Chapter 2

Automated Assignment of NMR Spectra

2.1 High-Resolution NMR of Proteins

2.1.1 Protein Structure Determination

NMR spectroscopy, X-ray crystallography and EM[1] spectroscopy are, at present, the most accurate means available to determine the three-dimensional structures of protein molecules to atomic scale. In many respects, these three techniques are complementary: NMR can be applied to soluble proteins (usually water-soluble proteins), crystallography to proteins which can be crystallised, and EM spectroscopy to proteins that can be regularly packed together in two dimensions, eg. on a lipid bilayer. NMR is limited to proteins smaller than 250 residues, the other techniques have, in principal, no upper size limit.

NMR techniques do not provide direct structural information. Instead, they yield large numbers of inter-atomic distance values, normally between protons, typically up to 5 pm[2]. These values are taken as constraints by *distance geometry* [8] or *molecular dynamics* programs, which use energy minimisation or simulated annealing techniques to obtain a structure.

2.1.2 The Basic Multidimensional NMR Experiment

The physical basis for the NMR experiment is described in detail in the first two chapters of [10]. Reviews of 2D techniques can be found in Kessler *et al*, 1988 [31] and Redfield, 1993 [60]. Oschkinat *et al*, 1994 [55] provide a comprehensive review of 3D techniques.

[1] Electron microscopy
[2] picometers

In order to be able to compare spectra produced on different spectrometers, which may have different base frequencies, a normalised measure of frequency has been established, the *chemical shift*:

$$\delta = 10^6(F_s - F)/F_b$$

where δ is the chemical shift, F the resonance frequency of a nucleus, F_s is the resonance frequency of a standard, eg. 3-trimethylsilylpropionate (TSP), and F_b the base frequency of the spectrometer. Chemical shift values are generally quoted in ppm or parts per million, hence the factor of 10^6 in the expression.

A set of connected spins which show scalar coupling is known as a *spin system*. In spectra such as TOCSY or COSY for example, where magnetisation is transferred through bonds, such spin systems manifest themselves as characteristic patterns of peaks. These are often recognisable in manual examination of the spectrum, and provide much useful information about the nature of the spin system, eg. to what *type* of amino acid it might belong.

Assignment is the process by which the frequency of each nucleus in the protein molecule is determined. This is done with the help of through-bond spectra, such as the 3D HCCH-COSY and HCCH-TOCSY for side chain assignment, and 3D HNCA and HN(CO)CA or CBCANH and CBCA(CO)NH for backbone assignment.

Spin systems determined via COSY or TOCSY spectra may be connected to each other via the inter-residual peaks in the backbone spectra, to effect *sequential assignment*. These peaks also form characteristic patterns (typically though, these patterns are simpler than in TOCSY or COSY spectra, since only backbone nuclei will be visible).

A 2D proton NOESY experiment can tell us about the through-space magnetisation transfers in a sample. The volume of a peak in such a spectrum is proportional to the distance between the contributing nuclei, to the power of minus six. This is the source of the distance constraints used by the distance geometry and molecular dynamics programs mentioned above.

2.2 3D and 4D Spectra: Advantages and Disadvantages

3D spectra provide a number of advantages over 2D spectra that make them particularly amenable to automated assignment procedures [55]. They also have a number of features that make manual assignment more difficult, and automated assignment more desirable.

A single peak in a 3D spectrum represents the magnetic interactions between three nuclei, and hence gives a relation between three chemical shifts. To obtain the same information from a 2D spectrum, one would need to find a pair of 2D cross peaks, having one chemical shift in common. Finding such pairs is not

usually completely straightforward, since, for instance, one of the chemical shifts in the pair may be degenerate with others, or two or more peaks may be very close together. Of course, ways may be found to circumvent such problems, eg. tracing through relayed peaks in a TOCSY spectrum, but such methods will still fail if the spectrum is very crowded. The 3D cross peak, on the other hand, gives the correlation between the three spins *a priori*.

