Chapter 3

Spin System Assignment

3.1 Implementation and Practical Considerations

3.1.1 Introduction

In many types of spectrum, peak patterns provide important information about the spin systems present. A program called *patt_recog* has been designed to search for these patterns in spectra. It may be applied to sidechain spectra such as COSY or TOCSY, and also to backbone spectra such as HNCA or HN(CO)CA. It yields results (putative chemical shift assignments), weighted by probability-like factors, which give estimates of correctness for the results.

As input, the spectra and the expected peak topology need to be supplied to the program. The peak topology, or pattern, describes which peaks are to be expected in the spectra; this information resides in a *pattern file*. Many patterns may be stored in a single pattern file, eg. patterns for alanine spin systems, threonine spin systems, etc. For each expected peak in a pattern, a *region of interest* is determined by the pattern file. This is a rectangular or cuboid portion of a spectrum, whose size is bounded by the chemical shift ranges of the spins contributing to that peak.

For each pattern, the program steps through three principal operations. These are:

- peak emphasis filtering;
- pattern search;
- heuristic filtering of results.

In the first step, a mask sensitive to the most probable peak shape is scanned through each region of interest. A data array is obtained, which assigns a

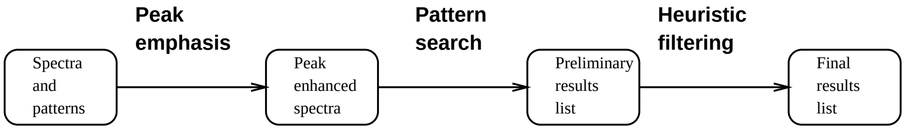


Figure 3.1: **Program flow diagram.**
The input required by the program consists of *spectra* and *pattern files*. *Peak emphasis* finds the peak plausibility factors in selected portions of the spectra (regions of interest). *Pattern search* exhaustively explores the chemical shift space for peak patterns within the regions of interest (see also Figure 3.6). A *heuristic filter* applies rules typically used by spectroscopists, in order to penalise or remove unlikely results.

number to each point in a region, quantifying the *plausibility* of there being a peak at that point. This quantity will be referred to as the *mask response.*

In the second step, a pattern search is performed, which looks for instances of the user-defined pattern within the regions of interest. Each found pattern, or result, has a *score* associated with it, which is calculated as a sum of mask response values. To accelerate execution time, thresholding may be applied to these mask response values in a dynamic way during the search. Internally, the search is implemented as a depth-first recursion, and thresholding allows a significant pruning of the search space. The results are put into a preliminary *results list*, which is ordered according to score.

Finally, heuristic filtering is applied to the preliminary result list, invoking rules that reorder or delete results, according to, for instance, chemical shift order, relative amplitudes of individual mask responses, etc. A program flow diagram is shown in Figure 3.1. The following section briefly discusses the pattern file; the subsequent three sections are devoted to the three stages of execution.

Figure 3.2 tabulates the symbols which will be used in the algorithms described below.

Symbol	Meaning
a	Current spectrum axis
A	Total number of axes (dimensions) in a spectrum
δ	A single chemical shift
$\underline{\delta}$	Chemical shift vector $\underline{\delta} = [\delta_1, \delta_2, \ ... \ \delta_N]$
Δ	Chemical shift difference
e	Current result number
E	Total results count
F	Penalisation factor ($0 \leq F \leq 1$)
k	Deletion criterion (logical, **true**=delete, **false**=retain)
$\mathbf{M}$	Matrix to map chemical shift list into a set of coordinates
n	Current spin number
N	Total number of spins in a spin system
p	Current region of interest number
P	Total region of interest count for a spin system
r	Current mask response
$\underline{r}$	List of all mask responses for a given value of $\underline{\delta}$ $\underline{r} = [r_1, r_2, \ ... \ r_P]$
R	Sum of mask responses (score) for a spin system
$\underline{R}$	List of all scores for all results $\underline{R} = [R_1, R_2, \ ... \ R_E]$ N.B. since results are ordered according to score, R_1 will be the minimum score value, and R_E will be the maximum.

Figure 3.2: **Symbols used in pattern search algorithms.**

3.1.2 The Pattern File

This file describes in detail the data files, peak topologies and parameters needed during the pattern search. Many "template" pattern files exist, which may be customised to suit individual requirements. The syntax of a pattern file is quite flexible; the *patt_recog* program has a built-in lexical analyser and parser that allow these files to be transcribed into internal data structures.

Pattern files divide into two parts, a global part, and a set of pattern definitions. The global part may contain any or all of the following:

- Information about the spectra being examined, including location on disk, spectrum type and size-related parameters;
- Directives to modify spectra non-linearly, both before or after peak emphasis processing;
- Parameters relating to the results generated, including maximum and minimum allowed results counts, and distance thresholds for the results clustering algorithm;
- Parameters for results list processing algorithms.

The pattern definitions may contain any or all of the following:

- Definitions of individual cross peaks, in terms of the spins contributing to them;
- Definitions of the masks to be used during peak emphasis filtering;
- Threshold values;
- Directives dividing the search into multiple levels;
- Limits on the search chemical shift ranges for each spin;
- Parameters for results list processing algorithms.

An example of the peaks that one might expect to find for threonine in a 2D TOCSY experiment is given in Figure 3.3; Figure 3.4 shows the pattern definition which would be used to search for these peaks.

3.1.3 Peak Emphasis Filtering

The pattern search takes appropriately processed data points directly from spectra. These spectra are first filtered, to emphasise genuine peaks and to de-emphasise features such as broad peaks, noise or variations in baseline. The filtering procedure works by calculating a *peak plausibility factor* for every point in all regions of interest in all spectra.

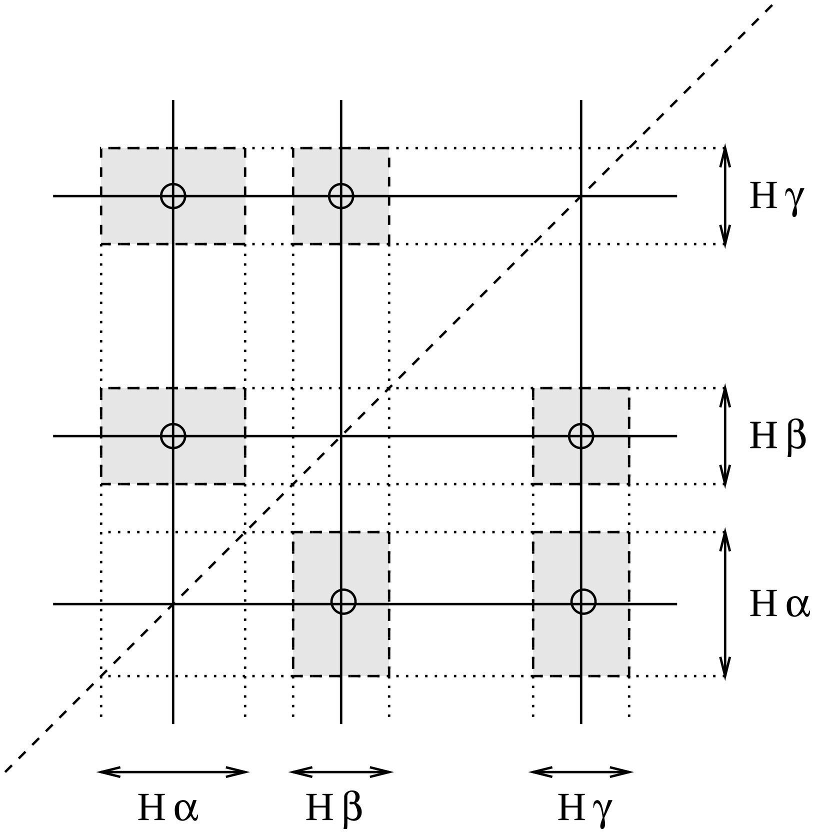


Figure 3.3: **Idealised peak pattern for threonine in a 2D TOCSY spectrum.** The small circles indicate *which* peaks are expected, and the grey patches show the *areas* in which the program will search for them. These areas are called *regions of interest*. The chemical shift ranges for each spin type determine their boundaries. In this example, the regions of interest in a 2D spectrum for peaks between Hα, Hβ and Hγ are shown. For clarity, these regions have been shown as distinct patches, but in reality, there will often be overlap.

```
Thr_spin_sys1
{
    /*          TOCSY SPECTRUM                */
    current_spect=“tocsy.2rr”;

    search_level=0;

    cross_peak=[[Thr,H_gamma_2-3],[Thr,H_beta]];
    cross_peak=[[Thr,H_beta],[Thr,H_gamma_2-3]];

    search_level=1;

    cross_peak=[[Thr,H_beta],[Thr,H_alpha]];
    cross_peak=[[Thr,H_alpha],[Thr,H_beta]];

    search_level=2;

    cross_peak=[[Thr,H_gamma_2-3],[Thr,H_alpha]];
    cross_peak=[[Thr,H_alpha],[Thr,H_gamma_2-3]];


    spin_ppm_range=[“Thr”,“H_alpha”,3.00,6.00];
    spin_ppm_range=[“Thr”,“H_beta”,3.40,5.00];
    spin_ppm_range=[“Thr”,“H_gamma_2-3”,0.05,1.57];
}
```

Figure 3.4: **Example pattern file.**
Cross peaks are specified in an abstract way, according to the spins contributing to them. The search process is split into multiple levels. The number of peaks searched for at a given level is limited, and the chemical shift values found at one level will constrain the search at the next one. Limits on the chemical shift ranges to be searched for each class of spin can be imposed also.

The peak plausibility factor of a given point in a spectrum is a measure which combines *peak intensity* at this point with the *degree of fit* between an ideal peak profile and the actual data points in the immediate neighbourhood of the point. This is done by convolving a mask function with the original spectrum [61].

Convolving a Spectrum With a Mask Function

For the 1D case, one may express the convolution operation quite generally thus:

$$f'(x) = \int_{-\infty}^{\infty} f(x - \epsilon)h(\epsilon)d\epsilon$$

where $f(x)$ is a function describing the original spectrum, $f'(x)$ is the spectrum after convolution, and $h(\epsilon)$ is the the masking function.

The convolution operation has a very interesting property: in regions of $f(x)$ which are similar to $h(\epsilon)$, $f'(x)$ (the mask response) will be large, whereas in regions of $f(x)$ which are not similar to $h(\epsilon)$, $f'(x)$ will be small. Hence, in order to find peaks, one chooses a function $h(\epsilon)$ which models an ideal peak. For instance, a Lorentzian function:

$$h(\epsilon) = 1/(1 + s\epsilon^2)$$

Or a Gaussian:

$$h(x) = e^{-\epsilon^2/2s^2}$$

In both cases, s is a factor which describes the half width of the model peak.

This is not quite ideal, though, because if such a mask function is convolved with a nonzero but constant spectrum $f(x) = c$, the mask response will also be nonzero. Practically speaking, this means that the mask response would be baseline dependent. In order to make the mask *zero sum*, the mean mask height is subtracted from the mask function during convolution:

$$f'(x) = \int_{-\infty}^{\infty} f(x - \epsilon)(h(\epsilon) - h_{mean})d\epsilon$$

For the finite discrete case, we can write:

$$f'(m) = \sum_{i=0}^{K} f(m - i)(h(i) - h_{mean})$$

$$h_{mean} = (\sum_{i=0}^{K} h(i))/K$$

where K is the size of the mask in data points.

Figure 3.5: **Convolution of a spectrum with peak detecting mask.** The original 1D spectrum is shown as a dashed line; the solid line shows the result of the convolution operation. Two properties of this convolution are worth noting. At position (a), an artificial broad peak has been inserted into the spectrum. After convolution with a narrow mask, this peak has been significantly flattened. At position (b), two peaks in the original spectrum are so close together that they have merged. Using the convolution approach, this clearly resolves into two peaks. The operation of the convolution algorithm is shown in detail at position (c). The horizontal arrow shows the direction of scan for mask convolution; a mask is shown in four consecutive positions underneath.

All of the above formulæ extend easily to multidimensional cases. Eg. in two dimensions:

$$f'(m,n) = \sum_{j=0}^{L} \sum_{i=0}^{K} f(m-i, n-j)(h(i,j) - h_{mean})$$

where L is the size of the spectrum in the second dimension. An example of how a 1D mask is employed in a real spectrum is shown in Figure 3.5.

Each region of interest has its own mask associated with it, and these masks may be different from region to region. This will be necessary, for instance, in cases where several different types of spectrum are being used simultaneously. The appropriate mask is convolved with each region of interest, generating a *convolved region of interest.* Convolved regions of interest are given arbitrary numbers, and are themselves stored in an array:

$$\underline{f'} = [f'_1, f'_2, ...f'_P]$$

where P is the total number of regions of interest. $\underline{f}'$ will usually undergo further processing, before being used as the input for the next step in the program's operation.

Further Processing of Convolved Regions of Interest

Once a convolved region of interest has been generated, it may be further processed to make the job of the pattern search algorithm easier. Non-linear functions may be applied to the whole convolved region of interest, certain portions of it may be deleted or scaled, and diagonal suppression may be applied to it.

Globally Applied Non-Linear Functions

In many cases, the default option of using raw mask responses, taken directly from the convolution process, produces unwanted effects. A small set of post-processing functions is available, which perform operations such as:

- Setting positive mask responses to 1 and all others to 0;

  This can be used where diagonal peaks are used in a pattern, but where their responses are ignored, because they tend to make a disproportionately large contribution to the overall score for the result. In this case, the post-processing function is combined with an absolute lower threshold of 1, which means that zero or negative parts of the diagonal will be ignored. Thresholding is discussed in Section 3.1.4 (page 30).

- Setting all positive mask responses to 0;

  This is most often used in forbidden peak penalisation (see Section 3.1.5, page 43), where the region of interest is convolved by a negative mask, producing negative values for unwanted peaks. It is undesirable that the program react to *dips* in the spectrum, since these would produce positive mask responses; this function can be used to inhibit this behaviour.

- Setting all mask responses above a given level to be equal to that level;

  This is used to prevent very large peaks from making a disproportionate contribution to the total score of a result.

- Scaling all mask responses above a given threshold by a logarithmic function.

  This function retains the essential peak shape, but reduces the responses to large peaks very significantly. This is useful because, rather than generating a flat-topped peak, as would be the case for the previous function, it produces a logarithmically scaled cap, which preserves the maximum point of the original peak.

Excluded or Scaled Zones

Spectra often contain linear (in 2D) or slice-shaped (in 3D) artifacts. The most common of these is the water line, but t_1/t_2 noise also sometimes arises. Features are available for deleting these completely (excluded zones) or for scaling them (scaled zones). In both cases, the position and width of these zones must be ascertained by the user, by a manual assessment of the spectra.

Diagonal Suppression

In spectra which have two or more axes upon which common spins occur, such as an H-H TOCSY, the diagonal peaks are usually very strong. This makes them attractive targets for the assignment code, and leads to the favouring of degenerate assignments, in which two or more spins are assigned to the same chemical shift. Suppression may be applied to each convolved region of interest, if the diagonal passes through it. Suppression is a maximum at the diagonal itself, becoming less severe with increasing distance from the diagonal. A choice of Gaussian or Lorentzian suppression profile is available.

