Diss. ETH Nr. 7723

STRUKTURERMITTLUNG VON POLYPEPTIDEN UND KLEINEN PROTEINEN AUF DER BASIS VON ^1H-NMR DATEN MITTELS COMPUTER-GRAPHIK UND OPTIMIERUNG

ABHANDLUNG

zur Erlangung des

Titels eines Doktors der Naturwissenschaften der

EIDGENÖSSISCHEN TECHNISCHEN HOCHSCHULE

ZÜRICH

vorgelegt von

MARTIN URS BILLETER

Dipl. Phys. ETH

geboren am 13. Februar 1955

von Männedorf ZH und Basel-Stadt

Angenommen auf Antrag von

Prof. Dr. K. Wüthrich, Referent

Prof. Dr. M. Engeli, Korreferent

1985

ZUSAMMENFASSUNG

Die vorliegende Arbeit beschreibt verschiedene Methoden zur Bestimmung dreidimensionaler Strukturen von Polypeptiden und kleinen Proteinen aufgrund von ^{1}H-NMR-Experimenten, hauptsächlich Distanzmessungen zwischen Protonen sowie Angaben über Dihedralwinkel.

In engem Zusammenhang mit der Identifizierung von lokalen Struktureinheiten steht die sequentielle Zuordnung der Protonenresonanzen in ^{1}H-NMR Spektren. Verschiedene Distanzen zwischen Protonen benachbarter Aminosäuren nehmen in verschiedenen regulären Sekundärstrukturen kleine, charakteristische Werte an. Damit, sowie aufgrund der Eindeutigkeit von relativ kurzen Spinsystemsequenzen in der Aminosäuresequenz der untersuchten Proteine, kann im Allgemeinen die Mehrzahl der Resonanzen einzelnen Protonen bzw. Protonengruppen (CH_2,CH_3...) zugeordnet werden.

Dieselben sowie weitere Proton-Proton Kontakte erlauben ebenfalls eine Identifizierung regulärer Sekundärstrukturen. In helikalen Abschnitten der Sequenz treten insbesondere kurze Proton-Proton Distanzen zwischen den Aminosäuren i und i+j (j=2,3,4) auf. β-Blätter lassen sich durch kurze Distanzen zwischen dem α-Proton einer Aminosäure und dem Amidproton der nächsten sowie durch ein Netzwerk von kurzen Distanzen zwischen Protonen benachbarter β-Stränge erkennen. Turns unterscheiden sich von Helices nur durch die kleinere Anzahl betroffener

Aminosäuren. Sämtliche für die Identifizierung regulärer Sekundärstrukturen verwendeten sequentiellen Proton-Proton Kontakte können auch in Random Coil Strukturen kurze Distanzwerte annehmen; die regulären Sekundärstrukturen sind jedoch durch ein repetitives Muster dieser kurzen Distanzen charakterisiert. Weiter sind Helices kleine und β-Blättern grosse Kopplunskonstanten $^3J_{HN\alpha}$ eigen. Die kombinierte Anwendung aller Kriterien erlauben eine weitgehende Lokalisierung der regulären Sekundärstrukturen.

Ein weiteres Instrument zur Untersuchung dreidimensionaler Strukturen ist das Graphikprogramm CONFOR. Es verwendet spezielle Hardware, unter anderem ein "line drawing" System mit einem Farbmonitor. Proteine bis zu einer Grösse von 1000 Atomen können auf dem Schirm in verschiedenen Darstellungen betrachtet und dabei in "real time" bewegt werden. Weiter können bis zu acht beliebige Dihedralwinkel gewählt werden und ebenfalls in "real time" geändert werden. Zudem werden in den Strukturen Verletzungen oberer Distanzgrenzen (aus NOE-Messungen) sowie eine Auswahl von Kollisionen zwischen den durch ihren Van der Waals Radius bestimmten Atomkugeln angezeigt. Dies erlaubt gezielte Konformationsänderungen bis zur vollständigen Elimination der Verletzungen. Sämtliche Strukturdaten sind jederzeit zugänglich, jedoch ermöglichen zahlreiche Massnahmen eine Beschränkung der gezeichneten Strukturen und Daten auf das im Moment Wesentliche. Eine wichtige Anwendung von CONFOR ist die detaillierte Untersuchung von mit automatischen Methoden erhaltenen

Strukturen, sowie auch deren Vergleich mit weiteren Proteinen.