3D spectra also have the advantage that they tend to "pull apart" peaks which, in a 2D spectrum, overlap. It means, for example, that peak shapes will be more predictable, a factor that greatly assists automated peak picking algorithms (see Figure 2.1).

And finally, 3D spectra may contain more redundancy than the corresponding 2D spectra, providing more scope for the application of noise reduction techniques.

Of course, all these good things don't come without a cost. Firstly, the time required to acquire a spectrum increases with the power of its dimensionality. A 3D spectrum can take many days to acquire, especially using a 600MHz spectrometer. The problem can be partially alleviated by restricting the data acquisition, such that only specific regions of interest in the spectrum are recorded. A second "trick" is to acquire at a high resolution along one of the axes (usually the time axis), and to acquire at a lower resolution, relatively speaking, along one or more of the other axes, and use linear prediction [2] to bring the resolution up to the same level as the primary axis. However, linear prediction can only extrapolate trends which already exist in the data, hence it is rather susceptible to noise.

The advantages of 3D significantly outweigh the disadvantages, however, and the technique is now widely used in the manual assignment of spectra; indeed, in many cases, with large proteins, it is not possible to assign spectra with 2D information alone. Both the advantages and the disadvantages of 3D are increased when one moves to 4D. In this case, computerised analysis becomes more than ever desirable, partly because of the sheer bulk of data present, and partly because of human difficulty in conceptualising 4D data spaces.

2.3 Spectra for Assignment

Nowadays, assignment of protein ^{1}H, ^{13}C and ^{15}N signals is accomplished by evaluating a set of a set of multidimensional NMR spectra, *viz.* HCCH-TOCSY ([19],[5]) and HCCH-COSY ([27], [25]) experiments, various versions of the X-filtered NOESY experiment ([20], [38]), and the so-called backbone experiments ([28]). The following is a typical set of spectra:

a: HCCH-COSY
HCCH-TOCSY
b: CBCANH

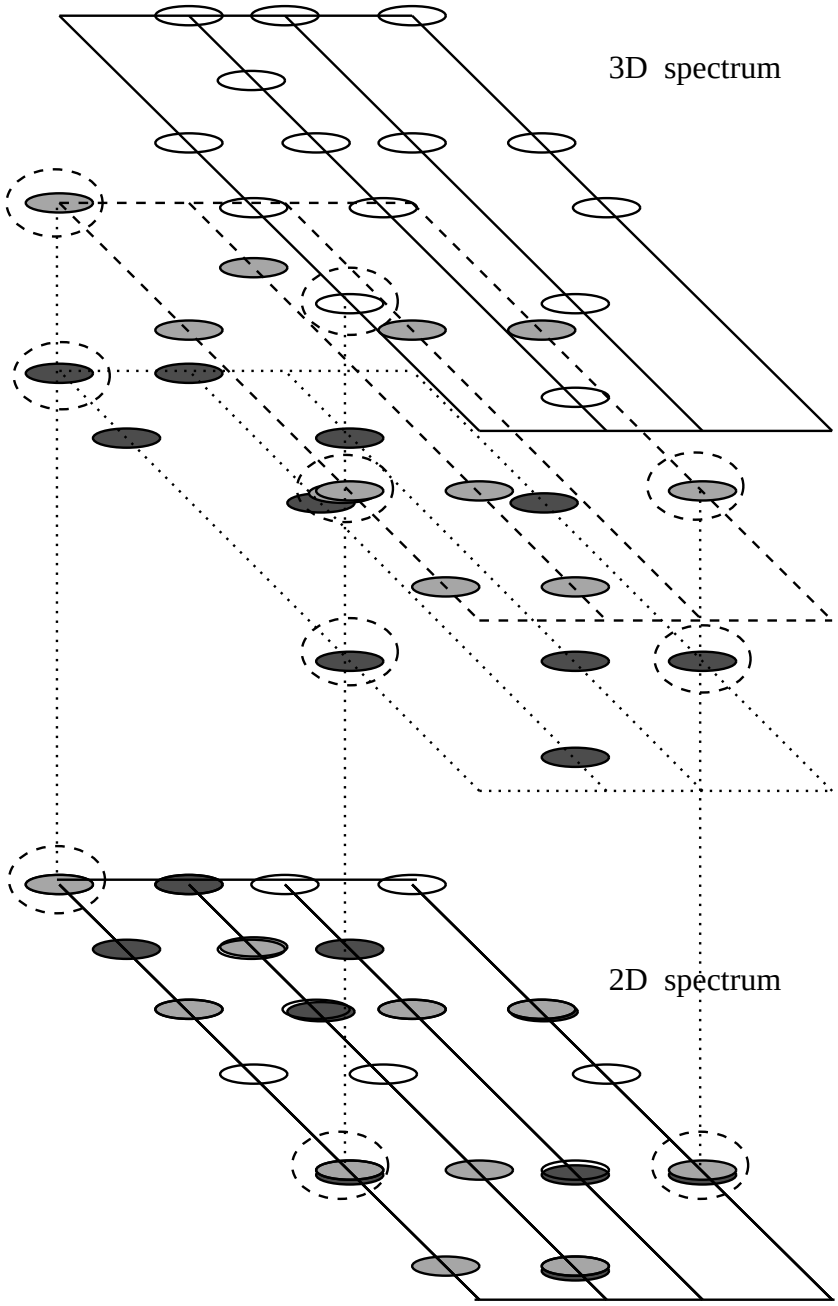

Figure 2.1: **Peak separation in 3D spectra.**
The ringed peaks, which are distinct in the 3D spectrum, overlap in the 2D spectrum.

CBCA(CO)NH
c: HBHA(CBCA)NH
HBHA(CBCACO)NH
d: HNCA
HN(CO)CA
e: HNCO
HN(CA)CO

These spectra may all be recorded using the same (^{13}C, ^{15}N - labelled) sample, or perhaps multiple samples, employing random fractional deuteration of non-exchangeable sites [50] as appropriate. In all cases, the cross-peaks indicate the presence of scalar couplings, and each individual combination or set of spectra reflects a particular pattern of couplings (with corresponding patterns of peaks) that can be found. One may search in all spectra simultaneously, or in various subsets. These subsets correspond to different starting points for the assignment procedure.

The *hierarchical* strategy, a generalisation of the strategy originally proposed by Wüthrich ([75]), starts by assigning as many spin systems as possible in set **a**, and then uses these to direct the search in the backbone sets **d** and/or **e** using a knowledge of the sequence as guidance. In a second cycle of assignment, this information can be used to constrain a new search in the side-chain spectra **a**, with relaxed searching conditions. One can also use spectra from sets **b** and **c** to obtain a more reliable connection between side-chain and backbone. This cyclic assignment procedure can be repeated as many times as required, to iteratively improve the global assignment.

Alternatively, one could start with the spectra in sets **d** and/or **e**, and perform a sequential assignment of the protein's backbone first, and then look into set **a** to find the sidechain assignments; this strategy is also widely used.

The choice of strategy for a given protein will depend on the complexity of the spectra and to some extent on the relaxation properties of the protein signals. Most effective are combinations which allow the correlation of three like chemical shifts to obtain sequential assignments and which provide in addition two or more chemical shifts which give handles on the side-chain. The combination of sets **a**,**b** and **e**, for example, fulfil these criteria, as backbone assignments can be obtained by correlating CO, Cα and Cβ resonances of neighbouring residues, and Cα and Cβ frequencies can be used to attach side-chain spin systems. For smaller proteins, sets **a** and **d** may be adequate, in other cases an individual combination may be chosen.

Extensions using multiplicity-edited experiments [62] and amino-acid specific backbone experiments ([14], [15]) (the two-dimensional form may suffice) may be incorporated as appropriate.

