5 Zusammenfassung

Die NMR-spektroskopische Strukturaufklärung von Peptiden erfordert die Analyse komplexer, mehrdimensionaler Spektren. Hierfür müssen als erstes die Signale in den Spektren eindeutig zugeordnet werden. Typischerweise wird die Interpretation im NH-Bereich von TOCSY-Spektren begonnen, da hier eine relativ gute Dispersion vorliegt. Jede Aminosäure eines Peptids, mit Ausnahme von Prolin, zeigt eine charakteristische Spur. Auf dieser sind im Regelfall ausgehend von dem NH-Signal alle weiteren Signale der betreffenden Aminosäure zu finden. Sowohl die Lage der NH- als auch der Hα-Signale sind stark sequenzabhängig. Da die manuelle Zuordnung der Signale einen wesentlichen Zeitfaktor bei der Interpretation der Spektren darstellt, ist es von Interesse, diesen Vorgang zu automatisieren.

In dieser Arbeit wurde untersucht, in wie weit künstliche neuronale Netze in der Lage sind, diese Zuordnung durchzuführen oder zu erleichtern. Ziel war es, aus den automatisch generierten Listen der Peaks im TOCSY Spektrum einen Vorschlag für den Typ der Aminosäure und für die Position im Peptid zu erhalten.

Zunächst sollte aus diesen Peaklisten automatisch der Typ der Aminosäuren bestimmt werden. Hierzu wurden neuronale Netze entworfen, deren Eingabeneuronen die chemische Verschiebung repräsentieren. Da hier nur die Spuren im NH Bereich ausgewertet wurden, ist mit 650 Eingabeneuronen der Bereich von 0 bis 6.5 ppm abgebildet worden. Zum Training der neuronalen Netze wurden Daten aus der *BioMagResBank,* die die NMR Spektren von ca. 1700 Proteinen und Peptiden enthält, verwendet.

Zur Bestimmung des Aminosäuretyps wurden verschiedene neuronale Netze getestet. Der Hauptunterschied dieser Netze lag in dem Verfahren zur Erzeugung der Trainingsmuster. Ein Ansatz, bei dem die Trainingsdaten auf der beobachteten statistischen Verteilung der jeweiligen chemischen Verschiebung beruhten, ergab nur eine

Erkennungsrate von 35 %. Aminosäuren, deren Seitenketten sehr ähnliche NMR Spektren liefern, wie z.B. Asparaginsäure und Asparagin, wurden daraufhin in Gruppen zusammengefaßt, wodurch sich insgesamt zwölf Klassen ergaben. Hierdurch konnte die Erkennungsrate auf 60 % erhöht werden. In einem weiteren Ansatz wurden die Muster direkt aus den Spektren der Datenbank erzeugt. Dabei wurden aus einzelnen Mustern durch geringe Variation der Signale weitere Muster erzeugt. Hier wurden Erkennungsraten von 40 % bzw. 65 % bei Gruppierung der Aminosäuren erreicht. Der beste Ansatz benutzte die in der Datenbank abgelegten Spektren, denen künstlich eine größere, dreiecksförmige Linienbreite von 0.05 ppm gegeben wurde. Faßt man auch hier ähnliche Aminosäuren zusammen, so liefern die neuronalen Netze Erkennungsraten von 80 % bis 90 %, je nach präsentiertem Peptid.

Als zweiter Schritt bei einer automatischen Interpretation ist die Zuordnung der jeweiligen Aminosäure zu der sequentiellen Position notwendig. Hierfür wurden drei verschiedene Varianten getestet, um die Aminosäuresequenz auf der Eingabeschicht eines neuronalen Netzes abzubilden. Es zeigte sich, daß die besten Ergebnisse mit neuronalen Netzen erzielt werden, die auf die Vorhersage der chemischen Verschiebungen einzelner Aminosäuren spezialisiert sind. Die beste Kodierung für die Struktur der jeweiligen Seitenkette war eine Abfolge einzelner Bits, die die funktionellen Gruppen respektive Atome kodieren. Die Standardabweichung σ für die auftretenden Fehler bei der Vorhersage eines Testdatensatzes lag hier bei 0.57 ppm für Amidprotonen bzw. 0.38 ppm für Hα-Protonen. Um inkohärente Vorhersagen zu behandeln, wurde ein Zuordnungsalgorithmus, der die Vorhersagen von vier mit unterschiedlichen Mustern trainierten Netzen benutzt, erarbeitet. Zusätzlich zu den neuronalen Netzen wurde aus den verfügbaren Daten ein Inkrementsystem zur Berechnung der chemischen Verschiebungen von amidischen und Hα-Protonen entwickelt. Dieses Inkrementsystem kann die chemischen Verschiebungen mit einer ähnlichen Genauigkeit berechnen wie die oben erwähnten neuronalen Netze (σ_{NH} = 0.53 ppm, $\sigma_{H\alpha}$ = 0.37 ppm).

Diese Genauigkeit ist für eine akkurate Vorhersage der Kreuzsignale allerdings nicht ausreichend. Die vorhergesagte Position eines Signals kann jedoch als Startpunkt für eine Suche im Spektrum herangezogen werden. Da der Typ der gesuchten Aminosäure bekannt ist, kann das Kreuzsignal, das dieser Aminosäure entspricht und am dichtesten an der Vorhersage liegt, der entsprechenden Seitenkette in der Sequenz zugeordnet werden. Die beste Methode benutzt zur Berechnung jeweils vier auf einzelne Aminosäuren spezialisierte Netze und das Inkrementsystem. Damit können ca. 25 % der Signale zugeordnet werden (absolute Erkennungsrate). Von diesen Zuordnungen sind im Durchschnitt 61 % korrekt (relative Erkennungsrate).