Als dritte Methode ist ein erst kürzlich beschriebener Optimierungsalgorithmus, der Ellipsoidalgorithmus, für die Berechnung von Polypeptidkonformationen eingesetzt worden. Er ist ursprünglich für die Lösung konvexer Probleme vorgesehen gewesen, zeigt aber bei nichtkonvexen Problemen aussergewöhnlich gute Konvergenzeigenschaften. Sein Prinzip ist sehr einfach. Im Konformationsraum wird das Volumen eines Ellipsoids, das die Lösung enthält, laufend reduziert. Dazu wird der Gradient im Mittelpunkt des Ellipsoides berechnet und ein neues Ellipsoid um die Hälfte des alten, die den negativen Gradienten enthält, gelegt. Folgende Eigenschaften unterscheiden diesen Algorithmus von anderen Optimierungsalgorithmen: Konvergenz wird erreicht durch die Volumenabnahme des Ellipsoides, welches mit einem konstanten Faktor, der nur von der Dimension des betrachteten Raumes abhängt, schrumpft. Die Richtung der einzelnen Iterationsschritte ist durch die Richtung des Gradienten und durch die Form des Ellipsoides, d.h. durch die früheren Iterationen, bestimmt. Randbedingungen (wie Distanzangaben aus NOE-Messungen) werden einzeln behandelt und nicht zu einem Pseudopotential aufsummiert. Dies erlaubt den parallelen Einsatz von Randbedingungen und einer Fehlerfunktion (z.B. Energie). Wie alle Anwendungen des Ellipsoidalgorithmus zeigen, wird bei konvergierenden Läufen das Resultat erreicht lange bevor das Ellipsoid so weit geschrumpft ist, das es im wesentlichen nur noch eine Konformation beinhaltet. Sämtliche in dieser Arbeit

besprochenen Anwendungen des Ellipsoidalgorithmus zielen auf ein globale Optimierung hin, d.h. es sind keine ausgewählten Startstrukturen verwendet worden.

Folgende Anwendungen dieser Methoden sind dargestellt: (i) Die Bestimmung der relativen Lage dreier Helices im DNA-Bindungsbereich des *lac* Repressors von E.coli anhand von 28 NOE-Distanzmessungen wurde unter ausschliesslicher Verwendung von CONFOR durchgeführt. (ii) Im Peptidhormon Glucagon wurde die Konformation eines Fragmentes bestehend aus elf Aminosäuren mit dem Ellipsoidalgorithmus bestimmt und mit früheren Distanzgeometrie-Rechnungen verglichen. (iii) Rechnungen am Polypeptid Conotoxin G1 berücksichtigten neben NOE-Daten auch zwei Disulfidbrücken sowie drei Kopplunskonstanten $^3J_{HN\alpha}$. Die zwei besten mit dem Ellipsoidalgorithmus berechneten Strukturen wurden interaktiv mit CONFOR noch weiter optimiert. (iv) Zuletzt ist noch der Einsatz von CONFOR bei der Bestimmung dreidimensionaler Strukturen vom Inhibitorprotein BUSI IIA erwähnt. Diese Strukturen wurden mit Distanzgeometrie-Rechnungen ermittelt, wobei CONFOR als komplementäres Instrument dieser automatischen Methoden eingesetzt wurde.

these short distances. Furthermore, helices are correlated with small and β-sheets with large coupling constants $^3J_{HN\alpha}$. With the combined use of all criterias most of the regular secondary structures can be localized.