2.4 Processing Spectra Prior to Assignment

Multidimensional spectra generally contain many "artifacts", such as solvent lines, t_1/t_2 noise and unresolved signals, especially around the diagonal line (or plane, in the case of 3D spectra). Even to an experienced spectroscopist, these features can be very confusing. For automated procedures, they also represent a problem. Hence, before any assignment can be attempted, a number of preprocessing steps are usually applied to the source data, with the intention of removing features which do not contribute any useful information.

Before Fourier transformation of the time domain data, several preprocessing options are available. For instance, it is likely that the user will want to increase the nominal resolution of the spectrum with the aid of zero-filling or linear prediction. Noise and baseline distortion produced by the solvent, usually H_2O, can be significantly diminished by use of the Karhunen-Loéve transformation [43]. After Fourier transformation, noise can be reduced by thresholding, and linear or non-linear filtering techniques. Ridges parallel to axes, such as those produced by t_1 "noise" are usually removed interactively, by having the user select rows in the spectrum that contain no peaks, just ridge, and subtracting from the rest of the spectrum. See, for instance, Glaser and Kalbitzer, 1986 [22] and Neidig *et al*, 1990 [47].

2.5 Strategies for Automated Assignment

Many different approaches to the automated assignment problem have been applied - for more complete reviews, see Kalbitzer *et al*, 1990 [26], Hoch *et al*, 1991 [24] and Zimmerman and Montelione, 1995 [80]. There are a number of characteristics which allow us to distinguish existing programs:

1. Degree and flexibility of automation. Most programs allow some user intervention. Some do this in a fairly crude way - eg. the software of van de Ven, 1990 [67] generates output files at various stages in its execution, which can be user-edited before execution is restarted. This allows filtering and correction of automatically generated results. On the other hand, Eccles *et al*, 1991 ([17] provide an comprehensive interactive environment within which to perform assignment; automated assignment is simply a component of this environment. Many programs only automate part of the assignment problem, for instance finding spin systems [71], or sequential assignment ([6],[44],[9]).

2. Dimensionality of the spectra that the programs can deal with. Some can deal only with two dimensional spectra (eg. Eads and Kuntz, 1989 [16]), though programs capable of dealing with three or more dimensions are now the norm (eg. Kleijwegt, 1991 [33]).

3. Type of spectra used. Spectra containing intra-residual (eg. COSY or TOCSY) and/or inter-residual (eg. HNCA or NOESY) information may be used. The latter type of spectrum is used to obtain information on sequential connections between residues, plus (possibly) secondary structure information ([53], [74], [73]). If isotopically labelled proteins are available [39], the inter-residue J-coupling between ^{13}C and ^{15}N can be exploited at the sequential assignment stage; this has an advantage over NOE data, in the sense that interactions between non-sequential residues can be eliminated. This also simplifies the data which is subsequently fed into the automated assignment program, and makes the assignment process less dependent on the structure of the protein.

4. Extent to which structural information is utilised in assignment. This ranges from none at all (by taking advantage of ^{13}C and ^{15}N couplings, as discussed in section 2.1.2), to conventional sequential assignment programs relying on some secondary structure, such as that of Cieslar *et al*, 1988 [12], to those that rely almost entirely on structural information, such as Oshiro and Kuntz, 1993 [56].

5. Some programs put a special emphasis on particular kinds of computing technique, for instance, expert systems ([78], [79]), neural nets [66], simulated annealing [6] or genetic algorithms [72].

2.5.1 Peak Picking Techniques

The features of especial interest to spectroscopists are, of course, the peaks in the spectrum. Hence, it has been seen as a natural data abstraction step to generate lists of peaks from spectra as input for automated assignment. Quite a few different approaches have been tried. The simplest approaches would be either to pick all points above a given threshold, or to use a maxima detecting routine. These approaches tend to generate vast numbers of peaks, most of them noise. Hence, more sophisticated approaches have been developed. For instance, using a user-defined shape (eg. ellipsoid) and searching for peaks of that shape in the spectrum [21]. Or using a library of example peaks, and comparing actual peaks to example peaks to generate a measure of "goodness" [35]. Or using learning algorithms, such as neural nets, to find peaks, after first training them with examples of good and bad peaks ([32], [13]). See also Neidig and Kalbitzer, 1990 [46] and Stoven *et al*, 1989 [65]. Some of these techniques are summarised in Figure 2.2