3.1.4 Pattern Search

The program searches for patterns of peaks in the convolved regions of interest in a systematic way. It does this by incrementing the chemical shift values of each spin one data point at a time, and for each increment, examining the mask responses in each region of interest. This is illustrated schematically in Figure 3.6. A "snapshot" of this process, taken from a threonine search in a 2D TOCSY spectrum is shown in Figure 3.7.

At all points where chemical shift coordinates intersect within regions of interest, the mask response values are taken, and added together. The resulting value is called a *score*. Each combination of chemical shifts obtained by these means represents a possible assignment; the calculated score value gives these "assignments" a quantitative figure of merit, based on the combined peak plausibility (mask response) values:

$$R = \sum_{p=1}^{P} \sigma(p, \underline{\delta})$$

where R is the score value, p the current region of interest number, P the total number of regions of interest, $\sigma(p, \underline{\delta})$ the function which retrieves the mask response value from the convolved regions of interest for the current set of spin chemical shift values, and

$$\underline{\delta} = [\delta_1, \delta_2, \ \dots \ \delta_N]$$

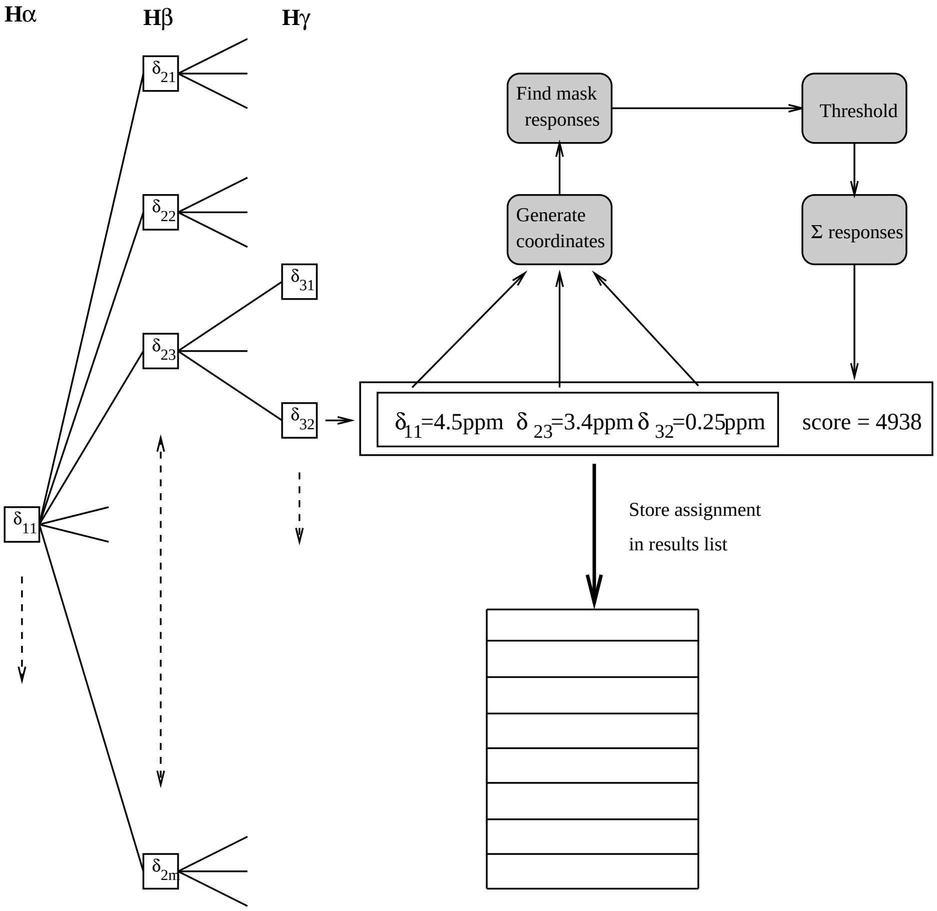


Figure 3.6: **Pattern searching.**
Exhaustive generation of chemical shift combinations is shown here as a tree. Only one $H\alpha$ value is illustrated; of course, there will actually be many $H\alpha$ values, one for each data point increment along the $H\alpha$ chemical shift range. Similarly for the $H\beta$'s and the $H\gamma$'s. Each combination of chemical shifts defines sets of coordinates in the regions of interest. The mask response values for each putative peak (taken from the convolved regions of interest) are summed, to give a score for this putative assignment. The assignment is stored in a list, ordered by score value.

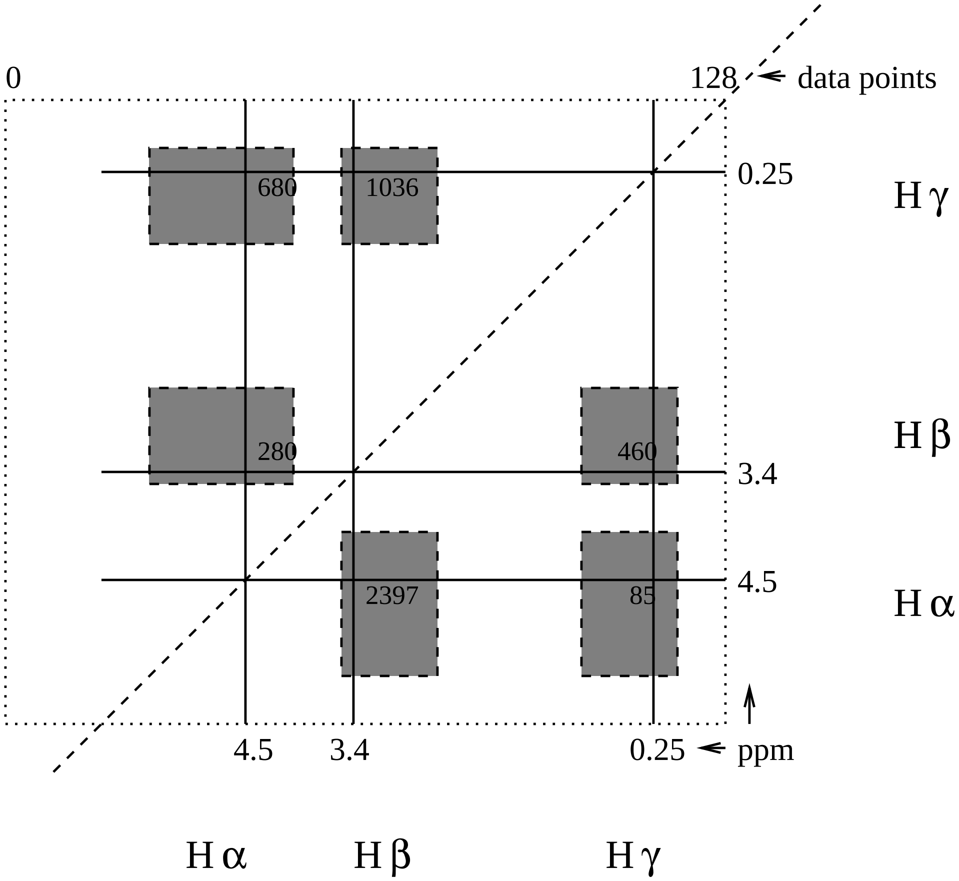


Figure 3.7: **Searching for threonine peaks in a 2D TOCSY spectrum.** This figure represents a "snapshot" taken from the search. The chemical shift values for each of the spins at this point in the search are shown in ppm. Internally, the program uses data point values; there will only be a finite number of possible chemical shift positions for each spin, due to the discrete nature of the spectrum. Where the chemical shift lines intersect within regions of interest, mask response values can be extracted from the convolved regions of interest. These values will be summed to give the overall score.

where N is the number of spins in the current pattern. $\sigma(p, \underline{\delta})$ uses p to select a particular convolved region of interest, and $\underline{\delta}$ to derive the chemical shift coordinates.

A region of interest has only as many dimensions as the spectrum that it was derived from. Hence, it follows that only a subset of the chemical shift values contained in $\underline{\delta}$ are needed to define the coordinates of a point within a region of interest. In order to extract these coordinate values, $\underline{\delta}'$, from $\underline{\delta}$, each region of interest has a *coordinate mapping matrix* ($\mathbf{M}$) associated with it. This is an $N * A$ matrix, where A is the dimensionality of a spectrum. It contains only 1's and 0's, and is multiplied with $\underline{\delta}$:

$$\underline{\delta}' = \mathbf{M}\underline{\delta}$$

For instance, if a region of interest was derived from a 3D spectrum, and corresponds to the first three spins in a list of six, the coordinate mapping matrix would work as follows:

$$\underline{\delta}' = [\delta_1\ \delta_2\ \delta_3] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \delta_1 & \delta_2 & \delta_3 & \delta_4 & \delta_5 & \delta_6 \end{bmatrix}$$

Hence, the function σ, used in calculating the mask response R in the expression above, can be defined as:

$$\sigma(p, \underline{\delta}) = f'_p(\underline{\delta}')$$

where f'_p is the zero-corrected mask convolution function for region of interest number p.

Avoiding a Combinatorial Explosion

The pattern finding algorithm just described could potentially lead to a combinatorial explosion. The number of chemical shift combinations goes up approximately to the power of the number of distinguishable spins, which for a spin system such as lysine, is large.

By prudent use of thresholding, however, the branches of the search "tree" can be severely pruned, with a corresponding increase in execution speed. An assignment which fails the threshold tests will not be retained. Two forms of thresholding are available: absolute and relative. Absolute thresholding will accept a mask response, r_p, which lies within given limits:

$$r_{pl} \leq r_p \leq r_{pu}$$

where r_{pl} is the lower threshold for region of interest p and r_{pu} is the upper threshold for region of interest p.

Relative thresholding tells the program to compare mask responses between two points in two convolved regions of interest; if the values are not similar enough, the assignment will be rejected. For two regions of interest, p and q, this can be expressed as:

$$t_l r_q \leq r_p \leq t_u r_q$$

where t_l and t_u are factors by which mask response r_q is multiplied to derive threshold values for comparison with r_p.

Search Levels

The peak definitions in a spin system pattern tell the search program which peaks to expect in the spectra. Each peak definition comprises of a set of spins contributing to that peak. For a 3D spectrum, such a definition will look something like this:

cross_peak=[[Gly,H_alpha_prime],[Gly,H_alpha],[Gly,C_alpha]];

As each definition is read in, it is added to a *peak list*. There is a one-to-one correspondence between the peak definitions in this list and the convolved regions of interest discussed in section 3.1.4 (page 27). For small spin systems such as glycine and alanine, it is quite acceptable to deal with all peaks in this list simultaneously. But, as the number of peaks grows, so does the memory consumption, because the convolved regions of interest for each peak need to be kept in memory. These regions take up typically 0.25 to 2 Mbytes. For, say, a leucine, one might be looking for 32 peaks, which could take up to 64 Mbytes. This will lead to swapping, and hence to a significant degradation in performance on many machines.

Furthermore, looking at *all* chemical shifts is wasteful. In the leucine case, for example, when one looks at the Cα planes of a HCCH-COSY spectrum, there are a limited number of "lines" of peaks that can be found corresponding to Hα chemical shifts, and hence, a limited number of sensible chemical shifts for Cα, Hα and Hβ spins. When looking into the Cβ planes of the spectrum, it would be sensible to limit the search to the Hα and Hβ chemical shifts already found, and only search over the full range for the Cβ chemical shifts.

How Search Levels Work

For the reasons mentioned above, the concept of the *search level* was introduced into the program. This strategy breaks up the search into a number of steps or levels. Within a given search level, only a limited number of peaks are searched for. Which peaks are searched for at which search level is determined by the pattern file.

The peak pattern search at a given search level will produce an intermediate list of results. Each result in this list will be used to generate a set of ranges, one for each spin in the spin system. The range for a given spin is calculated by taking the corresponding chemical shift value from the result, and then computing a narrow "error band" around that value. If the spin does not yet have an assigned chemical shift value, the full chemical shift range for that spin type will be used. These ranges constrain the search at the next search level. Thus, at search level L+1, the program can use chemical shift values found at search levels 1, 2, ... L as constraints, instead of using the full chemical shift ranges. This is illustrated in Figure 3.8.

Some Problems With the Search Level Concept

Each search level can be regarded as acting as a "filter" for the results from the level preceding it. As long as the spectra are "perfect", in the sense that all expected peaks are present and reasonably large, nothing will go wrong. However, if the spectrum contains weak or missing peaks, all partial assignments which require these peaks will be lost, even though peaks present in later search levels may be sufficient to fully define all chemical shifts for that spin system.

To prevent this potential loss of information, two options are available for "relaxing" the severity of the search:

- **Don't recycle results.** At the end of a search level, after the limiting ranges have been generated, the results list is normally "recycled" - ie. emptied - to make space for the results generated at the next search level. If many peaks in the next search level are weak or even missing, then it is probable that some of the "good" assignments found at the current search level will be filtered out, because the program will be unable to find the corresponding peaks. However, the user has the option to say that the results list will not be recycled, but passed on. In this case, the good assignments will survive, though some spins may not have any chemical shift values assigned to them. The difference between recycling and non-recycling search levels is illustrated in Figure 3.9.

- **Automatically adjust thresholds.** Alternatively, if the desired peaks are present at the current search level, but much weaker than expected, it is possible to make the program automatically adjust the threshold value, so that they can be found. The program is supplied with minimum and maximum acceptable results counts, via the pattern file. It will count the results generated, and if there are too few, reduce the thresholds. This process is iterative - if there are still too few after the first adjustment, it will be adjusted down even lower. If there are too many results, the thresholds will be increased.

