüthrich, 1992).

Zur Intensitätsbestimmung von Kern-Overhausereffekten wird eine Methode vorgeschlagen, die gut aufgelösten Signalen Referenzlinienformen entnimmt, um mit ihrer Hilfe die Intensitäten überlappender Signale zu finden (Eccles *et al.*, 1991). Die Messung skalarer Spin-Spin-Kopplungskonstanten durch inverse Fourier-Transformation von In-Phase-Multipletts wird einge-

führt (Szyperski *et al.*, 1992a). Überlappende Resonanzfrequenzen erschweren sowohl die Resonanzzuordnung als auch Intensitätsbestimmung von Kern-Overhausereffekten. Ein Verfahren zur Auflösung von Zweideutigkeiten in der Resonanzzuordnung, implementiert im Programm ASNO, verwendet dazu eine vorläufige dreidimensionale Struktur (Berndt *et al.*, 1993).

Das Programm DIANA wurde für die Berechnung der dreidimensionalen Struktur von Proteinen aufgrund geometrischer Einschränkungen, d. h. oberer Distanzschranken aus Kern-Overhausereffekten, unterer sterischer Distanzschranken und Diederwinkeleinschränkungen aus skalaren Spin-Spin Kopplungskonstanten, entwickelt, wobei die einzigen Freiheitsgrade des Moleküls die drehbaren Diederwinkel sind. Eine Startkonformation mit zufälligen Einstellungen der Diederwinkel wird schrittweise an die Einschränkungen anpaßt (Güntert *et al.*, 1991a). Weil es sich dabei um ein kompliziertes Minimierungsproblem in einem hochdimensionalen Raum handelt, treten oft lokale Minima auf. Ein wirkungsvolles Verfahren, die Häufigkeit ihres Auftretens zu verringern und damit die Konvergenz zu beschleunigen, ist die Verwendung redundanter Diederwinkeleinschränkungen (Güntert & Wüthrich, 1991).

Eine Reihe von Hilfsprogrammen umgeben das zentrale Programm DIANA: CALIBA zur Umwandlung von Signalintensitäten in obere Distanzschranken, HABAS (Güntert *et al.*, 1989) und GLOMSA zur Bestimmung stereospezifischer Zuordnungen, deren Einfluß auf die Genauigkeit der berechneten Struktur diskutiert wird, und COFIMA zur Bearbeitung und Formatumwandlung verschiedener Datentypen (Güntert, 1992).

Die Strukturbestimmungen von zwei Proteinen — der *Antennapedia* (C39S) Homeodomäne (Güntert *et al.*, 1991b) und von Hirudin(1–51) (Szyperski *et al.*, 1993b) — illustrieren die Anwendung und Nützlichkeit der vorgestellten neuen Methoden zur Strukturbestimmung aufgrund von Kernresonanzdaten in Lösung.

Summary

This dissertation introduces new computational methods for the determination of three-dimensional protein structures from nuclear magnetic resonance data (Ernst *et al.*, 1987). The method introduced for this purpose a few years ago (Wüthrich, 1986) needs innovations and improvements at different stages — the nuclear magnetic resonance experiments, the processing and evaluation of spectra, and the structure calculations — in order to apply it efficiently to small proteins, or to larger proteins. The new methods developed during the dissertation are presented in the order in which they are employed in the course of a protein structure determination, and their use is explained by means of practical applications. The emphasis lies on the calculation of three-dimensional protein structures with distance geometry calculations.

The product operator formalism is an important tool to develop new schemes for nuclear magnetic resonance experiments. Its first complete implementation in *Mathematica*, a programming language designed for symbolic computation, simplifies the simulation of nuclear magnetic resonance experiments with spin-1/2 systems significantly (Güntert *et al.*, 1993).

The new program PROSA for the processing of multi-dimensional nuclear magnetic resonance data is presented. It provides efficient and flexible implementations of the common data processing steps such as linear prediction, digital filtering, Fourier transformation, and automatic phase and baseline correction (Güntert *et al.*, 1992). Due to a reliable automatic phase correction method no graphics display is needed, such that the program can operate on supercomputers for which it is especially optimized. New is also the FLATT method for baseline correction (Güntert & Wüthrich, 1992).

For the determination of the intensity of nuclear Overhauser effects a method is proposed that uses reference line-shapes from well-separated peaks in order to find the intensities of overlapping peaks (Eccles *et al.*, 1991). The measurement of scalar spin-spin coupling constants by inverse Fourier transformation of in-phase multiplets is introduced (Szyperski *et al.*, 1992a). Overlapping resonance frequencies make both the assignment and the quantification of nuclear Overhauser effects difficult. A method to resolve ambiguities in resonance assignments, implemented in the program ASNO, uses a preliminary three-dimensional structure (Berndt *et al.*, 1993).

The distance geometry program DIANA was developed for the calculation of three-dimensional protein structures on the basis of geometric restraints, *i. e.*, upper distance bounds from nuclear Overhauser effects, steric lower distance bounds, and dihedral angle constraints from scalar spin-spin coupling constants. The only degrees of freedom of the molecule are its rotatable dihedral angles. A start conformation with random values of the dihedral angles is adapted to the constraints in a step-wise fashion (Güntert *et al.*, 1991a). Local minima are abundant for such a complex optimization problem in a high-dimensional space. The use of redundant dihedral angle constraints is an effective method to improve the convergence by reducing the occurrence of local minima (Güntert & Wüthrich, 1991).

Several auxiliary programs surround the central program DIANA: CALIBA converts signal intensities into upper distance bounds, HABAS and GLOMSA determine stereospecific assignments, whose influence on the precision of the calculated structures is discussed, and COFIMA is used to manipulate different types of data that is relevant for the structure calculation (Güntert, 1992).

The structure determinations of two proteins — the *Antennapedia(C39S)* homeodomain (Güntert *et al.*, 1991b) and Hirudin(1–51) (Szyperski *et al.*, 1992b) — illustrate the application and advantages of the new methods for the structure determination on the basis of nuclear magnetic resonance data in solution.