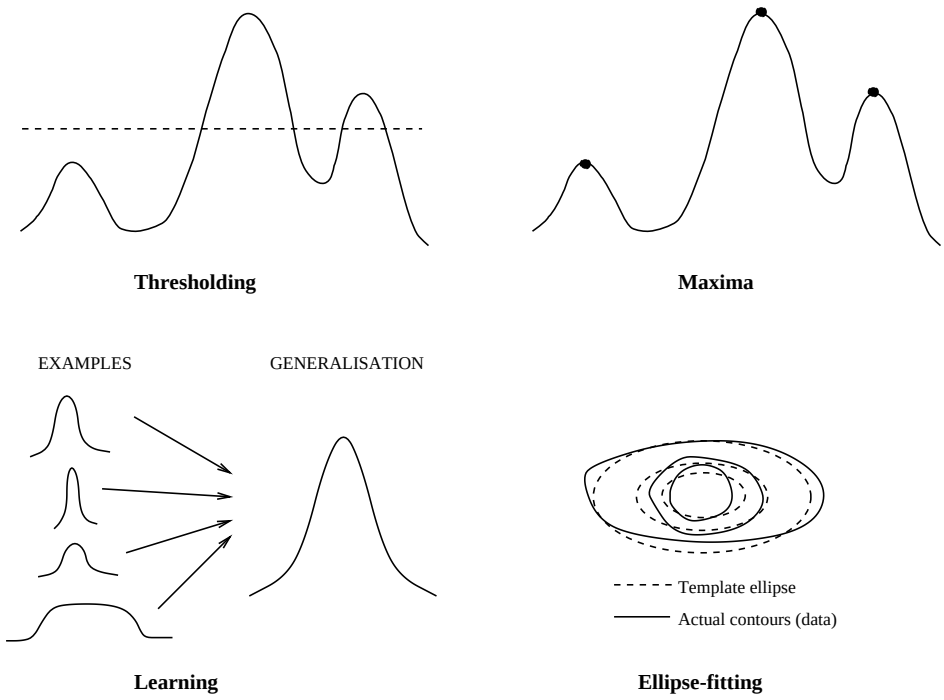

Figure 2.2: **Some peak picking techniques.**
Thresholding and maximum detection are relatively crude methods, which tend to deliver noise and artifact peaks as well as signals. Learning techniques or ellipse-fitting allow the user more control over the peak profile that the peak picker will use, and hence improve the quality of the final set of picked peaks.

2.5.2 Some Assignment Strategies

Cross Peak Fine Structure

In spectra of sufficiently high resolution, the fine structure in 2D cross peaks caused by spin-spin couplings may be modified to produce quite characteristic and distinct patterns of line splittings, which can be identified by pattern matching techniques ([58], [57] , [41], [37], [18]). These kinds of techniques are useful for small molecules, such as short polypeptides, and have been fairly extensively explored, but are generally not applicable to proteins, because the fine structure is usually absent in protein spectra.

Hierarchical Assignment

Cieslar et al, Eads et al, and others, have made attempts to automate the traditional hierarchical assignment approach ([76], [12], [16], [68]). Such programs search first for spin systems in through-bond spectra (eg. COSY). They then combine them, using through-space connectivities (eg. from a NOESY spectrum) or through-bond connectivities in doubly labelled samples (eg. HNCA), in the sequential assignment. Some information on secondary structure is usually employed by these programs.