  Program behaviour is different, depending on whether the results count is too large or too small:

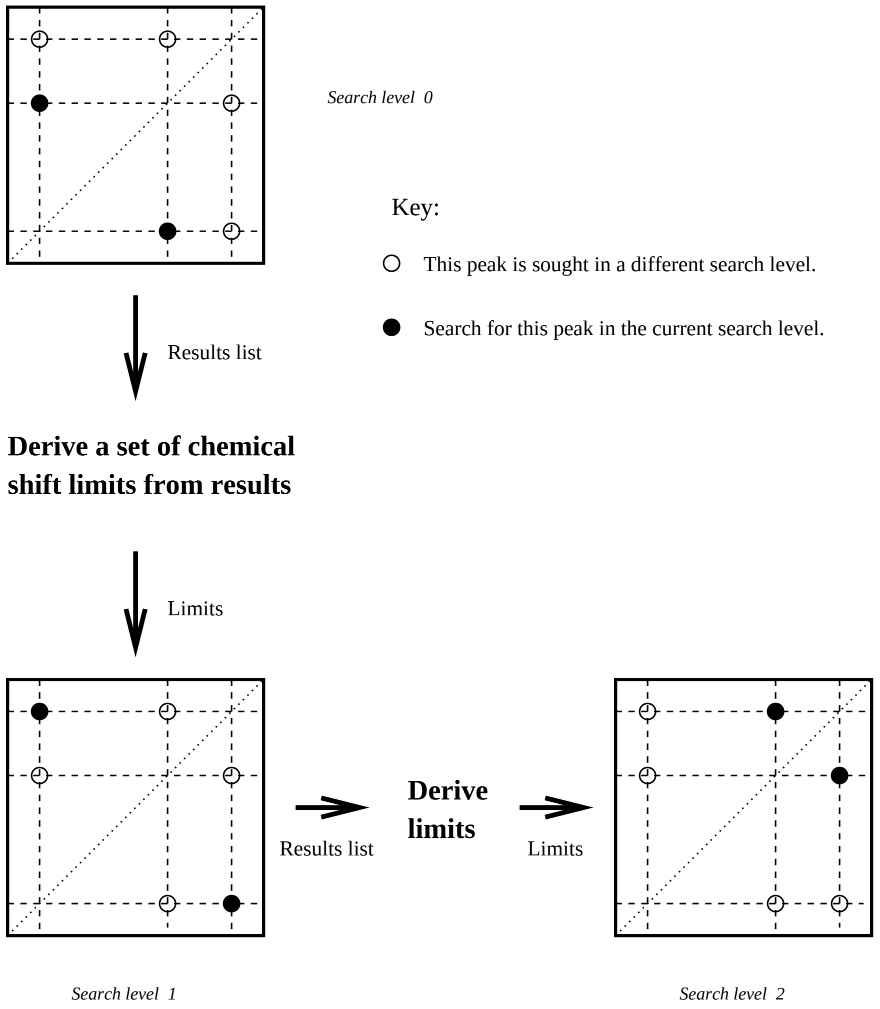

Search level 0
Key:
○ This peak is sought in a different search level.
● Search for this peak in the current search level.
Results list
Derive a set of chemical shift limits from results
Limits
Results list
Derive limits
Limits
Search level 1
Search level 2

Figure 3.8: **Search levels.**

The search pattern for an AMX spin system in a 2D COSY spectrum is shown. In the top panel, the situation at search level 0 is indicated: two peaks are searched for, the Hβ/Hα and the Hα/Hβ peaks. The results produced at this search level - there may be several hundred - are then converted into a set of ranges. In search level 1, searching for the Hβ'/Hα and the Hα/Hβ' peaks will be restricted within these ranges. In this case, this will mean that small search bands around the Hα chemical shifts found at search level 0 only will be searched. By the time the program reaches search level 2, all three chemical shifts, viz. Hα, Hβ and Hβ', will be known, and the restricted search for the Hβ'/Hβ and Hβ/Hβ' peaks will simply act as a filter to remove poor assignments from the previous two search levels.

i) If it is *too large*, the program generates mask response histograms for each of the peaks active at the current search level, using the results list as the data source. The program checks the results count continuously during the search and can apply this algorithm as soon as it detects that there are too many results. It must therefore make an estimate of the number of results *expected* at the current stage of the search, making the assumption that a) at the end of the search, the maximum number of results will be found and b) the results count grows linearly with the number of chemical shift combinations tested. It checks through the histogram until it finds a mask response value that will give exactly the estimated results count, and uses this as the new threshold value. Figure 3.10 shows how this works. Assume that the frequency distribution of the mask responses can be represented as a function $\kappa(r)$ of the mask response, r, and that we would like to find a new lower threshold value, $r_{l(new)}$. The integral of this function would then be:

$$\int_0^{r_{l(new)}} \kappa(r)dr = E_{max} - E_{expected}$$

Where actual results count E_{max}, and expected results count $E_{expected}$, are known values. Since $\kappa(r)$ is a non-analytic function, there is no general solution for $r_{l(new)}$, and the program finds it by starting with a value of $r_{l(new)} = 0$, and incrementing until the above equation is satisfied.

ii) If the results count is *too small*, the program multiplies the thresholds for all peaks active at the current search level by a factor equal to the desired number of results divided by the actual number of results. This makes the assumption that the results count is linearly dependent on reciprocal of the threshold value. This algorithm is only applied at the end of a search level, the program does not check if there are too few results during the search. Symbolically,

$$r_{l(new)} = r_{l(orig)} E/E_{min}$$

where $r_{l(orig)}$ is the original lower threshold value, $r_{l(new)}$ is the adjusted (new) lower threshold value, E_{min} the minimum allowed results count, and E the actual results count.

These two algorithms effectively put a feedback loop into the system, and turn it into what is known as a "bang-bang" servomechanism in control theory terminology [1]. In order to prevent uncontrollable oscillations between minimum and maximum results counts, a number of limiting and damping parameters are employed.

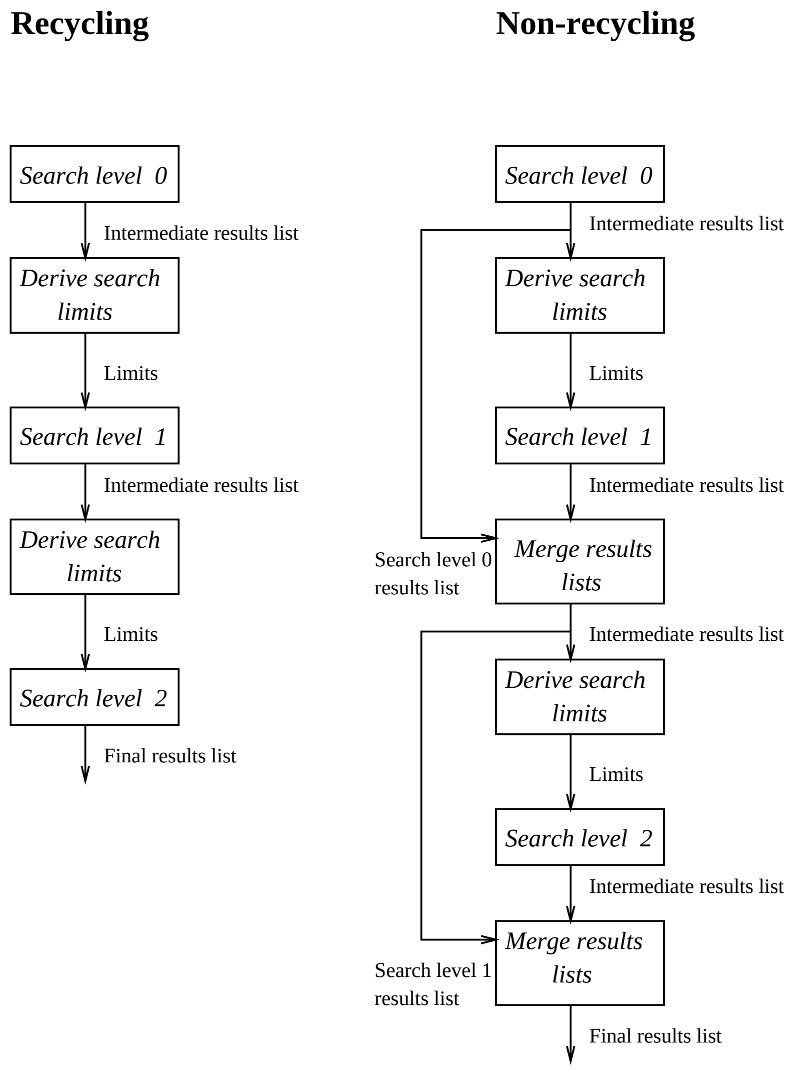


Figure 3.9: **Recycling and non-recycling search levels.**
The two flow diagrams show the difference between the situation where results lists are recycled at the end of each search level, and the situation where they are not. In the latter case, the results generated at search level N are retained, and are merged with the results generated at search level N+1, *before* limits are generated for search level N+2. This means that the information generated at search level N will still be available at search level N+2, even if the peaks searched for at search level N+1 were weak or nonexistent.

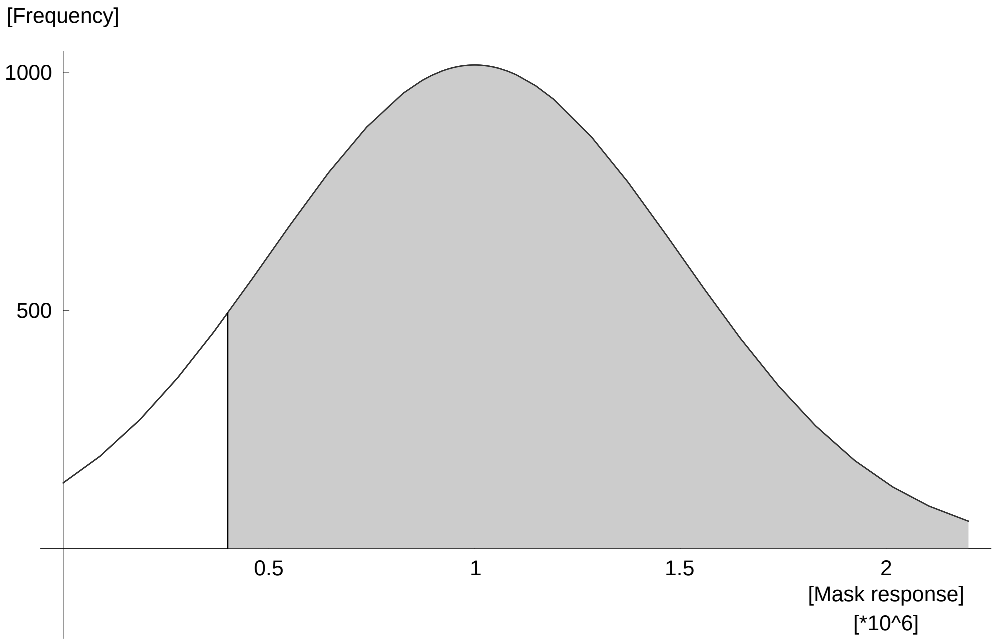


Figure 3.10: **Using a mask response histogram to choose a new threshold.** This illustrates how the threshold for a *single* peak is adjusted. The graph shows a (hypothetical) distribution of mask response values as obtained from the current list of results. It is assumed that the results list hit its upper limit during the search at a given search level. The total area under the curve represents the limiting results count. The area of the shaded part of the graph represents the number of results that would be *expected* at the point in the search where the limit was hit. The position of the line at the left-hand side of this shaded part determines the lower threshold that would be needed to achieve the desired results count. The underlying assumption of this method is that the shape of the mask response distribution will not change significantly during the search process.

Generating Results Lists

The assignments produced by the search algorithm will be stored sequentially in a *results list*, ordered according to score. To avoid storing results with only marginally different chemical shift values, the incoming assignment is checked against existing ones, to see if there are any from the same spin system, but with a higher score. If not, the new assignment is stored, otherwise it is recycled. This can be represented symbolically thus:

$$k_f = T(e, f, N, \Delta) \wedge (R_f \leq R_e)$$

Where k_f is a logical function returning the value **true** if result number e is to be clustered with the new result (f), **false** otherwise. $T(e, f, N, \Delta)$ is also a logical function which determines if all N chemical shift values in result f are within the chemical shift difference Δ of the corresponding chemical shift values in result e:

$$T(e, f, N, \Delta) = (\sum_{n=1}^{N} t(e, f, n, \Delta) > N)$$

The arithmetic function $t(e, f, n, \Delta)$ determines whether the chemical shifts for spin n in results e and f are less than chemical shift difference Δ apart:

$$t(e, f, n, \Delta) = \begin{cases} 1 & \text{if } |\delta_{en} - \delta_{fn}| \leq \Delta \\ 0 & \text{otherwise} \end{cases}$$

The results list forms the input for the next step in the program.

3.1.5 Result List Filtering

Typically, the pattern search process will generate between 50 and 2000 results, depending on the pattern being searched for and the exact search control parameters being used. Although the correct results would be expected at the high-scoring end of the results list, manual verification of the full list would be a daunting task. Hence, a suite of heuristic results list filtering algorithms has been written, to "clean up" the list, and to make sure that the correct results are more likely to be the highest scoring ones. The algorithms embody criteria similar to those used during manual assignment of spin systems. There are two classes of filtering algorithm: *penalising* and *deleting*. The penalising type multiplies the score for a given result by a value which lies between zero and one, called a penalisation factor (F). The penalisation factor is a measure of how "sensible" a result looks according to a particular criterion. The deletion type removes results which do not look "sensible". The deletion criterion k is a logical value, **true** if the current result is to be deleted, **false** if not. The parameters required by the filtering algorithms may be supplied to the program via the pattern file.

The following penalisation algorithms are available (see Figures 3.11 and 3.12 for a graphical summary):

- **Too-far.** Penalise in cases where two chemical shifts are more than a given distance apart. This is usually applied to the protons in methylene groups, since their chemical shifts are rarely more than 1 ppm apart. Eg. if the pattern file defines the search ranges for $H\beta/H\beta'$ in an AMX spin system as being from 0.5 ppm to 3.0 ppm, the program might find candidate patterns with $H\beta$=0.53 ppm and $H\beta'$=2.94 ppm. Such wide separations of chemical shift values seldom arise in real spectra, and should therefore be penalised. Symbolically,

$$F = \begin{cases} F_s & \text{if } |\delta_1 - \delta_2| > \Delta_u \\ 1 & \text{otherwise} \end{cases}$$

  where Δ_u is an upper limit on chemical shift difference, and F_s is a constant penalisation factor, which will be applied if the penalisation criterion holds.

- **Std-dev.** Find the standard deviation of the mask response values for the current result. The larger this value, the more severe will be the penalisation, thus favouring results where all mask responses are similar. For result number e:

$$F_e = 1 - \frac{S(\underline{r}_e)}{S_{max}}$$

  where S_{max} is the maximum standard deviation of the mask responses over all results, E,

$$S_{max} = \max_{e=1}^{E} S(\underline{r}_e)$$

  $S(\underline{r}_e)$ is the standard deviation of the mask responses for result e:

$$S(\underline{r}_e) = \sqrt{\sum_{p=1}^{P} (r_m - r_p)^2}$$

  r_p is the mask response for region of interest number p, and r_m is the mean mask response for the current result:

$$r_m = \frac{\sum_{p=1}^{P} r_p}{P}$$

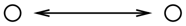
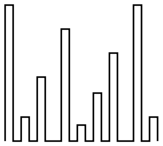
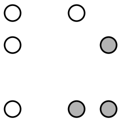
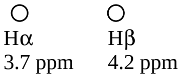
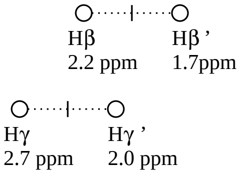
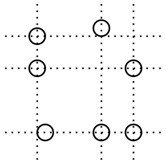
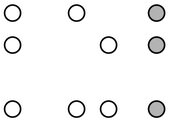
Name	Situation (graphical representation)	Situation (in words)	Action
Too-far		Chemical shifts too far apart.	Penalise if chemical shift separation greater than a given threshold.
Std-dev		Mask responses show a large diversity	Penalise according to the extent of the diversity.
Thresh-count		Some mask responses responses are below threshold.	Penalise according to the number of below-threshold mask respoonses.
Nucleus-order		First spin should have greater chemical shift than second.	Apply a standard penalisation factor.
Quad-order	 Hβ 2.2 ppm Hβ ' 1.7ppm Hγ 2.7 ppm Hγ ' 2.0 ppm	Mean chemical shift of first pair of spins should be greater than mean chemical shift of second pair of spins.	Apply a standard penalisation factor.
Peaks		Some chemical shift line intersections are not at the centres of peaks in the spectrum.	Penalise according to the number of non-centred peaks.
Excess-peaks		More peaks are found than would be expected for the current pattern.	Penalise according to the number of excess peaks.

