Diss. ETH No. 18043

Automated assignment of amide resonances in NMR spectra of proteins with known crystal structure

A dissertation submitted to

ETH ZURICH

for the degree of

Doctor of Sciences

presented by

VENIAMIN GALIUS

dipl. Biophysiker, Humboldt-Universität zu Berlin

born on October 5, 1976

citizen of Germany

accepted on the recommendation of

Prof. Dr. Gerhard Wider, examiner

Prof. Dr. Frédéric Allain, co-examiner

2008

Summary

Major advances have been made in genomic studies in the last decade leading to a large number of newly identified protein sequences. For the functional and structural characterization of the encoded proteins fast methods are now required. X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy are the two main techniques dedicated to the structural analysis of proteins, and over 86% of all available protein structures have been obtained by X-ray crystallography. However, to reveal the function of a protein, not only its structure, but also its interactions with other molecules must be studied. Over the last decade solution NMR has developed into a powerful method for investigating biomolecular interactions, since it is very sensitive to conformational changes of the molecule. This salient feature made NMR an indispensable tool for rapid detection of binding and structural changes on atomic level both in protein folding and in interaction studies highly relevant in pharmaceutical research. For the complete interpretation of the information content of NMR spectra, e.g. for the mapping of binding-induced spectral changes to the molecular structure, assignment of signals (or resonances) in the spectra to the corresponding atoms in the molecule has to be carried out first. Resonance assignment normally represents the most time consuming and challenging task in the process of an NMR study, even in cases when the 3-dimensional structure of the protein is already available from X-ray crystallography.

For studies of molecular interactions it is often sufficient to assign only the resonances (or chemical shifts) of atoms in the protein backbone. Conventional backbone resonance assignment procedures rely on sets of triple-resonance experiments which correlate ^{1}H, ^{15}N and ^{13}C nuclei between neighboring amino acid residues in the primary sequence via covalent bonds in the protein backbone. In this way resonances of sequential residues can be identified and connected into fragments which are then mapped onto the primary sequence of the protein using amino acid specific chemical shifts derived from experimental statistics. With increasing molecular weight of the protein the conventional assignment approach may fail due to low sensitivity of triple-resonance experiments. The conventional assignment approach is also not applicable in situations when ^{13}C labeling is extremely costly, e.g. when expression has to be performed in eukaryotic cells. Conventional assignment approaches also do not integrate already available structural information.

The thesis describes a novel computer-based algorithm NAXERA which assigns backbone amide resonances of only ^{15}N-labeled proteins based on their known 3-dimensional structure and a single ^{15}N-resolved nuclear Overhauser enhancement (NOE) spectrum. The assignment is performed by matching interproton distance-networks obtained from an NOE spectrum and those derived from the crystal structure, while minimizing the difference between experimental and predicted amide chemical shifts. NAXERA is applicable to deuterated proteins in large molecular complexes, because the experimental input is limited to NOE-correlations between exchangeable amide protons, and the NOE experiments are sensitive also in large molecules.

The algorithm was validated using real and partially simulated data of 6 different proteins with molecular weights ranging from 8.5 to 110 kDa. For these proteins NAXERA yields assignments for 47 to 90% of observable amide resonances with 87 to 100% accuracy depending on the quality of the input. Robustness of the algorithm against incomplete or inaccurate input data has also been investigated in the thesis. The completeness and accuracy of the assignment result is shown to improve when additional experimental data from residue type specifically isotope-labeled samples or samples partially aligned in the magnetic field is used.

Partial alignment of the protein by means of various established liquid crystal media enables the observation of residual dipolar couplings (RDCs). However, RDC data can not be directly correlated to the molecular structure, because RDC values also depend on the magnitude and orientation of the alignment tensor with respect to the magnetic field, which is not known *a priori*. Only the magnitude of the alignment tensor but not its orientation can be reliably estimated from the RDC values of unassigned resonances using statistical methods or steric simulations. We developed a method for deriving orientation-independent assignment constraints from RDC data. The method does not require any computationally demanding optimization routines and thus can be easily incorporated into an automated assignment algorithm as shown for NAXERA.

NOE observation in spectra of large molecular systems is often prevented by signal overlap. We designed a sensitivity-optimized 2-dimensional ^{15}N-resolved NMR experiment, which resolves NOE-crosspeaks overlapped in conventional 2-dimensional NOE spectra. The utility

of the experiment for assignment of resonances in large molecules is demonstrated by its application to a 121 kDa protein-DNA complex.

Zusammenfassung

Im letzten Jahrzehnt wurden bedeutende Fortschritte in genomischen Studien erzielt, die zu einer grossen Zahl an neu identifizierten Proteinsequenzen geführt haben. Es sind nun schnelle Methoden für die Charakterisierung der Struktur und Funktion der neu identifizierten Proteine gefragt. Röntgenkristallographie und magnetische Kernspinresonanz (NMR) sind die zwei wichtigsten Verfahren für die Analyse der Proteinstruktur, wobei über 86% aller heute bekannten Proteinstrukturen durch Röntgenkristallographie gelöst wurden. Um die Funktion eines Enzyms jedoch aufzuklären, muss nicht nur seine Struktur, sondern auch seine Wechselwirkung mit anderen Molekülen untersucht werden. Lösungs-NMR hat sich in den letzten zehn Jahren zu einer wichtigen Methode für die Untersuchung biomolekularer Wechselwirkungen entwickelt. Dank der hohen Empfindlichkeit gegenüber konformationellen Änderungen wurde Lösungs-NMR zu einem unersetzlichen Werkzeug für die schnelle Detektion von Bindungsreaktionen und strukturellen Veränderungen auf atomarem Niveau sowohl in Proteinfaltungs- als auch in pharmakologisch hochrelevanten Wechselwirkungsstudien. Für eine vollständige Interpretation des Informationsgehalts von NMR Spektren, um z. B. durch Bindung hervorgerufene spektrale Änderungen auf das Molekülstrukturmodell zu übertragen, müssen zuerst die Signale (sprich Resonanzen) im Spektrum den entsprechenden Atomen im Molekül zugeordnet werden. Die Resonanzzuordnung stellt normalerweise den aufwendigsten und schwierigsten Teil einer NMR-Studie dar, selbst in Fällen, in denen die räumliche Struktur des Proteins durch Röntgenkristallographie bereits aufgeklärt wurde.