All attempts at automating the complete assignment of proteins have, so far, concentrated on the automation of the hierarchical assignment methodology. Two somewhat different approaches will be given as examples:

- **A Graph-Theoretical Approach.** For example, Oschkinat *et al*, 1991 [54] have developed a program which works on 3D TOCSY-TOCSY and 3D NOESY-NOESY spectra. First it looks for spin system patterns in a peak list extracted from the TOCSY-TOCSY spectrum, using techniques taken from graph theory. This is a 2-step process: pattern subunits (corresponding to quadruplets of protons) are first identified in the spectrum; they are represented by the program internally as small graphs. This is illustrated in Figure 2.3. Then these subunits are assembled into larger groups (spin systems) via shared 3D cross peaks, using graph combinatorial methods. The NH, Hα, and Hβ resonances of these spin systems are identified by peak intensities. Assignment of the spin systems to specific amino acids is done manually by examining the side-chain lengths and chemical shifts. The sequential phase of the assignment then commences, by first pairing residues via the TOCSY-NOESY spectrum, using NOEs from NH protons in one residue to various main chain protons (ie. NH, Hα and Hβ) in another residue; then, the pairs are connected to form plausible chains. Only chains which fit the sequence for the protein are retained. An almost complete recursive search is done for all possible chains of pairs of residues, with the exception that if two possibilities share a peak, one of the possibilities is eliminated. The connectivities of the residues are given

ratings at various levels, which, when combined, allow the best chains of residues to be selected.

- **The CLAIRE Program Suite.** Another approach, also employing the hierarchical assignment methodology, is that of Kleywegt *et al*, 1990 [34]. They present a package of programs for automated assignment. The starting point is the generation of peak patterns, corresponding to the main chain portions of residue spin systems. This algorithm uses a 2D HOHAHA spectrum, searching first for all peaks in the Hα/NH region. Each peak is used as a seed point to search for further peaks, eg. Hβ/Hα or Hβ/NH peaks, in a systematic way. This is illustrated in Figure 2.4. Then, possible patterns are scored according to various measures, including number of cross peaks (as a fraction of the total possible number). Patterns containing similar chemical shift values to existing patterns are discarded, as are low-scoring patterns. Pattern pairs are generated, using NOE information. Sequential assignment is done by using the pattern pairs, plus additional information about expected chemical shift values obtained from Groß and Kalbitzer, 1988 [23].

Assignment Using Backbone Spectra from Doubly-Labelled Samples

Producing labelled proteins has become easier and cheaper over the past decade, and a whole family of new experiments has grown up to exploit this, eg. HNCA/HN(CO)CA, HNCO/HN(CA)CO, etc. Numerous automated assignment programs have been written to use these types of spectra. Physical and technical limitations mean that, so far, these experiments are limited to spins on or close to the backbone - side chain data beyond Hβ/Cβ cannot be provided. This places an inherent limitation on the scope of the information that can be extracted.

An early example of such a project is Powers *el al*, 1992 [59]. They use extensive isotope labelling and many different 3D experiments. Three programs are reported: CAPP, PIPP and PEAK-SORT. CAPP is an automated peak picker, which identifies peaks by fitting ellipses to contours. PIPP is a manual peak list editor, which allows removal of, or addition to, peaks picked by CAPP. PEAK-SORT first finds intra-residue connectivities by comparing picked peaks in various spectra. The spectra are specific enough that the nuclei between which magnetic transfers takes place can be specified. Hence, construction of spin systems is relatively straightforward. Having done this, the program searches for inter-residue connectivities, generating a list of pairs of residues, the residues being arbitrarily numbered. The program is capable of chaining these residues together, but degeneracy makes such automatically generated chains suspect, and manual cross-checking is required to make the procedure reliable.