Figure 3.11: **Graphical summary of penalisation algorithms**

(see Figure 3.12 for the remaining algorithms).

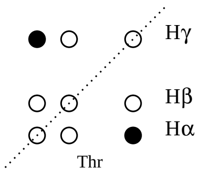
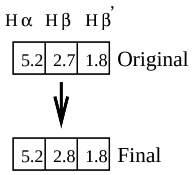
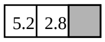
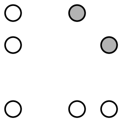
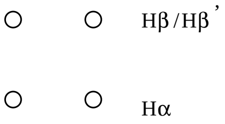
Name	Situation (graphical representation)	Situation (in words)	Action
Forbidden-peaks		Peaks exist at the intersections of chemical shift lines which shouldn't be there, eg. H α/Hγ peaks in a COSY.	Penalise according to the number of these "forbidden" peaks.
Chem-shift-migr		In a multi-search level pattern, the chemical shift for a given spin may change from one search level to the next.	Penalise according to the number of spins which have "migrated" in this way, and according to the extent of the migration.
Uninst-chem-shft		The program may fail to find one or more chemical shifts.	Penalise according to the number of missing chemical shifts.
Uninst-resp		The program may fail to find one or more peaks.	Penalise according to the number of missing peaks.
Deg-peaks-pen		Peaks belonging to different spins may overlap.	Penalise according to the number of overlapping peaks and the degree of overlap.

Figure 3.12: **Graphical summary of penalisation algorithms**

(continued from Figure 3.11).

- **Thresh-count.** Apply penalisation if any mask responses are below absolute threshold values. This feature allows results to be accepted even if one or more of the mask response values is below threshold, a situation which can arise, for instance, if the program performs dynamic threshold adjustment during search. The degree of penalisation depends on the number of sub-threshold mask responses. Symbolically,

  $$F = 1 - \frac{P_T}{P}$$

  where P_T is the number of under-threshold mask responses,

  $$P_T = \sum_{p=1}^{P} t(r_p)$$

  $t(r_p)$ is the thresholding function:

  $$t(r_p) = \begin{cases} 1 & \text{if } r < r_u \text{ and } r > r_l \\ 0 & \text{otherwise} \end{cases}$$

  r_u and r_l are the upper and lower mask response thresholds respectively.

- **Nucleus-order.** Penalise results which contain badly ordered chemical shifts. For instance, in the case of serine, the ranges for $H\alpha$ and $H\beta$ are heavily overlapping, but the $H\beta$ chemical shifts tend to be lower than the $H\alpha$ chemical shifts. Hence, if a spin system were found in which the $H\beta$ chemical shift is *greater* than the $H\alpha$ chemical shift, penalisation could be applied. Symbolically,

  $$F = \begin{cases} F_s & \text{if } \delta_j > \delta_i \\ 1 & \text{otherwise} \end{cases}$$

  where δ_j and δ_i are the chemical shifts of spins j and i respectively, and where δ_j would normally be expected to be larger than δ_i.

- **Quad-order** A mechanism similar to **Nucleus-order** exists for penalising two badly ordered *pairs* of spin chemical shifts. Consider, for instance, the typical chemical shift ranges for the protons in the lysine side-chain spin system. The ranges for the Hβ/Hβ' and the Hγ/Hγ' spins show a lot of overlap, and it is difficult to sensibly specify an ordering for single chemical shifts. But *in general*, the *mean* chemical shift for the Hβ/Hβ' pair is larger than the *mean* chemical shift for the Hγ/Hγ' pair. This leads fairly naturally to the following penalisation scheme:

$$F = \begin{cases} F_s & \text{if } \delta_j + \delta_i > \delta_k + \delta_l \\ 1 & \text{otherwise} \end{cases}$$

where δ_j and δ_i are the chemical shifts of one of the pairs of spins, and δ_k and δ_l are the chemical shifts of the other pair of spins, and where the mean chemical shift for the δ_j/δ_i pair would normally be expected to be larger than the mean chemical shift for the δ_k/δ_l pair.

- **Peaks.** Penalise results whose chemical shift values don't lie exactly at the maximum points of spectrum peaks. This will tend to promote "perfect" results, those where all peaks are distinct and precisely aligned. This feature is mainly useful in spectra with a high degree of dispersion. Symbolically,

$$F = \frac{P_o}{P}$$

where P_o is the number of mask responses for the current result which lie on maxima in the spectrum:

$$P_o = \sum_{p=1}^{P} o(p)$$

$o(p)$ is a function which decides if mask response number p lies exactly on a maximum. Assume that a list of T coordinates for the maxima in the spectrum has been generated and $\underline{\gamma}_t$ representing coordinate number t in this list, and that associated with each peak p is a peak coordinate $\underline{\delta}_p$. Then one can define

$$o(p) = \begin{cases} 1 & \text{if } \prod_{t=1}^{T} q(\underline{\delta}_p, \underline{\gamma}_t) = 0 \\ 0 & \text{otherwise} \end{cases}$$

where the function $q(\underline{\delta}_p, \underline{\gamma}_t)$ determines if coordinate sets $\underline{\delta}$ and $\underline{\gamma}$ are identical:

$$q(\underline{\delta}, \underline{\gamma}) = \prod_{a=1}^{A} z(\delta_a, \gamma_a)$$

where A is the total number of dimensions in the spectrum under consideration, and we define $\underline{\delta} = [\delta_1, \delta_2, \ \ldots\ \delta_a, \ \ldots]$ and $\underline{\gamma} = [\gamma_1, \gamma_2, \ \ldots\ \gamma_a, \ \ldots]$, with $z(\delta, \gamma)$ being the comparison function for two individual chemical shift coordinates:

$$z(\delta, \gamma) = \begin{cases} 1 & \text{if } \delta = \gamma \\ 0 & \text{otherwise} \end{cases}$$

- **Excess-peaks.** This function penalises results where, at a given chemical shift, there are more peaks in the spectrum than would be expected for the spin system currently being searched for. For example, in the case of glycine, the peak pattern which being searched for in a 2D-COSY or -TOCSY spectrum may also match peaks from AMX spin systems. Hence, it makes sense to look within a given range of ppm values for other large peaks which might indicate participation in a larger spin system, and penalises according to the number of "excess" peaks found. Symbolically,

  $$F = \begin{cases} \frac{P_{expected}}{P_{found}} & \text{if } P_{expected} < P_{found} \\ 1 & \text{otherwise} \end{cases}$$

  where $P_{expected}$ is the expected number of mask responses along coordinate corresponding to the chemical shift of one of the spins, and P_{found} is the total number of peaks actually found on this line,

  $$P_{found} = \sum_{\delta=\delta_l}^{\delta_u} \rho(\delta)$$

  where δ_l and δ_u are lower and upper limits on the range of chemical shift values that will be examined, and $\rho(\delta)$ is a thresholding function for chemical shift position δ on the line,

  $$\rho(\delta) = \begin{cases} 1 & \text{if } r_\delta > r_l \\ 0 & \text{otherwise} \end{cases}$$

  r_δ is the mask response at the current chemical shift position, r_l is the lower threshold value.

- **Forbidden-peaks.** The pattern definitions will normally be set up to search for peaks in the spectra. But there will also be cases where one wishes to search for peaks which should *not* be there, so-called "forbidden" peaks. For instance, peaks which may be present in a TOCSY spectrum, but should not be present in a COSY spectrum. This is achieved by using mask specifications with negative scale values. If a forbidden peak is found, it will therefore produce a *negative* response. These peaks will hence make a negative contribution to the overall score for a result, a desirable feature. However this may not be enough to prevent the result from scoring highly, if the other peaks are strong. A special penalisation algorithm has therefore been written to overcome this - each forbidden peak found causes the result's score to be multiplied by a fractional penalisation factor. Symbolically,

$$F = 1 - \frac{P_F}{P}$$

where P_F is the number of forbidden peaks:

$$P_F = \sum_{p=1}^{P} \eta(r_p)$$

$\eta(r_p)$ is a function which returns 1 if a peak p is forbidden, 0 otherwise:

$$\eta(r_p) = \begin{cases} 1 & \text{if } r < r_l \text{ and } r < 0 \\ 0 & \text{otherwise} \end{cases}$$

r_l is a lower mask response threshold value, this will be a negative number, and it ensures that the peak is big enough to be worth penalising.

- **Chem-shift-migr.** One of the features of the pattern search algorithm is that it allows a certain error margin around peaks, so that peaks with slight displacements in different spectra will still match with each other. This is generally a useful feature, but it can lead to undesirable side-effects. For instance, the skirt of a peak in one spectrum may be incorrectly matched with the top of a peak in another spectrum. This can happen if the lower mask response threshold is small, for instance. In order to minimise the effects of such errors, a penalisation algorithm is available. It looks at the extent to which chemical shift values have "migrated" from the originally chosen chemical shift during the pattern search procedure. This migration is caused by the program trying to find the best possible compromise chemical shifts to fit many peaks in all spectra, over multiple search levels. The program stores the initial chemical shift values found for each spin, and at the end of the search, computes the differences between the initial and final chemical shift values. The larger the mean value of this difference, the more the result will be penalised. Symbolically,

$$F = 1 - \frac{\sum_{n=1}^{N} \Delta_n}{N\Delta_t}$$

where N is the total number of spins in the spin system, Δ_t is a threshold chemical shift difference value for the current spectrum axis, and Δ_n the chemical shift migration for spin n:

$$\Delta_n = abs(\delta_{n(initial)} - \delta_{n(final)})$$

Where $\delta_{n(initial)}$ and $\delta_{n(final)}$ are the initial and final chemical shift values for spin n respectively.

- **Uninst-resp.** In order to allow for missing peaks or even missing groups of peaks, search levels may be "skipped" by the program. This will lead to some mask responses having no value - they are uninstantiated. Results with many uninstantiated mask responses are considered to be "poor", even though they may score highly, and occupy top positions in the results list. They can be penalised with the following algorithm:

$$F = 1 - \frac{P_q}{P}$$

where P_q is the count of uninstantiated mask response values,

$$P_q = \sum_{p=1}^{P} \phi(r_p)$$

r_p is mask response number p, and $\phi(r)$ determines if a chemical shift value is uninstantiated - by convention, $-\infty$ is used as a "flag" to indicate this:

$$\phi(r) = \begin{cases} 1 & \text{if } r = -\infty \\ 0 & \text{otherwise} \end{cases}$$

- **Uninst-chm-shft.** In the extreme case, skipping search levels in the manner described for **Uninst-resp** can result in chemical shift values not being found either. Results with many missing (or uninstantiated) chemical shifts are undesirable, and this algorithm can be used to penalise them. Symbolically,

$$F = 1 - \frac{N_q}{N}$$

where N_q is the number of uninstantiated chemical shift values,

$$N_q = \sum_{n=1}^{N} \phi(\delta_n)$$

δ_n is the chemical shift for spin number n in the current result, and $\phi(\delta)$ determines if a chemical shift value is uninstantiated - by convention, $-\infty$ is used as a "flag" to indicate this:

$$\phi(\delta) = \begin{cases} 1 & \text{if } \delta = -\infty \\ 0 & \text{otherwise} \end{cases}$$

- **Deg-pks-pen.** Geminal protons may sometimes have similar or identical chemical shifts (degeneracy). Hence, it is undesirable to delete results containing degeneracy, since this can remove valid results. However, due to the way the search algorithm is implemented, the program has a strong tendency to find degenerate chemical shifts, even in cases where it might be obvious to an observer examining the spectra manually that degeneracy is not present. Degenerate peaks penalisation can be used to counteract this tendency. It counts the number of degenerate mask responses positions for a given result. The bigger this number, the more vigorous the penalisation. Symbolically,

$$F = 1 - \frac{P_d}{P}$$

where P_d is the number of degenerate mask responses,

$$P_d = \sum_{p=1}^{P} \zeta(p)$$

$\zeta(p)$ determines if mask response number p overlaps any other mask response, q:

$$\zeta(p) = \begin{cases} 1 & \text{if } \sum_{q=1}^{P} \prod_{a=1}^{A} \mu(\delta_{ap}, \delta_{aq}) > 0 \\ 0 & \text{otherwise} \end{cases}$$

where A is the number of dimensions in the current spectrum, and the function $\mu(\delta_{ap}, \delta_{aq})$ determines if two chemical shift values on dimension a of the spectrum are close to each other:

$$\mu(\delta_{ap}, \delta_{aq}) = \begin{cases} 1 & \text{if } |\delta_{ap} - \delta_{aq}| > \Delta_t \\ 0 & \text{otherwise} \end{cases}$$

Δ_t is a threshold chemical shift difference value for the current spectrum axis, in general, equal to half the expected peak size (eg. on a ^{13}C axis, $\Delta_t \approx 0.6$ppm).

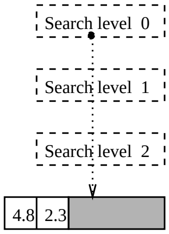
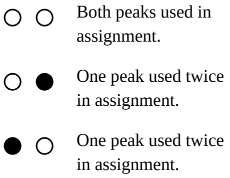
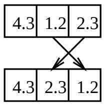
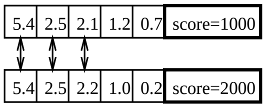
Name	Situation (graphical representation)	Situation (in words)	Action
Sl-diff	 Search level 0 Search level 1 Search level 2 4.8 2.3	If there are many non-recycling search levels in the current pattern, then some results will survive over many search levels.	Delete results which have survived for more than a given number of search levels.
Ovrlp-clus	 Both peaks used in assignment. One peak used twice in assignment. One peak used twice in assignment.	There are situations where a pair of peaks might be the most obvious assignments for a spin system, but the program also finds assignments for each of the single peaks too.	Retain the result whose chemical shifts for the peaks are different, and delete the results with degenerate chemical shifts for these peaks.
Exs-penal	$H\beta = H\beta' = 2.8$ $H\gamma = H\gamma' = 1.2$ $H\delta = H\delta' = -0.3$	Many pairs of spins have degenerate chemical shifts.	Delete.
Sym		Two results exist, which differ only in that the chemical shifts for two spins are exchanged.	Delete the result with the lowest score.
Sec-clus	 5.4 2.5 2.1 1.2 0.7 score=1000 5.4 2.5 2.2 1.0 0.2 score=2000	Two or more results exist with many chemical shifts in common	Delete the lowest scoring results.

Figure 3.13: **Graphical summary of deletion algorithms**

(see Figure 3.14 for the remaining algorithms).

