Chapter 1

Introduction and Overview

Protein structure determination by NMR [76] has become an indispensable tool in biological laboratories and automation of the structure determination process has become an urgent necessity. In the context of genome projects, for example, structure determination on a large scale is required, which would not be practicable without a high degree of automation.

Modern high-resolution NMR generates spectra in 2, 3 or 4 dimensions. The peaks in these spectra represent magnetisation transfers between two or more nuclei, and are called *cross peaks*. The allocation of peak frequencies to specific nuclei is known as *assignment*. Having assigned the frequencies in the NMR spectra, a special type of spectrum, the NOESY, is used to determine interatomic distance values, from which a structure can be calculated.

Spectral assignment is one of the most time-consuming steps in the determination of protein structure by NMR. In 1986, Wüthrich published his groundbreaking book on NMR ([75]), in which a systematic approach to the assignment of spectra was laid out. It involved locating peaks in a key region of a spectrum, drawing horizontal and vertical lines from this peak until they met other peaks, then repeating the process with the newly found peaks. This *hierarchical* approach is still used even today. Its simplicity made it seem ripe for automation.

The earliest work led to the development of so-called "electronic drawing boards", such as ANSIG ([36]) or EASY ([17]). These allowed a user to sketch horizontal and vertical lines connecting peaks in spectra displayed on a computer screen, without the need for paper or light table. They also gave facilities for saving the assignments made to disk. Such programs are still widely used, for instance XEASY ([4]).

In parallel, hierarchical approaches to assignment using peak lists were being developed ([12], [16], [67], [7], [49], [77]), which attempted to apply the same ideas fully automatically to 2D spectra. The development of 3D and 4D spectra a few years later led to an extension of these 2D assignment strategies ([54], [68])

and thence to new programs for automated assignment ([33], [40]). Subsequent developments have been built on new algorithms, new types of spectra, and the incorporation of structural information.

In this work, a new approach to the assignment of NMR spectra is introduced. The structure of the thesis is as follows. In Chapter 2, the theoretical and practical background of NMR are discussed, with a special emphasis on assignment. Some of the earlier work in automated assignment mentioned above is also examined in more detail, and the drawbacks of these techniques are illustrated.

In Chapter 3, a program, *patt_recog*, is introduced, which modifies the spin system assignment step of the hierarchical assignment strategy, such that all operations within this step are performed concurrently. This means that peak-picking, spin-system detection and spin-system assignment are combined into a single procedure. The program does this by looking for patterns of peaks characteristic for known amino acids, and scoring the patterns it finds according to their closeness to the ideal pattern for a given amino acid. Extensive experimental data is presented to demonstrate the efficacy of this approach.

Chapter 4 discuses the automation of the sequential assignment problem. A second program, *chain*, is introduced, which can take result lists produced by *patt_recog*, and generate plausible sequential assignments based on them. It does this by growing fragments from all possible points along the sequence of the protein, taking the results from a search for *pairs* of residues in backbone spectra as building blocks. Experimental data is also presented.

The conclusions of this work are presented in Chapter 5.