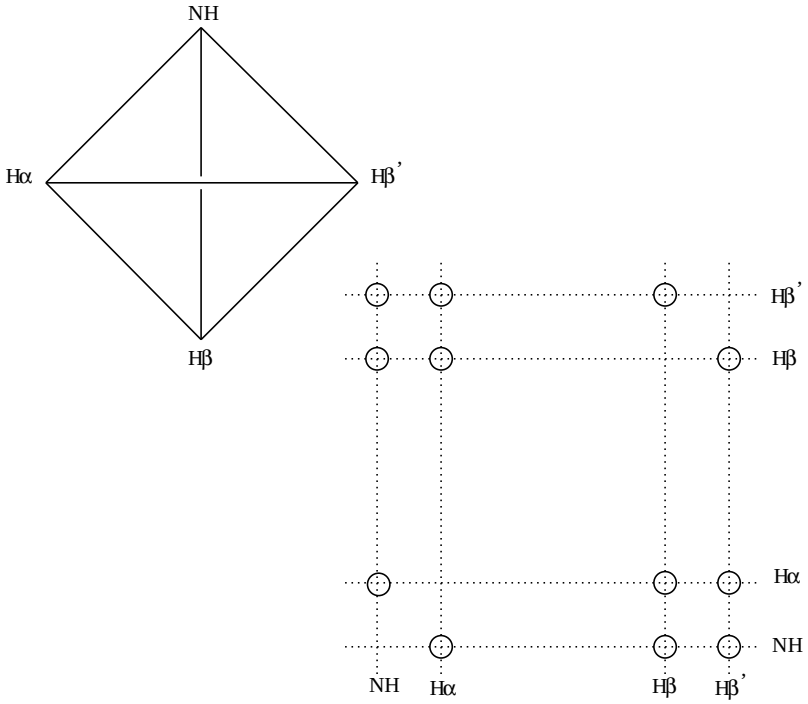

Figure 2.3: **Representing a group of spins as a graph.**
The graph in the top left hand side of the figure can be translated into the 2D TOCSY search pattern in the bottom right of the figure. The arcs in the graph allow us to determine which spins are magnetically coupled; an undirected arc means that magnetisation can be transferred in both directions.

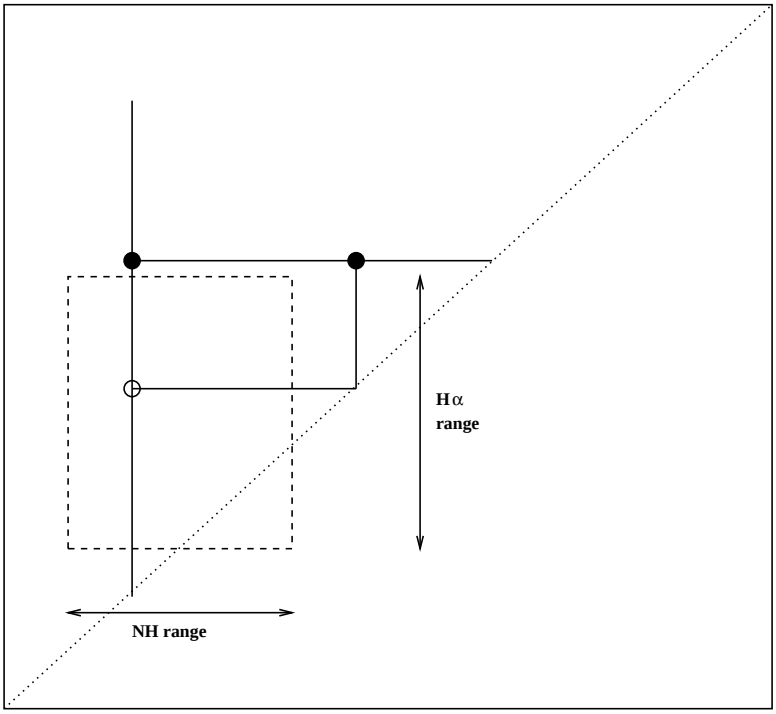

Figure 2.4: **Systematic search for spin system peaks.**
The figure shows a notional HOHAHA spectrum. The region enclosed by a dashed box is the Hα/NH region, and the empty circle represents one of the peaks found by the program in this region. The program then searches upwards at the Hα chemical shift, to see if it encounters a peak. If so, it also checks for a peak at the same chemical shift on the NH line (shown in black). The program continues in this manner until no more peaks can be found.