The graphics program CONFOR is a further instrument for the work on three-dimensional structures. It works on the basis of specialized hardware, among which is a colour "line drawing" system. Proteins consisting of up to 1000 atoms can be displayed on the screen in various ways and can be rotated in real time. In addition, a maximum of eight dihedral angles can be varied simultaneously in real time. Violations of upper distance limits (from NOE-measurements) and of lower distance limits (sum of van der Waals radii) are monitored in the structures, thus allowing conformational changes to eliminate all violations. All structural data are accessible at any time, but there are numerous ways to reduce the visible structures and data to the actually essential information. An important use of CONFOR is the investigation of structures obtained with automatic algorithms and the comparison with other proteins.

As a third method a recently proposed optimization algorithm known as ellipsoid algorithm is used for the determination of polypeptide conformations. This algorithm was originally designed for convex problems, but very good results are obtained with nonconvex problems. The basic ideas are quite simple. The volume of an ellipsoid in the conformation space containing the solution is constantly reduced. For this purpose the gradient of the

these short distances. Furthermore, helices are correlated with small and β-sheets with large coupling constants $^3J_{HN\alpha}$. With the combined use of all criterias most of the regular secondary structures can be localized.

The graphics program CONFOR is a further instrument for the work on three-dimensional structures. It works on the basis of specialized hardware, among which is a colour "line drawing" system. Proteins consisting of up to 1000 atoms can be displayed on the screen in various ways and can be rotated in real time. In addition, a maximum of eight dihedral angles can be varied simultaneously in real time. Violations of upper distance limits (from NOE-measurements) and of lower distance limits (sum of van der Waals radii) are monitored in the structures, thus allowing conformational changes to eliminate all violations. All structural data are accessible at any time, but there are numerous ways to reduce the visible structures and data to the actually essential information. An important use of CONFOR is the investigation of structures obtained with automatic algorithms and the comparison with other proteins.

As a third method a recently proposed optimization algorithm known as ellipsoid algorithm is used for the determination of polypeptide conformations. This algorithm was originally designed for convex problems, but very good results are obtained with nonconvex problems. The basic ideas are quite simple. The volume of an ellipsoid in the conformation space containing the solution is constantly reduced. For this purpose the gradient of the

center point of the ellipsoid is used to compute a new ellipsoid enclosing the half of the old ellipsoid that contains the negative gradient. The following features are characteristic to the ellipsoid algorithm: Convergence is obtained through a constant decrease of the ellipsoid volume, the shrink factor being a sole function of the space dimension. The individual iteration steps are defined by the gradient direction and by the shape of the ellipsoid, i.e. by the previous iterations. Constraints (e.g. distance bounds from NOE-measurements) are considered individually thus avoiding the use of a pseudopotential. This allows the parallel use of constraints and an objective function like the energy. As all applications shows, the result in converging runs is obtained long before the ellipsoid has shrunk until it describes essentially one single conformation. In all applications discussed here, the ellipsoid algorithm has been used for global optimization, i.e. no selected starting conformations were used.

These methods have been tested on the following problems: (i) Three helices in the binding domain of *lac* repressors from E.coli have been positioned relative to each other with the sole use of CONFOR. (ii) The conformation of an eleven residue long fragment of the peptide hormon glucagon has been determined using the ellipsoid algorithm and compared to earlier results obtained with the use of distance geometry calculations. (iii) Work on the polypeptid conotoxin G1 was based on distance bounds from NOE-data and the known location of two disulfide bridges as well as

on three coupling constants $^3J_{HN\alpha}$. The two best structures computed with the ellipsoid algorithm were further optimized interactively with CONFOR. (iv) Finally, the use of CONFOR for the determination of three-dimensional structures of the inhibitor protein BUSI IIA is mentioned. These structures were computed using distance geometry, and CONFOR was helpful as a complementary instrument to the automatic methods.