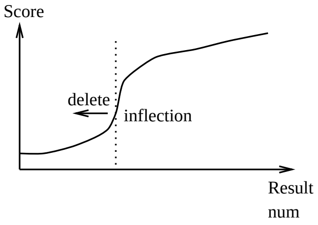
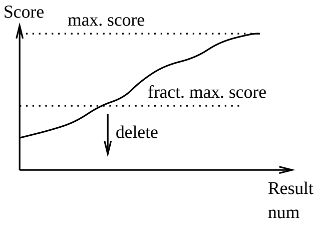
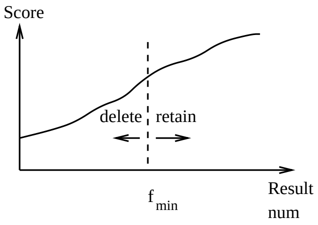
Name	Situation (graphical representation)	Situation (in words)	Action
Lo-score		The action of multiple penalisations on the score of the results in the results list can lead to a point of inflection in the score values.	Delete results after point of inflection.
Fract-score		Results whose scores are below a certain fraction of the score of the most highly scoring result are probably not very interesting.	Delete all results under the fraction of the maximum score.
Cut		A fixed number of the lowest scoring results are not interesting.	Delete them.

Figure 3.14: **Graphical summary of deletion algorithms**

(continued from Figure 3.13).

The following deleting algorithms are available (see Figures 3.13 and 3.14 for a graphical summary):

- **Sl-diff.** If results recycling is disabled, then results with many uninstantiated chemical shifts or responses can survive over many search levels. This is by no means desirable behaviour - it is usually better to stop results from being passed on from one search level to the next once they are more than, say, 4 search levels old. This is realised in the following deletion algorithm:

  $$k_f = \quad s_f \leq (s - \Delta s)$$

  where s_f is the search level at which result number f was generated, s is the current search level, and Δs is the maximum allowed difference between current search level and that of a result.

- **Ovrlp-clus.** This algorithm looks for triplets of results for which, in a given pair of spins, one has different chemical shifts and the other two identical ones. The result with the *different* chemical shifts is retained, whilst the two results with degenerate chemicals shifts are deleted. Eg.

#	Cα	Hα	Cβ	Hβ	Hγ	Hγ'	score
1	56.9	4.36	36.2	3.45	2.18	2.18	20813
2	56.9	4.36	36.2	3.45	1.56	2.18	17993
3	56.9	4.36	36.2	3.45	1.56	1.56	13329

  In this case, it is highly likely that result 2 is correct, but the other two are not. Hence, the program retains result 2 and discards the others. Symbolically,

  $$\begin{aligned} k_f = k_g \quad = \quad & (\delta_{fi} = \delta_{fj}) \wedge (\delta_{gi} = \delta_{gj}) \wedge \\ & (\delta_{fi} \neq \delta_{gi}) \wedge (\delta_{ei} = \delta_{fi}) \wedge \\ & (\delta_{ej} = \delta_{gj}) \wedge \Lambda(e, f, g, i, j) \end{aligned}$$

  where result numbers e, f and g are being considered, and comparing spin numbers i and j. δ_{qz} represents the chemical shift value for spin number z in result number q. $\Lambda(e, f, g, i, j)$ is the function which determines whether the remaining spins (those which are not i or j) have similar chemical shifts for all three results:

$$\Lambda(e,f,g,i,j) = \lambda(e,f,i,j) \wedge \lambda(e,g,i,j)$$

$\lambda(s,t,i,j)$ determines whether the chemical shifts for the *two* results s and t are similar, apart from the chemical shifts i and j:

$$\lambda(s,t,i,j) = (\prod_{n=1}^{N} s(s,t,i,j,n) > 0)$$

$s(s,t,i,j,n)$ determines whether the chemical shift for spin n are identical between results s and t:

$$s(s,t,i,j,n) = \begin{cases} 1 & \text{if } \mu(\delta_{sn},\delta_{tn}) \text{ and } n \neq i \text{ and } n \neq j \\ 0 & \text{otherwise} \end{cases}$$

δ_{zn} is the chemical shift for spin number n in result z. The function μ is defined above, under **Deg-pks-pen**.

- **Exs-penal.** A result containing more than two or three chemical shift degeneracies has a rather low probability of being correct. This algorithm removes such over-degenerate results. Symbolically,

$$k = (M > M_x)$$

where M_x is the *maximum* allowable degenerate pair count (equal to the number of pairs of spins with overlapping search ranges), and M is the count of the actual number of degenerate pairs:

$$M = \sum_{n=1}^{N}\sum_{q=1}^{N} \mu'(\delta_n,\delta_q)$$

$$\mu'(\delta_n,\delta_q) = \begin{cases} 1 & \text{if } \mu(\delta_n,\delta_q) \text{ and } n \neq q \\ 0 & \text{otherwise} \end{cases}$$

μ is the proximity function, defined above under **Deg-pks-pen**.

- **Sym.** If two spins have large overlapping chemical shift ranges (for instance, the Hβ/Hβ' spins), then the program can produce two results with identical (or very similar) score values, but with transposed chemical shifts between the two spins. One of these, the one with the lowest score, is deleted. The deletion condition for result f is given by:

$$\begin{aligned} k_f &= (\delta_{fi} = \delta_{ej}) \wedge (\delta_{fj} = \delta_{ei}) \wedge \\ &\quad (|R_f - R_e| < D) \wedge (R_f \leq R_e) \end{aligned}$$

where results with numbers e and f are compared, and i and j are the numbers of two spins with overlapping chemical shift ranges (such spins are found automatically). D is a maximum allowable difference in score value - results with very different scores are not compared.

- **Sec-clus.** The results lists produced by the pattern search algorithm tend to contain many results with chemical shift sets which differ only slightly from one another. In general, only a small number of these results will correspond to genuine assignments; the rest will be "near misses". Secondary clustering is designed to remove the "near misses" and retain the genuine assignments. It is called "secondary" clustering to distinguish it from the weaker form of clustering which takes place when a new result is inserted in the results list (see 3.1.4, page 37). Results are clustered according to degree of chemical shift overlap. Two "chemical shift difference bands" are provided; these specify the maximum allowable differences in chemical shifts between corresponding spins in two results. For any two given results to be clusterable, a given fraction of the differences in chemical shifts must lie within one band, whilst the rest must lie within the other band. If two results are considered clusterable, then the lowest scoring result will simply be deleted. Symbolically,

  $$k_f = T(e, f, N_l, \Delta_l) \wedge T(e, f, N_u, \Delta_u) \wedge (R_f \leq R_e)$$

  where $T(e, f, N_z, \Delta_z)$ is a logical function which determines if N_z or more spins in result f are within the chemical shift difference Δ_z of the corresponding spins in result e - this has already been defined in Section 3.1.4 (page 37). The remaining parameters are specified in the pattern file. The two chemical shift difference values, Δ_l and Δ_u, correspond to the lower and upper chemical shift difference bands respectively. The number of spins in each of these bands are N_l and N_u respectively. N_u is calculated as:

  $N_u = N - N_l$.

  where N is the total number of spins in the spin system.

- **Lo-score.** Results in a results list are ordered according to their score. As one descends the results list, the scores gradually decline. If there is a sudden decline in score, it indicates that multiple penalisations have cut in simultaneously, and it is likely that subsequent results will be of poor quality. Such results can be deleted using the condition:

$$k_f = (R_f < R_{low})$$

where f is the current result number, and R_{low} is the minimum acceptable score value. This value is obtained from the ordered list of scores for all results, $\underline{R}$ as follows. Starting with the highest scoring result, number P, set the current result number $p = P$. Now descend this list by decrementing p until the condition $R_p - R_{p-1} \geq D_{low}$ is satisfied. Then, we define $R_{low} = R_p$. D_{low} is the maximum allowable score difference value, and it is calculated from the highest scoring result thus:

$$D_{low} = R_P * \epsilon_{low}$$

The constant ϵ_{low} is a value between 0 and 1, defined in the pattern file.

- **Fract-score.** Delete results less than a given fraction of the maximum score. For result number f:

  $$k_f = (R_f < R_P * \epsilon_{low})$$

  where R_P is the score for the highest scoring result, and ϵ_{low} is obtained from the pattern file.

- **Cut.** Only accept a fixed number of the highest scoring results. For result number f:

  $$k_f = (f < E - f_{min})$$

  where f_{min} is the minimum results list size.

3.1.6 Pattern Design

As will be readily apparent from the preceding sections, there are many things to be taken into account when designing a pattern, not just the peaks that should be found, but also setting up convolutional masks, selecting threshold values, and setting results list processing parameters.

Masks are always given a default height of unity to start with. Mask sizes, in ppm in all dimensions of the spectra, are established by eye, as follows: i) Contouring of the spectra is set up such that obvious signal peaks are visible, but noise is not. ii) Ten peaks are selected at random from each spectrum, and their size at the lowest contour measured in each dimension. iii) Average size values are found for each dimension; these will then be used to establish the mask sizes for each spectrum.

Calibration peaks must then be selected, so that i) the mask responses from different spectra can be normalised to each other (by adjusting the mask scales away from unity), and ii) a noise threshold can be established. A number of program runs may need to be performed on these peaks to get things right.

Selecting penalisation factors for the various results list penalisation algorithms involves making a quantitative estimate of the importance of the various algorithms; the *more* important an algorithm is deemed to be in producing correctly ordered results, the *smaller* the penalisation factor.

Results list deletion algorithms serve two somewhat different purposes: i) to present a non-redundant list of results to the user at the end of a pattern search, and ii) to reduce the number of results passed from one search level to the next (this increases speed and reduces the probability that "good" results get lost due to results list overflow). In general, the fewer deletions that take place between search levels, the better, since deleting incomplete results is based on incomplete information and therefore prone to errors. The choice of how many deletion algorithms are applied to intermediate results sets depends on how many search levels a pattern has, and how low the thresholds have been set, since both of these factors have a strong bearing on the size of intermediate results lists.

Once these choices have been made, and a pattern designed and tested, one will often find that the program has not behaved quite in the way expected. This is most often due to the fact that there are many interactions between the various parts of the pattern searching process which cannot all be anticipated beforehand. Hence, some tuning of the various parameters is often necessary.

For a given set of spectra (eg. the HCCH-COSY plus HCCH-TOCSY pair presented in Section 3.5.1, page 72), this basic design process will need to be done only once, when the spectra are first encountered. The next time a similar set is encountered, one can safely assume that most of the parameters are correct, but one will normally still need to do some fine-tuning, eg. to make sure that suppression of peaks on the water line (which is not always in the same place in different spectra) is set up correctly.

In Section 3.4 (page 66), an example is given of how different design choices can affect program performance.

3.2 Introduction to the Experimental Work

This section summarises the data gathered during the use of *patt_recog* in an experimental context. The first three sections (3.2.1, 3.2.2 and 3.2.3) describe the administrative software for the assignment program suite, and explain the presentation of pattern files and results in this chapter. The following two sections (3.3 and 3.4) give a systematic program performance analysis and an illustration of how pattern design affects performance respectively. Subsequent sections (3.5 and 3.6) provide case studies, where the program was applied to a couple of proteins, one already assigned, the other unassigned.

3.2.1 The CATCH23 Suite

This is a suite of programs providing an environment for automated assignment. The following are the main programs in the suite:

catch23	Administrative interface to suite.
patt_recog	Performs spin system assignment.
chain	Performs sequential assignment.
daurelia	Displays spin system assignments graphically.

The program *catch23* is a shell-like administrative program, which coordinates the operation of the other programs; it greatly simplifies the user-interface, and also performs essential bookkeeping operations. It can be used for editing pattern files, starting spin system or sequential assignment runs and for displaying or printing results.

The program *patt_recog* performs spin system assignment and has already been described in detail in this chapter. *chain* performs sequential assignment; this forms the subject of chapter 4.

A modified version of the Bruker program AURELIA [48] called *daurelia* has been produced, with an additional top-level menu, allowing the viewing the results generated by CATCH23. This allows the user to view the putative assignments superimposed upon the original spectra.

It works by reading in the results file for a given pattern search run, and using the information in this file to select which spectra to display. With three-dimensional spectra, a single slice is displayed from each spectrum. A list of the results is also displayed, each result being shown as a set of chemical shifts and a score, and if the user clicks on one of them, the assigned chemical shifts will appear as lines plotted over the selected spectra. The user also is able to (metaphorically) step along the sidechain in a 3D heteronuclear spectrum, to see how the program has performed on different ^{13}C or ^{15}N planes. This makes it possible to assess whether or not the search strategy has worked successfully for each result. Results may be "marked" with a status "good", "bad" or "unknown"; these statuses are preserved in the results list after leaving "daurelia", to be used either for future reference or for automatic results list statistics calculations.

3.2.2 A Note on the Presentation of Patterns

The pattern files used to drive the pattern search process can be quite large - for instance, the pattern file for the experiments described in Section 3.5.1 (page

72) covers 80 sides of A4 when printed out. However, in this thesis, a compact standard representation for the information in a pattern file is used. First, the essential parameters describing the spectra are given - spectrum type and size in data points and ppm. This is followed by a list of the results list processing parameters, using the symbolic terminology developed in chapter 3. Then a list of the chemical shift ranges employed (but only up to $C\beta/H\beta$) is presented.

Finally, a graphical representation of the peaks searched for in the various spectra is given. This uses a grid of lines in which each line represents a spin and peaks that the program should find are shown as circles at the intersections between lines. The sign of the peak being searched for is indicated by a full circle for positive peaks and a broken circle for negative peaks. Negative peaks are important in CBCANH and CBCA(CO)NH spectra, but the program also uses them to identify "forbidden" peaks, such as (apparent) $H\alpha/H\gamma$ peaks in a COSY spectrum, which could indicate a faulty assignment. Numbers are given in square brackets adjacent to each peak, to show the search level at which the peak is searched for. The columns of carbon spin names on the right hand side of the figures indicate the carbon planes within which the corresponding row of peaks will be found.

Figure 3.15 (page 58)is a good representative example. The column of labels on the left and the row of labels underneath the figure indicate the proton spins to which the associated lines belong. The column of labels on the right hand side indicates the ^{13}C plane in which the peaks will appear. The numbers in square brackets are the search levels at which the peaks are searched for by *patt_recog* - this pattern contains three search levels, numbered 0, 1 and 2.

3.2.3 Criteria Used in the Assessment of Results Lists

In order to allow assessment and comparison of the results lists produced by the *patt_recog* program, three criteria will be used throughout this chapter:

- **Correct as percentage of manually assigned.** The number of correctly assigned results, divided by the number of results *expected* from the manual assignment, and multiplied by 100. A value of 100 for this criterion means that the program found all of the spin systems found by manual assignment.

- **Correct as percentage of results found.** The number of correctly assigned results, divided by the total number of results *found* by the program, and multiplied by 100. A value of 100 for this criterion means that no false positive results were found, ie. all of the results found are correct.

- **Centre of mass of correct results.** Correct results in the results list are given a weight of 1, incorrect results are given a weight of 0. The centre of mass is calculated relative to the low score end of the results list, divided by the total number of results in the list, and multiplied by 100.