Automated-Main Chain Directed Assignment

At least one attempt has been made to automate the main-chain directed assignment strategy of Englander and Wand ([70], [64], [49]). It requires COSY and NOESY data. From the COSY data, peak picking is done and then the main-chain subsets of all spin systems, ie. the peaks from the NH, Hα and Hβ protons, are found. Now the program looks for patterns in the NOESY data. It looks for characteristic patterns of NOESY peaks associated with i) α helices, ii) anti-parallel β sheets, and iii) parallel β sheets (in that order), ie. it uses secondary structure in the assignment process. At each stage in the search, successfully identified patterns allow sets of peaks to be deleted from the NOESY data, thus reducing processing at subsequent stages. Overlapping patterns can be fitted together. The characteristic patterns mentioned above contain between 3 and 8 NOE contacts.

Structure Based

In principle, given a good enough NOESY spectrum, one could determine the structure of the protein under study without any prior assignments, and then use the spatial distribution of the protons to determine how they should be assigned.

A number of programs have been written based on this principle ([42], [56], [45], [51], [52] and [3]). Because of the noise, artifacts and ambiguity inherent in NOE data, all of these programs need additional data. Some methods require a structure from a homologous protein whose structure is already known as a starting point for the distance geometry calculations; others require that some spins already be assigned by other techniques as initial constraints. Raw NOESY data is never used as input; in all cases, a list of picked peaks is needed. This list is usually manually edited before being passed to the program. The advantage of this approach is that structure determination and assignment are tightly coupled with each other; this imposes constraints on assignments which are not available in other methods.

Sequential Assignment

Most sequential assignment programs described in the literature use a maximisation algorithm to obtain a best fit between a set of spin systems and the sequence for the protein.

One such project is that of Bernstein *el al*, 1993 [6]. Input to the program is a set of pre-assigned spin-systems, and tables of NOE contacts between them. These are obtained manually from ^{15}N HMQC-TOCSY and ^{15}N HMQC-NOESY spectra. The method uses an "energy minimisation" technique. Initially, the known spin systems are assigned randomly to the residues. A score, measuring the "goodness-of-fit", is computed for each spin-system/residue match. This score can be based on many things, but at a minimum, it will be

based on i) the similarity between the spin system and the residue, and ii) the strength of the NOE connections to the preceding residue in the chain. Once the initial energy has been computed, as the sum of all these scores, the minimisation proceeds. Two short segments from random points in the sequence are selected. Within each, the assignments are systematically rearranged, such as to minimise the global energy measure. The process is repeated, for some 10^5 cycles, until no further energy minimisation occurs.

2.6 Design Criteria

The long-term aim of such developments is the construction of a system which can perform the complete assignment process automatically, with little or no manual interference. A number of fundamental criteria should be taken into consideration when designing the architecture of such a system:

- A set of modular tools should perform the basic assignment tasks. Each tool should deliver results weighted according to quality factors.
- The tools should have a low sensitivity to baseline offsets, t_1/t_2-noise, small positional differences between resonances in different spectra, spectral resolution, the line widths of the peaks and the signal to noise ratio.
- Where possible, iterative adjustment of parameters should be performed by the software automatically.

These demands on the system would require that the maximum possible amount of information, as contained in the spectra themselves, be retained at all stages. Traditionally, assignment procedures rely on peak lists extracted from the multidimensional spectra. This approach has a number of problems. Firstly, in spectra with low S/N or strong t_1/t_2-noise, a considerable amount of spurious data may be generated, which would make the successful application of any combinatorial approach difficult. Secondly, in crowded spectra containing significant spectral overlap, peaks tend to merge together, and it is difficult to interpolate correct cross-peak positions in areas where this has occurred. Hence a peak list obtained from these spectra can be incomplete, and very likely imprecise with regard to picked frequencies, at important points. For this reason, highly automated procedures should work with the original spectra.

2.7 Using Patterns to Guide Search

Frequently, automated assignment programs reported in the literature must have their input extensively modified by hand before they can do anything useful. One goal of the work presented in this thesis is to remove this kind of manual intervention; the program takes normal spectra as input.