Für molekulare Wechselwirkungsstudien ist die Zuordnung von Resonanzen (d.h. chemischen Verschiebungen) der Atome aus dem Proteinrückgrat häufig ausreichend. Konventionelle Methoden der Resonanzzuordnung beruhen auf einer Reihe von Triple-Resonanz-Experimenten, welche die ^{1}H-, ^{15}N- and ^{13}C-Kerne aus sequenz-benachbarten Aminosäuren über kovalente Bindungen im Proteinrückgrat miteinander korrelieren. Auf diese Art können Resonanzen von sequentiellen Aminosäuren identifiziert und zu Fragmenten zusammengeschlossen werden, welche anschließend mittels statistischer Werte der aminosäurespezifischen chemischen Verschiebungen entlang der bekannten Aminosäuresequenz des Proteins positioniert werden. Mit wachsendem Molekulargewicht kann die konventionelle Zuordnungsmethode wegen der geringen Empfindlichkeit von Triple-

Resonanz-Experimenten versagen. Sie ist auch nicht brauchbar in Situationen, in denen die ^{13}C-Anreicherung extrem kostenintensiv wäre, wie z.B. bei Proteinexpression in eukaryotischen Zellen. Die konventionelle Methode sieht eine Verwendung der häufig bereits vorhandenen räumlichen Proteinstruktur bei der Zuordnung nicht vor.

In der vorliegenden Dissertationsschrift wird der neuartige Computer-Algorithmus NAXERA beschrieben, welcher Amidresonanzen aus dem Rückgrat eines lediglich mit dem ^{15}N-Isotop markierten Proteins auf der Basis einer vorhandenen dreidimensionalen Struktur und eines einzigen ^{15}N-aufgelösten nuclear Overhauser enhancement (NOE) Spektrums zuordnet. Der Algorithmus sucht dabei nach Übereinstimmungen zwischen den vom NOE Spektrum abgeleiteten und den aus der Kristallstruktur gewonnenen Protonendistanznetzwerken, womit der Unterschied zwischen gemessenen und vorhergesagten chemischen Verschiebungen der Amidgruppen minimiert wird. NAXERA kann für die Resonanzzuordnung von deuterierten Proteinen in großen molekularen Komplexen verwendet werden, weil der benötigte experimentelle Input auf NOE-Wechselwirkungen zwischen austauschbaren Amidprotonen beschränkt ist und NOE Experimente auch bei großen Molekülen empfindlich sind.

Der Algorithmus wurde mit realen Daten und teilweise simulierten Daten von insgesamt 6 Proteinen mit Molekulargewichten von 8.5 bis 110 kDa validiert. Bei diesen Proteinen liefert NAXERA Zuordnungen für 47-90% aller beobachtbaren Amidresonanzen, wobei der Anteil korrekter Zuordnungen von 87-100% in Abhängigkeit der Qualität der Inputdaten variierte. Die Toleranz der Methode gegenüber unvollständigen oder verrauschten Inputdaten wurde ebenfall untersucht. Die Vollständigkeit und die Richtigkeit der Zuordnung konnten durch Hinzunahme zusätzlicher experimenteller Daten von aminosäuretyp-spezifisch isotopenmarkierten oder teilweise im Magnetfeld ausgerichteten Proben verbessert werden.

Die partielle Ausrichtung von Proteinen im Magnetfeld mittels etablierter anisotroper Flüssigkristallmedien ermöglicht die Beobachtung von dipolaren Restkopplungen (RDCs). RDC-Werte sind winkelabhängig, können jedoch nicht direkt zur Molekülstruktur korreliert werden, weil sie ebenfalls vom Betrag und der Orientierung des Ausrichtungstensors abhängen, der *a priori* nicht bekannt ist. Nur die Amplitude, nicht aber die Orientierung des Tensors kann zuverlässig aus RDC-Werten von unzugeordneten Resonanzen durch statistische Analyse oder sterische Simulationen vorhergesagt werden. In der vorliegenden Arbeit wird eine Methode beschrieben, die es ohne Kenntnis der

Ausrichtungstensororientierung erlaubt, aus RDC-Werten der zugeordneten Aminosäurereste Einschränkungen der RDC-Werte für andere Aminosäurereste zu berechnen. Die Methode erfordert keine rechenintensiven Optimierungsroutinen und kann problemlos in einen automatisches Zuordnungsverfahren integriert werden, wie es am Beispiel von NAXERA demonstriert wird.

Die Beobachtung von NOEs in Spektren von großen Molekülen ist häufig durch Überlapp von Signalen erschwert, a. wenn die chemischen Verschiebungen der wechselwirkenden Protonen oder ihrer gebundenen Heteroatome ähnlich sind. Es wurde ein empfindlichkeitsoptimiertes zweidimensionales NMR Experiment für die Beobachtung von NOEs zwischen Amidprotonen, die an Stickstoffatome mit degenerierten chemischen Verschiebungen gebunden sind, entwickelt. Die Brauchbarkeit des Experiments für die Resonanzzuordnung in großen Proteinen wurde mit der Anwendung bei einem Protein-DNS-Komplex von 121 kDa gezeigt.