A value of 100 for this criterion means that all of the results at the high-scoring end of the results list are correct, and indicates that the ranking of the results according to score is correct.

3.3 Program Performance Analysis

In order to quantify the performance of the *patt_recog* program in the face of noise and perturbations, a number of experiments were performed, both with with artificially generated data and with real data.

The tests with artificial spectra were aimed at quantifying the *patt_recog* program's response to noise and missing data. The same test pattern was used in all cases - a generic pattern for finding AMX-type spin systems. This pattern was also used by the program which generates artificial spectra (see Figure 3.15).

The tests with real spectra used HCCH-COSY and -TOCSY spectra simultaneously, with different degrees of chemical shift mismatch between the spectra, in order to quantify the degradation in performance as the mismatch is increased.

3.3.1 Performance of the Program in the Presence of Noise

The synthetic spectra used in these tests were constructed in three stages:

1. As a basis for the artificially generated data, a pseudo 3D HCCH-COSY spectrum was generated. The starting point was a block filled with zero-amplitude data points, with 512, 256 and 256 data points in the w1 (proton), w2 (^{13}C) and w3 (proton) dimensions respectively. The sweep widths were notionally 8 ppm, 80 ppm and 8 ppm in the w1, w2, and w3 dimensions, and the corresponding low-field end offset values were -1.48 ppm, 4.5 ppm and -1.48 ppm.

2. Artificial peaks were inserted into this block, using the Hα, Hβ/β', Cα and Cβ chemical shifts from the assignments of 22 of the AMX[1] spin systems in the N-terminal thioredoxin-like domain of protein disulphide isomerase [30] as a basis. These assignments were selected to be i) non overlapping and ii) non-degenerate. The peak pattern is shown in Figure 3.15. All peaks were given a Lorentzian line shape, and set to the arbitrary amplitude of 10^6, with a width of 0.115 ppm in both proton dimensions, and 1.54 ppm in the carbon dimension. Running *patt_recog* on this spectrum (searching for an AMX pattern) produced a perfect results set, ie. all spin systems were correctly found, and there were no false positives.

[1] D1, D7, H8, N16, F17, Y26, Y32, W35, F31, C36, Y46, D66, Y77, Y82, F87, F88, N90, D92, Y99, D107, N110, W111

3. The synthetic spectrum just described was degraded by the addition of various percentages of noise peaks. These peaks were placed with a uniform random spatial distribution into the spectrum, with amplitude variations also following a uniform random distribution over the range 0 to 10^6. Peak width was the same as the "signal" peaks. In other words, the "noise" had, in many cases, a comparable appearance to the signal. Percentage noise was calculated by dividing the number of peaks inserted into the spectrum by the volume of the spectrum (in data points), and multiplying by 100.

Pseudo spectra were generated with 0.5%, 1%, 2%, 3.5% and 5% noise. Three sets of spectra were generated, with different random seed values for the noise generator, so that comparisons could be made more reliable by averaging. All spectra were examined by an experienced spectroscopist, who had extreme difficulty in finding spin systems in the 0.5% noise spectrum, and who was unable to find any spin systems in the spectra containing 1% or more noise. An example Cβ plane from one of the assignments (D92) in a 1% noise spectra is shown in Figure 3.16. It is scarcely possible to discern any signal in this figure by eye. In Figure 3.17, the same plane is shown, with the chemical shifts found by *patt_recog* drawn in.

A graph showing the mean performance criteria at these different noise levels is shown in Figure 3.18. As can be seen, the three performance measures show that performance degrades with increasing perturbation level, but each measure shows a somewhat different degradation profile:

- The correctly assigned results as percentage of those manually assigned (the "o" symbols in the figure) shows an exponential drop off with increasing noise. However, the program always manages to correctly assign *some* spin systems, even at 5% noise, where it is impossible for a human to see any signal at all.

- The correctly assigned results as as percentage of the total number of results found (the "$\diamond$" symbols in the figure) show a double-exponential decay, remaining above 80% even at 2% noise. This means that for low to medium noise levels, nearly all of the results found are reliable.

- The centre of mass of the correct results (the "+" symbols in the figure) never falls below 70%. This shows that, even in the presence of very severe noise contamination, the position of a result in the results list gives a strong indication as to its reliability.

This series of experiments demonstrates that *patt_recog* is capable of finding spin systems even in the presence of significant amounts of noise, where manual assignment would have been impossible.

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	256
Sub-block (data pts)	16	4	16
Sweep width (ppm)	8	80	8
Offset (ppm)	-1.48	4.5	-1.48

Results list processing parameters:

Sym:	Active.		
Exs-penal:	M_x=1		
Too-far:	n_1=0	n_2=4	Δ_u=1.10
Thresh-count:	Active.		
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Cα	Hα	Cβ	Hβ
46.5→64.0	2.87→6.20	23.4→45.0	1.20→4.30

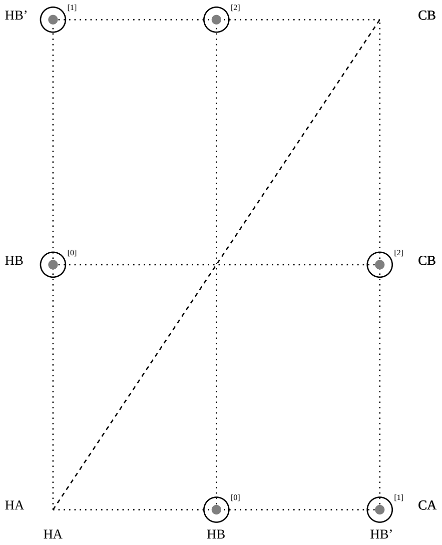


Figure 3.15: **Pattern of peaks utilised in artificial spectra.**
The artificial spectrum is meant to mimic a HCCH-COSY, so for the AMX spin system, one would expect to see a Hβ/Hα/Cα and a Hβ'/Hα/Cα peak in the Cα plane, plus Hα/Hβ/Cβ, Hα/Hβ'/Cβ, Hβ'/Hβ/Cβ and Hβ/Hβ'/Cβ peaks on the Cβ plane.

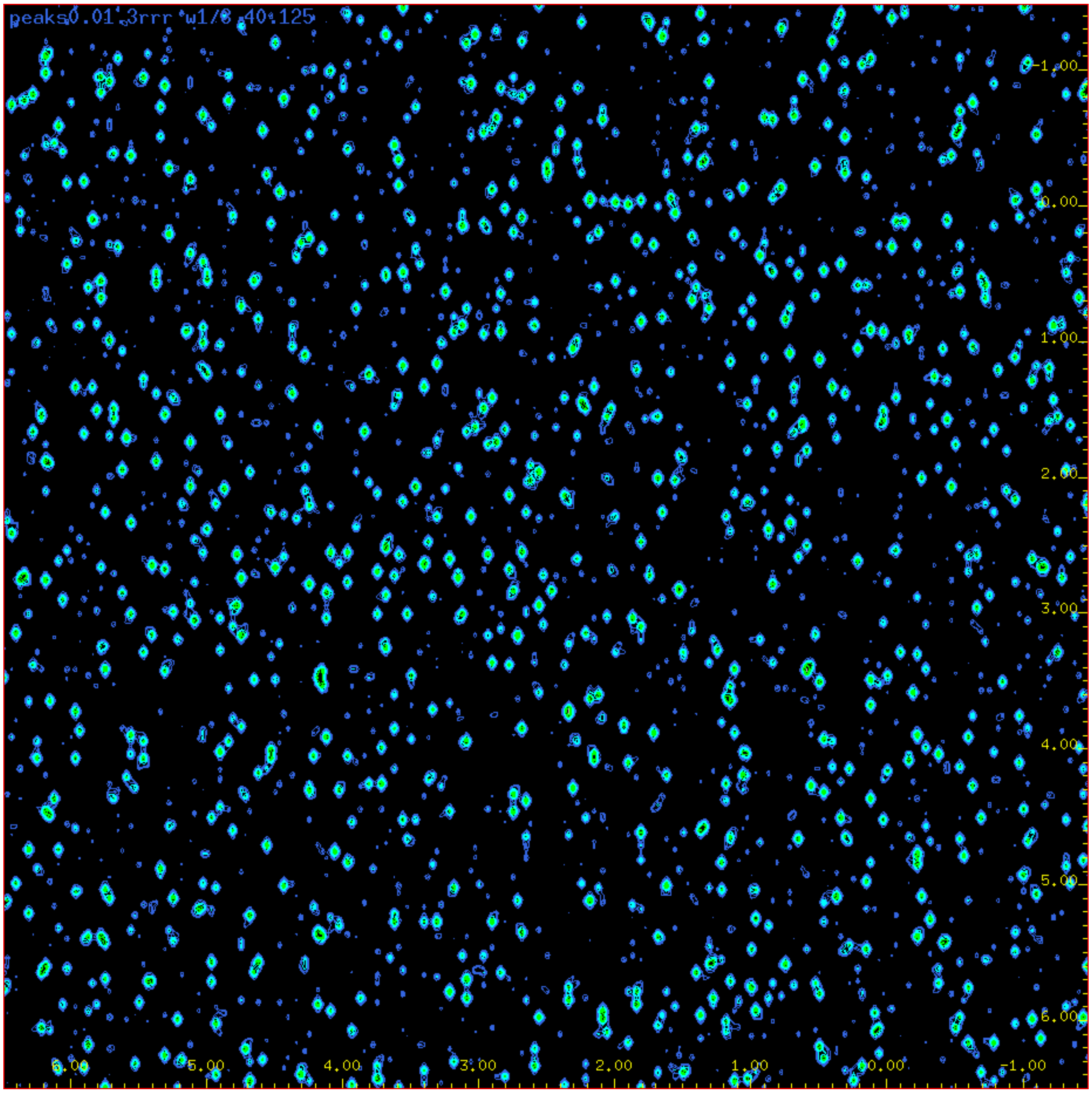


Figure 3.16: **Cβ slice from the 1% noise corrupted spectrum.** This slice contains four peaks belonging to the D92 spin system, but these are almost impossible to find by eye.

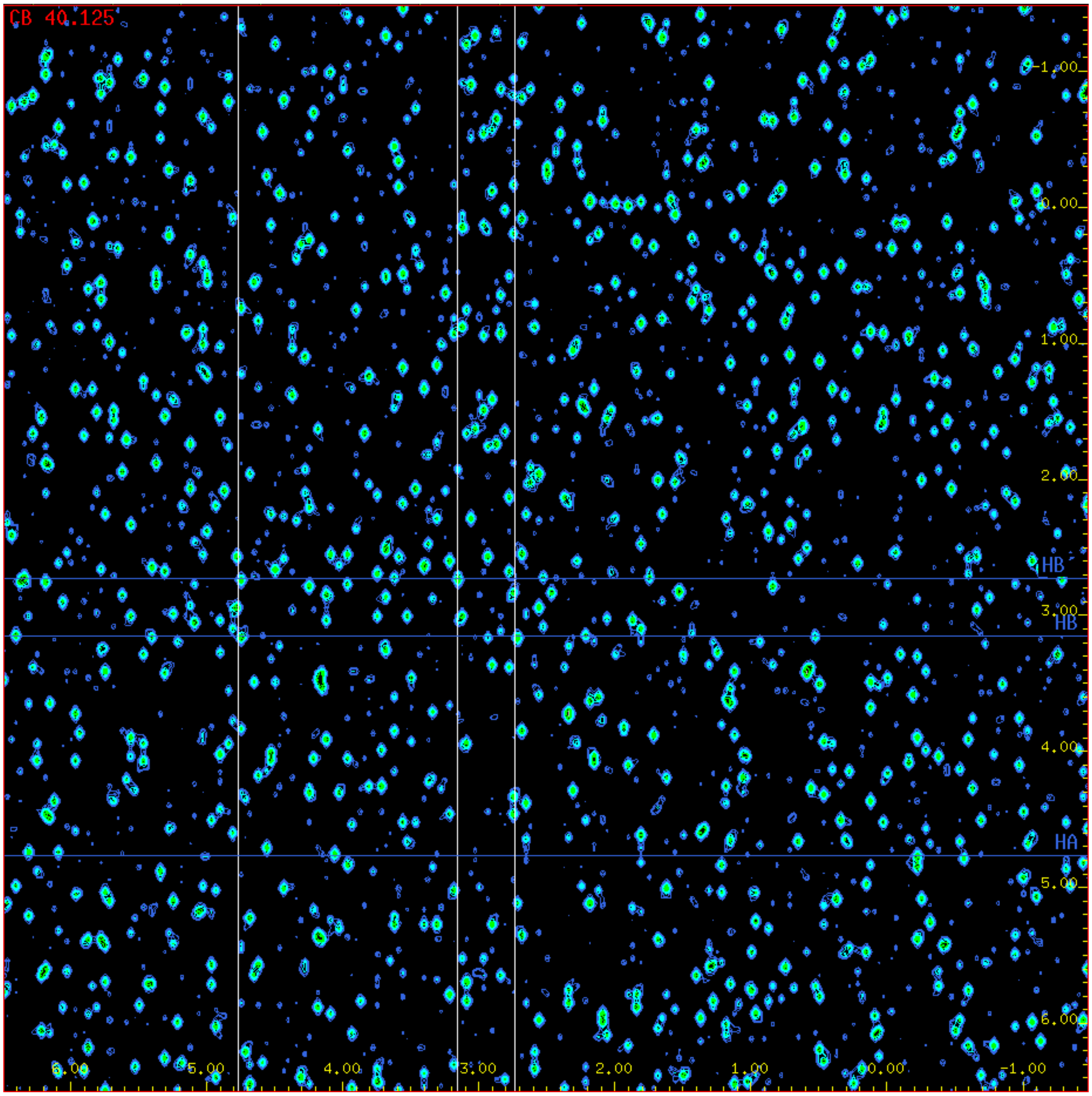
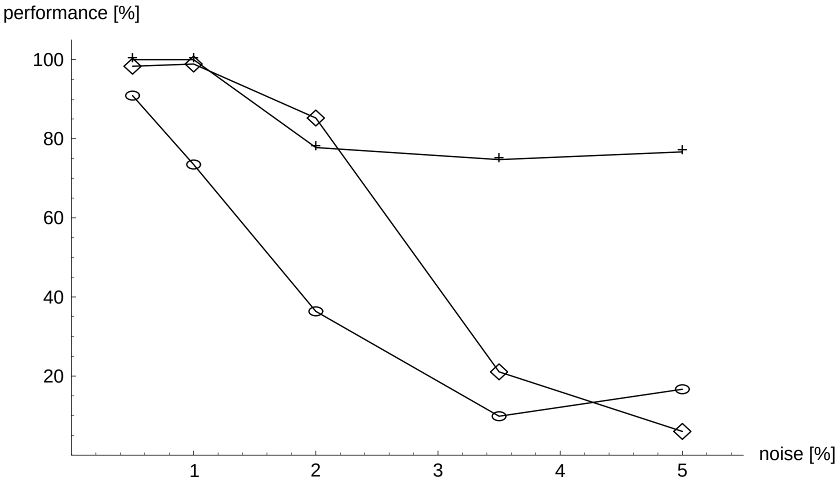


Figure 3.17: **Automatically assigned spin system in the presence of noise, showing automatically found chemical shifts for D92.**
This is the same Cβ slice as shown in Figure 3.16, but now the lines showing the Hα, Hβ and Hβ' chemical shifts found by the *patt_recog* program in this spectrum have been added.