Um die relative Erkennungsrate zu verbessern, wurden zusätzlich NOE-Daten verwendet. Diese geben weitere Informationen über Konnektivitäten innerhalb der Signale eines Spektrums. Dabei wurde nur der NH/Hα-Bereich in den entsprechenden NOESY-Spektren betrachtet und nur sequentielle NOEs berücksichtigt. Nach Tests zeigte sich, daß durch die zusätzliche Information falsche Zuordnungen weitgehend vermieden werden können. Signalzuordnungen, die durch Einsatz neuronaler Netze und der NOE-Information gewonnen werden, wurden akzeptiert, wenn mindestens 60 % der Methoden dieselbe Vorhersage lieferten. Dadurch steigt die absolute Erkennungsrate auf 31 % und die relative Erkennungsrate auf 91 %.

Neuronale Netze stellen somit ein Hilfsmittel dar, den Aminosäuretyp einer Spur automatisch zu bestimmen und für etwa ein Drittel der Aminosäuren auch deren Position in der Sequenz festzulegen. Mit diesen Eckpunkten ist eine weitere manuelle Zuordnung der Spektren wesentlich leichter und schneller durchführbar.

6 Summary

NMR based structure determination of peptides requires the analysis of complex, multidimensional spectra. First of all, the signals in the spectra must be unambiguously assigned. The NH region in TOCSY-spectra is usually the starting point for theses tasks due to the comparatively good dispersion in this area. Each amino acid other than proline gives rise to one characteristic trace in this region. Within this trace all signals belonging to one amino acid can be found emanating from the NH Signal. The chemical shifts of both NH and Hα protons depend strongly on the peptide sequence. Since the manual assignment of the signals is the major bottleneck in spectra analysis, an automation of this process is highly desirable.

The ability of artificial neural networks to perform or, failing this, facilitate this assignment was investigated in this work. Automatically generated peak lists from TOCSY spectra were to be used to obtain a sound proposal for the type of amino acid and the position within the sequence.

In the first step the type of amino acid should be determined automatically. To this end, artificial neural networks were designed which represent chemical shifts on their input layer. Since only traces within the NH region were analysed only the spectral range from 0 to 6.5 ppm was mapped to an input layer consisting of 650 neurons. Data for the training of neural networks was obtained from the *BioMagResBank*, which holds spectra of about 1700 proteins and peptides.

Various neural networks were tested to determine the amino acid type. The main difference between these networks was the method used in generating the patterns for training of the network. One approach, which used the observed statistical distribution of chemical shifts for the creation of patterns, resulted in recognition rates of 35 %. Some amino acids, e.g. aspartic acid and asparagine, give rise to nearly identical traces. Utilizing a classification of the 20 amino acids into subclasses

containing such similar residues, the recognition rate could be raised to 60 %. In another approach the patterns were generated out of the spectra deposited in the database. Here several additional patterns were generated out of one dataset by varying the chemical shift values by small amounts. The recognition rates of these networks were 40 % respectively 65 % when the aforementioned classification of amino acids was used. The best approach was to use the deposited spectra while artificially broadening the lines to 0.05 ppm with a superimposed triangle. Here the best recognition rates lie between 80 and 90 percent, depending on the presented peptide.

The second step in an automated interpretation is the sequential assignment of the amino acids. Three different coding schemes for the mapping of an amino acid sequence to the input layer of a neural network were developed for this task. Neural Networks, which were specialised for the prediction of chemical shifts of single amino acids, proved to perform best. The best representation of the structure of amino acids was a sequence of bits representing functional groups or atoms in the sidechain. The standard deviation σ of the error occurring when predicting a testset was 0.57 ppm for amidic protons and 0.38 ppm for Hα-protons. To treat incoherent predictions, an algorithm which averages the results of four different networks was developed. In addition to the neural networks an increment system for the fast calculation of chemical shifts was derived from the available data. The performance of this increment system was similar to the neural networks described before (σ_{NH} = 0.53 ppm, $\sigma_{H\alpha}$ = 0.37 ppm).

The precision of these methods is not sufficient for an accurate prediction of spectra. However, the predicted position of a crosspeak can serve as a starting point for a search within the spectrum. Since the type of the sought amino acid is known, the appropriate crosspeak which is nearest to the prediction can be assigned to the corresponding residue in the sequence. The best method uses four specialized neural networks and the increment system for each amino acid. With this combination, about 25 % of all signals can be assigned (absolute recognition rate). Out of

these assignments an average of 61 % are correct (relative recognition rate).

Additional NOE-data was used to increase the relative recognition rate. This data provides information about connectivities between signals in a spectrum. Only the NH/Hα-area was taken into account and all signals were taken to be sequential NOEs. Using this additional information, most wrong assignments could be prevented. Assignments, which were based on different neural nets and NOE-validations, were accepted when at least 60 % of all used methods yielded the same result. Thereby, the absolute recognition rate could be raised to 31 %. The relative recognition rate reached 91 % under this conditions.

Neural Networks thus can aid in the automatic determination of amino acid type and the sequential assignment of about one third of the TOCSY traces. The resulting assignments can be used as a vantage point for further analysis of a spectrum, which can thus be accomplished easier and faster.