Figure 3.18: **Performance of the *patt_recog* program in the presence of noise.** Three measures are used to characterise the performance of the program, all normalised and expressed in percent, and indicated by three different symbols in the graph. The "o" symbol shows how many correctly found assignments there are compared to the number of assignments expected from the sequence. The "◇" symbol shows the number of correct assignments compared to the total number in the results list. The "+" symbol shows the centre of mass of the correct results, measured from the low-scoring end of the results list, and is hence a measure of the effectiveness of the results list penalisation algorithms. The larger this value, the more correct results appear at the high-scoring of the list.



3.3.2 Performance of the Program with Missing Peaks

For these tests, synthetic spectra were also used. Generating these spectra was a two-step process; the first step was identical to the first step in the generation of noise-corrupted spectra, ie. production of an empty block. The second step was similar to the second step in the generation of noise-corrupted spectra, the same 22 AMX assignments for the N-terminal thioredoxin-like domain of protein disulphide isomerase were used to insert peaks into the block, but some peaks were omitted at random. Five tests were done, with 1%, 5%, 10%, 20% and 50% of peaks missing. As before, these tests were repeated three times with different seeding values for the pseudo-random number generator, to allow some averaging.

A graph showing the mean performance criteria at these different percentages of missing peaks is shown in Figure 3.19. As can be seen, performance degrades with increasing number of missing peaks, but this degradation follows quite a different pattern in comparison to noise corruption:

- The correctly assigned results as percentage of those manually assigned (the "o" symbols in the figure) shows no initial drop off with increasing number of missing peaks, but then starts dropping sharply after the 20% level. A "crossover" effect also occurs at this point; even though 20% of peaks are missing, we are getting 75% of all spin systems correct. But when 50% of all peaks are missing, we only get 24% of all spin systems correct.

- The correctly assigned results as as percentage of the total number of results found (the "⋄" symbols in the figure) show no significant change with increasing number of missing peaks, ie. the program does not show a significant tendency towards finding false positives if peaks are missing.

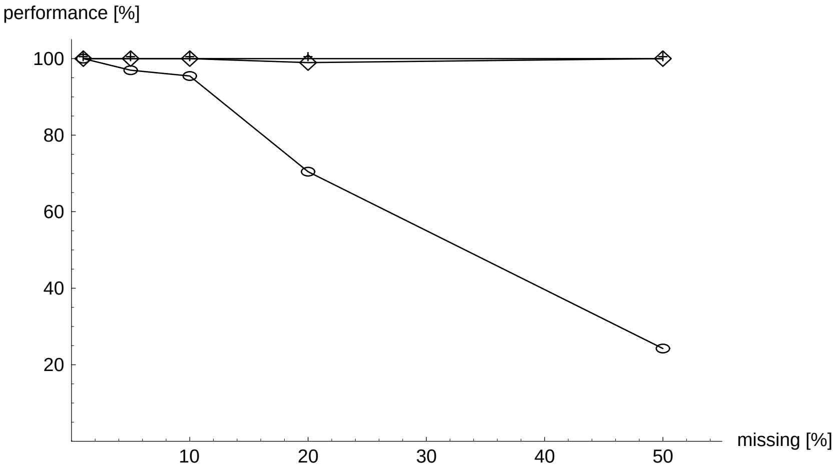
- The centre of mass of the correct results (the "+" symbols in the figure) stays constantly at 100%, reflecting the lack of false positives just mentioned.

3.3.3 Performance of the Program with Perturbations between Spectra

One of the facets of the search algorithm, namely, the fact that it searches in the original spectra, rather than in peak lists, should give it a certain amount of robustness in the face of small mis-calibrations between multiple spectra.

In order to quantify this, the HCCH-COSY and the HCCH-TOCSY from the N-terminal thioredoxin-like domain of protein disulphide isomerase [30] were used. The pattern file used is shown in Figure 3.27 (page 77). Starting from no perturbation at all, the perturbation between the two spectra was incremented

Figure 3.19: **Performance of the *patt_recog* program with missing peaks.** Three measures are used to characterise the performance of the program, all normalised and expressed in percent, and indicated by three different symbols in the graph. The "o" symbol shows how many correctly found assignments there are compared to the number of assignments expected from the sequence. The "◇" symbol shows the number of correct assignments compared to the total number in the results list. The "+" symbol shows the centre of mass of the correct results, measured from the low-scoring end of the results list, and is hence a measure of the effectiveness of the results list penalisation algorithms. The larger this value, the more correct results appear at the high-scoring of the list.



systematically, and the performance of the program checked for each increment. Similar experiments were performed for all three axes of the spectra, giving data for both proton and ^{13}C spins. On the proton axes (axes 0 and 1), perturbations were incremented in steps of 0.02 ppm over the range -0.14 ppm to +0.14 ppm. For the ^{13}C axis (axis 2), perturbations were incremented in steps of 0.2 ppm over the range -1.4 ppm to +1.4 ppm.

In principal, perturbations along both proton axes should have much the same effect on the performance of the program; furthermore, a positive perturbation of δ should have much the same effect as a negative perturbation of $-\delta$. Hence, it should be possible to remove some of the statistical variation by averaging all performance measures for ± 0.02 ppm along both proton axes, ditto for all other perturbation increments.

The graph of performance vs. perturbation for protons is shown in Figure 3.20. As can be seen, the three performance measures show that performance degrades with increasing perturbation, but each measure shows a somewhat different degradation profile:

- The correctly assigned results as percentage of those manually assigned (the "o" symbols in the figure) shows little degradation in performance for perturbations less than or equal to 0.04 ppm; for larger perturbations, performance shows an an almost linear drop off with increasing perturbation. However, even with a perturbation of 0.1 ppm between the two spectra, the program still manages to find over 30% of the spin systems correctly.

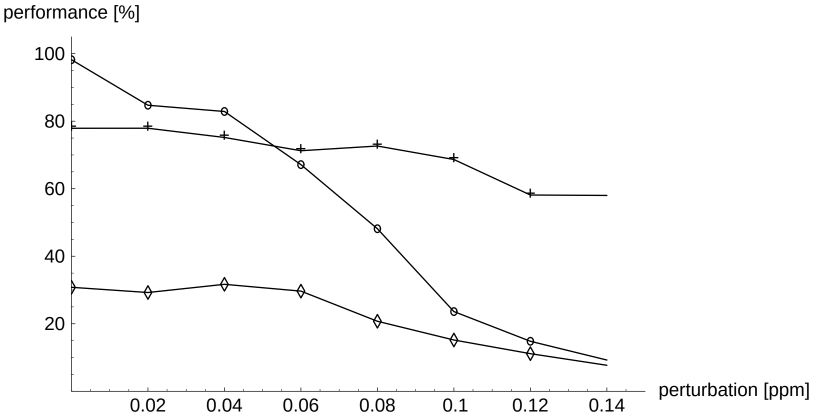
- The correctly assigned results as as percentage of the total number of results found (the "$\diamond$" symbols in the figure) show very little change with increasing perturbation, ie. the program does not find more false positives as a result of perturbation.

- The centre of mass of the correct results (the "+" symbols in the figure) never falls below 70%. This shows that, even in the presence of significant perturbation, the position of a result in the results list gives a strong indication as to its reliability.

In principal, a positive perturbation of δ along the ^{13}C axis should have much the same effect as a negative perturbation of $-\delta$ along this axis. Hence, it should be possible to remove some of the statistical variation by averaging all performance measures for ± 0.2 ppm along the ^{13}C axis, ditto for all other perturbation increments.

The graph of performance vs. perturbation for ^{13}C is shown in Figure 3.21. As can be seen, the three performance measures show that performance degrades with increasing perturbation, but each measure shows a somewhat different degradation profile:

Figure 3.20: **Performance of the *patt_recog* program in the presence of perturbation on the ^{1}H axes between spectra.**
Three measures are used to characterise the performance of the program, all normalised and expressed in percent, and indicated by three different symbols in the graph. The "o" symbol shows how many correctly found assignments there are compared to the number of assignments expected from the sequence. The "◇" symbol shows the number of correct assignments compared to the total number in the results list. The "+" symbol shows the centre of mass of the correct results, measured from the low-scoring end of the results list, and is hence a measure of the effectiveness of the results list penalisation algorithms. The larger this value, the more correct results appear at the high-scoring of the list.



- The correctly assigned results as percentage of those manually assigned (the "o" symbols in the figure) shows little degradation in performance for perturbations less than or equal to 0.6 ppm; for larger perturbations, performance shows an an almost linear drop off with increasing perturbation. However, even with a perturbation of 1.4 ppm between the two spectra, the program still manages to find over 40% of the spin systems correctly. The "steps" in the graph are a result of the poor resolution of the spectrum along this dimension - a data point is approximately 0.3 ppm wide.

- The correctly assigned results as as percentage of the total number of results found (the "⋄" symbols in the figure) show very little change with increasing perturbation up to 1 ppm, after which there is a sharp drop off. Ie. as long as the perturbation between the two spectra is less than the width of a peak, the program does not find more false positives as a result of perturbation.

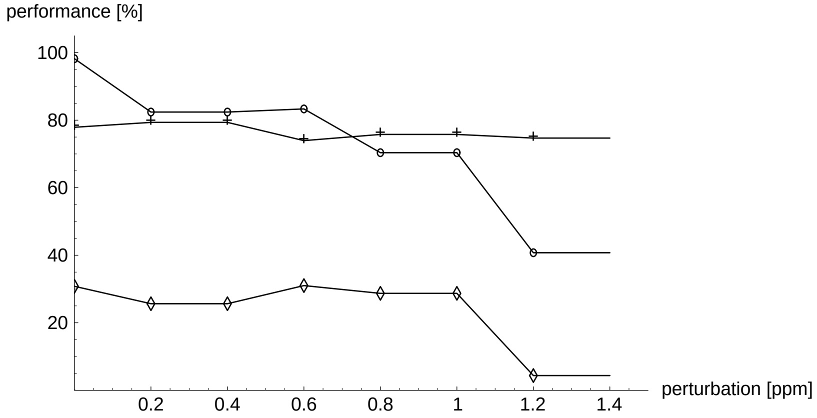
- The centre of mass of the correct results (the "+" symbols in the figure) never falls below 70%. This shows that, even in the presence of significant perturbation, the position of a result in the results list gives a strong indication as to its reliability.

3.4 Pattern Design, A Case Study: The Second PDZ Domain of Rat Postsynaptic Density Protein

The patterns and results lists used in this section were kindly made available by Joachim Walter ([69]). The spectra used were obtained from a ^{14}N labelled sample of the second PDZ domain of rat postsynaptic density protein ([11]) A total of five spectra were used, *viz.* the HSQC-TOCSY, the HSQC-NOESY, the HAHN and the HBHN spectra, plus, for some experiments, a 2D homonuclear TOCSY. These spectra provide, in addition to the amide proton and nitrogen chemical shifts, the Hα chemical shifts and the Hβ chemical shifts.

In order to illustrate the effect of design choices on program performance, a set of four experiments was performed. An "isoleucine" pattern was designed that searched for spin systems with amide H and N spins, plus Hα and a *single* Hβ. The chemical shift ranges were arranged such that this pattern would find not just isoleucines, but alanines as well. These experiments were set up to explore the effects of excess peaks penalisation and the introduction of additional information via the 2D TOSCY spectrum. The experiments performed are as follows:

Figure 3.21: **Performance of the *patt_recog* program in the presence of perturbation on the ^{13}C axis between spectra.**
Three measures are used to characterise the performance of the program, all normalised and expressed in percent, and indicated by three different symbols in the graph. The "o" symbol shows how many correctly found assignments there are compared to the number of assignments expected from the sequence. The "◇" symbol shows the number of correct assignments compared to the total number in the results list. The "+" symbol shows the centre of mass of the correct results, measured from the low-scoring end of the results list, and is hence a measure of the effectiveness of the results list penalisation algorithms. The larger this value, the more correct results appear at the high-scoring of the list.



1. Only the HSQC-TOCSY, HSQC-NOESY, HAHN and HBHN spectra used, with excess peaks penalisation inactive in all spectra (see Figure 3.22 for pattern).
2. All five spectra used (HSQC-TOCSY, HSQC-NOESY, HAHN, HBHN and 2D TOSCY) but with excess peaks penalisation inactive in all spectra (see Figure 3.23 for pattern).
3. All five spectra used (HSQC-TOCSY, HSQC-NOESY, HAHN, HBHN and 2D TOSCY) but excess peaks penalisation active in the HAHN and HBHN spectra only (see Figure 3.23 for pattern).
4. All five spectra used (HSQC-TOCSY, HSQC-NOESY, HAHN, HBHN and 2D TOSCY) plus excess peaks penalisation active in all spectra (see Figure 3.23 for pattern).

Because these spectra have not yet been assigned, and because the quantity of results was large, results were assessed by randomly selecting 15 results from each results set and examining them manually. Results were sorted into two classes, "good" and "bad". Statistics for the results sets were then calculated by estimating the total number of "good" results by scaling up the count obtained from the sample:

$$E_g = E\ E_{gs}/15$$

where E is the total results count, E_{gs} is the number of "good" results in the sample and E_g is the estimated total "good" results count. The expected number of "good" results for all experiments is about 50 (based on the length of the sequence); the actual results are summarised in the following table:

Experiment	Total result count	estimated "good" count
1	448	58
2	171	80
3	32	24
4	77	46

In the case of experiment 1, the number of "good" results is acceptable, but the total result count is far too large, it would be unreasonable to expect a human operator to examine all of these results by hand. The absence of TOCSY data is the main reason for this proliferation of results; wherever spin systems have degenerate amide proton and nitrogen chemical shifts, there will be many peaks on a single strip in the HAHN spectrum and on the corresponding strip in the HBHN spectrum. This allows the program to generate large numbers of results by simple combinatorics.

Spectrum size parameters:

	HSQC-TOCSY			HNHA				HNHB	
	w1	w2	w3	w1	w2	w3	w1	w2	w3
Size (data pts)	256	384	256	256	384	256	256	384	256
Sub-block (data pts)	32	128	16	32	128	16	32	128	16
Sweep width (ppm)	49.526	6.280	13.886	49.525	6.249	7.574	49.525	6.249	10.414
Offset (ppm)	95.046	4.779	-2.114	95.046	4.800	2.980	95.047	4.800	-0.391

Results list processing parameters:

Cut: $epsilon_{low}$=0.70
Exs-penal: M_x=1

Chemical shift ranges:

N	H	Hα	Hβ
100.0→135.0	6.10→10.70	3.00→6.10	-0.30→5.50

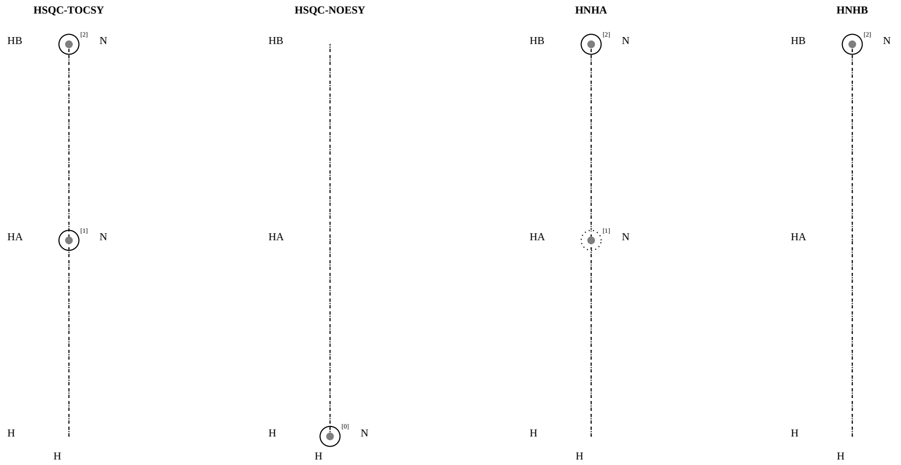


Figure 3.22: **“Isoleucine” pattern for test 1.**

Spectrum size parameters:

	HSQC-TOCSY			HNHA			HNHB		
	w1	w2	w3	w1	w2	w3	w1	w2	w3
Size (data pts)	256	384	256	256	384	256	256	384	256
Sub-block (data pts)	32	128	16	32	128	16	32	128	16
Sweep width (ppm)	49.526	6.280	13.886	49.525	6.249	7.574	49.525	6.249	10.414
Offset (ppm)	95.046	4.779	-2.114	95.046	4.800	2.980	95.047	4.800	-0.391

	2D TOCSY	
	w1	w2
Size (data pts)	1024	512
Sub-block (data pts)	64	64
Sweep width (ppm)	16.664	13.886
Offset (ppm)	-3.515	-2.116

Results list processing parameters:

Cut:	$epsilon_{low}$=0.70
Exs-penal:	M_x=1
Excess-peaks:	$P_{expected}$=2 r_l=140

Chemical shift ranges:

N	H	Hα	Hβ
100.0→135.0	6.10→10.70	3.00→6.10	-0.30→5.50

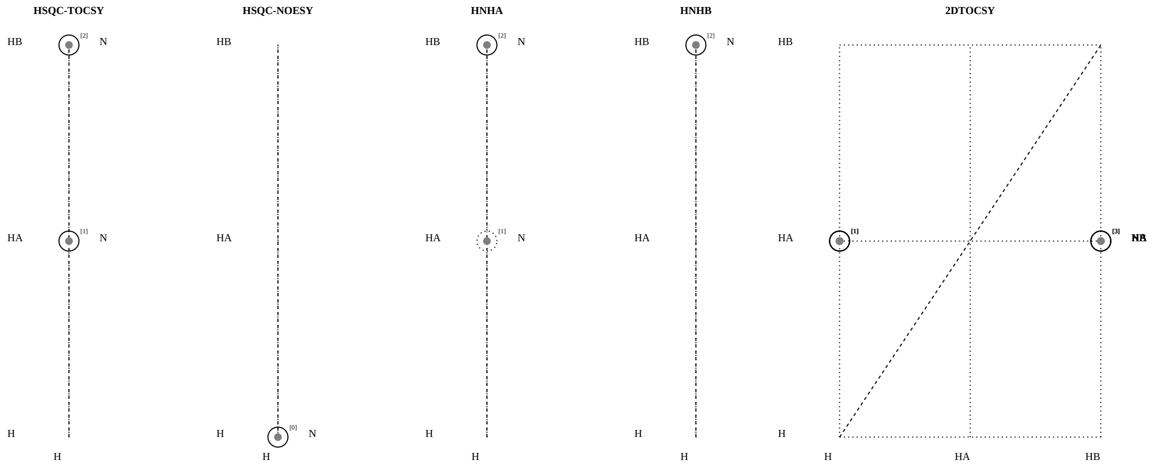


Figure 3.23: **"Isoleucine" pattern for tests 2, 3 and 4.**

Experiment 2 produces too many "good" results, suggesting unwanted redundancy in the results list. It differs from experiment 1, in that information from the 2D TOCSY is now available. This eliminates a large number of the combinations that were being found in experiment 1, and is responsible for the reduction in results count.

In experiment 3, only about half of the expected "good" results are found. This experiment is different from experiment 2, in that excess peaks penalisation is applied to the HAHN and HBHN spectra. The low results count is a result of heavy penalisation by the excess peaks penalisation algorithm - for this pattern, heavily penalised results were deleted. The problem is caused by degenerate strips - degeneracy will lead to there being more peaks in a strip than needed by the pattern for a single spin system. The additional peaks will therefore be treated as "excess" and cause all results generated from this strip to be penalised.

The pattern for experiment 4 was the one chosen as the best in this design exercise, because the number of "good" results comes close to the expected count, and because the total results count is small. The problem with the excess peaks in experiment 3 was overcome by putting excess peaks penalisation into the 2D TOCSY as well, and insisting that the peaks be found both in the HAHN or the HBHN *and* in the 2D TOCSY.

3.5 Results Obtained with the N-terminal Thioredoxin-like Domain of Protein Disulphide Isomerase, Two Case Studies

Two sets of pattern search experiments are presented in this section: the first concentrates on side-chains, and the second on the backbone. In the first set, the patterns for glycine, alanine, the AMX spin systems[2], serine, the GLX spin systems[3], threonine, valine, proline, isoleucine and leucine, were tested on the HCCH-COSY and -TOCSY spectra. In the second set, the six backbone spectra[4] were used to look for sequence fragments containing one amide nitrogen and proton, and the associated CO/Cα/Cβ spins. The unique Cα chemical shift range of glycine was also used during the construction of this latter pattern set.

In designing a search pattern, one may aim for different kinds of results, eg. a small number of results all of which are definitely correct, or a larger number of results amongst which there may be some incorrect ones, but will contain all possible correct instances of the pattern being searched for. Pattern searches will be performed with relatively strict conditions if a small results list is desired, or with more loose constraints if a more inclusive results list is to be

[2] aspartame, aspartic acid, cysteine, histidine, phenylalanine, tryptophan and tyrosine

[3] glutamine, glutamic acid and methionine

[4] HNCA, HN(CO)CA, HNCO HN(CA)CO, CBCANH, CBCA(CO)NH

produced. In this latter case, additional selection would then take place during the sequential assignment.

In the following subsections, the basic patterns for relatively tightly constrained side-chain and backbone searches will be described, together with the parameters for the heuristic result list filtering algorithms, and the results produced will be analysed. The spectra used come from the N-terminal thioredoxin-like domain of protein disulphide isomerase, a protein containing 120 residues, kindly provided by Johan Kemmink[5]. See [30] and [29] for full details of experimental conditions.

3.5.1 Evaluating Side Chain Patterns

The HCCH-COSY and -TOCSY spectra were recorded in an orthogonal manner, ie. in one case all cross peaks should occur along F_1 and in the other along F_2. In this way, some effects of t_1 noise could be compensated for. A mixing time of 24 ms was used for the HCCH-TOCSY. The sample was dissolved in D_2O, with only mild water suppression by irradiation applied. The resolution for both COSY and TOCSY was 13.4 Hz per data point on the proton axes, and 39.1 Hz per data point on the carbon axis.

As a key feature of the side-chain pattern search, limits were imposed on the chemical shift ranges scanned by the program, in order to make the search more specific, and to speed up operation. The limits used in the tests described below are based on chemical shift values published in the literature ([23], [74]).

A summary of the results presented in this section is shown in Figure 3.24. It can be seen that the centre of mass measure is always high, showing that the ordering of the results in the results list is a good indication of their reliability. Also, in the majority of cases, more than 70% of the expected residues were found. However, in a number of cases, the number of correct results found was small relative to the total number of results found. It would thus be reasonable to say that, as long as one uses only the highest scoring results, one has a very good chance of selecting valid assignments. The results of these pattern searches are discussed in detail below.

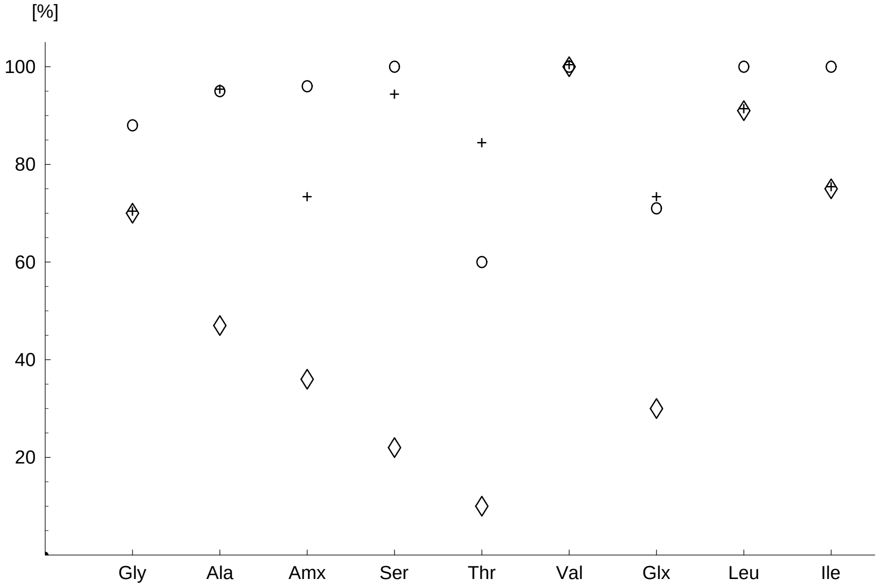
The results are tabulated in Appendix A.

Gly
Number of results expected (from the sequence): 8.
Actual results count: 10
Number of correct results found: 8.

The glycines in these spectra show no degeneracy in the $H\alpha/H\alpha'$ chemical shifts, hence the program was easily able to identify all chemical shifts correctly. Figure

[5] European Molecular Biology Laboratory, Heidelberg

Figure 3.24: **Summary of side chain results.**
Three measures are used to characterise the performance of the program with a given pattern, all normalised and expressed in percent, and indicated by three different symbols in the graph. The "o" symbol shows how many correctly found assignments there are compared to the number of assignments expected from the sequence. The "◇" symbol shows the number of correct assignments compared to the total number in the results list. The "+" symbol shows the centre of mass of the correct results, measured from the low-scoring end of the results list, and is hence a measure of the effectiveness of the results list penalisation algorithms. The larger this value, the more correct results appear at the high-scoring of the list.



3.25 shows the pattern and chemical shift ranges for glycine, along with the results list filtering parameters.

Ala
Number of results expected (from the sequence): 20.
Actual results count: 43
Number of correct results found: 19.

Alanine constitutes one of the simplest patterns, but the β-carbon has a very characteristic chemical shift, allowing it to be distinguished from other spin systems. Figure 3.26 shows the pattern and chemical shift ranges for alanine, along with the results list filtering parameters. The correct results are concentrated at the top of the results list, indicating that the ranking introduced by the results list filtering heuristics is correct. Nonetheless, the results list also contains many non-alanines, for various reasons. First, the alanine pattern is *so* simple, that peaks from other spin systems are also found. For example, the Hγ/Hδ/Cδ peaks of proline, or the Hβ/Hα/Cα peaks of leucine, fall within the search ranges for the Hβ/Hα/Cα peaks of alanine. In some of these cases, partially correct alanine spin systems are found, with a valid peak in the Cβ plane, but a false peak in the Cα plane. Such results *could*, in principle, be filtered out by the secondary clustering mechanism (see previous chapter), but then alanines with overlapping chemical shifts, such as A50 and A105, would not be distinguished by the program. Finally, some of the results *look* like alanines when manually inspected, but do not fit into the results from the backbone spectra. It is possible that these are due to minor conformations, caused by proline cis-trans isomerisation.

AMX
Number of results expected (from the sequence): 27.
Actual results count: 73
Number of correct results found: 23.

The AMX pattern matches many other spin system patterns, of which it is a subset, such as glutamine, glutamic acid and methionine. Nevertheless, the normalised centre of mass of the results, 0.726, confirms the impression given by a manual inspection of the results list, namely, that most of the correct results are at the high-scoring end of the results list. This was achieved using forbidden peak penalisation, allowing the program to penalise results which were actually subsets of larger spin systems. Figure 3.27 shows the pattern and chemical shift ranges for the AMX spin systems, along with the results list filtering parameters.

Ser
Number of results expected (from the sequence): 4.
Actual results count: 18

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	256
Sub-block (data pts)	16	4	16
Sweep width (ppm)	8	80	8
Offset (ppm)	-1.48	4.5	-1.48

Results list processing parameters:

Ovrlp-clus:	i=1	j=0	Δ_t=0.15
Sym:	Active.		
Sec-clus:	N_l=-1	Δ_l=0.10	Δ_u=2.00
Fract-score:	ϵ_{low}=0.001		
Exs-penal:	M_x=1		
Too-far:	n_1=1	n_2=0	Δ_u=1.70
Thresh-count:	Active.		
Excess-peaks:	$P_{expected}$=1	r_l=10000	
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Cα	Hα
38.5→48.5	3.00→4.36

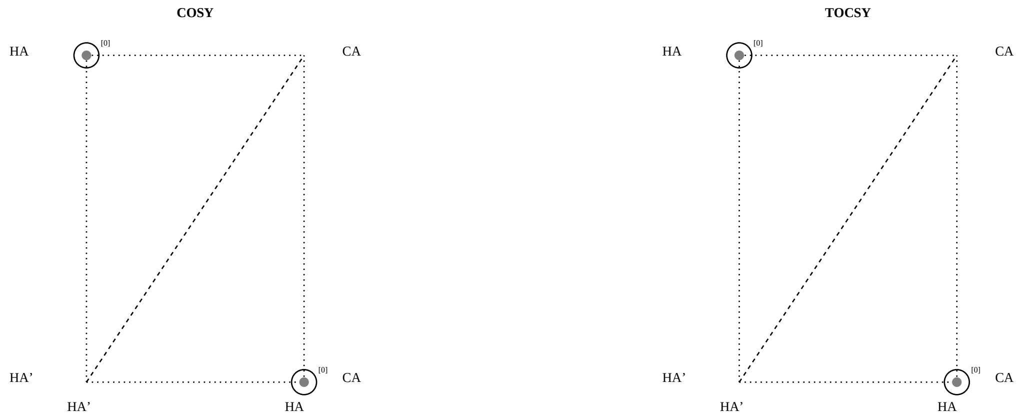


Figure 3.25: **Glycine pattern.**

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	256
Sub-block (data pts)	16	4	16
Sweep width (ppm)	8	80	8
Offset (ppm)	-1.48	4.5	-1.48

Results list processing params:

Sym:	Active.		
Sec-clus:	N_l=2	Δ_l=0.10	Δ_u=1.00
Fract-score:	ϵ_{low}=0.001		
Std-dev:	Active.		
Thresh-count:	Active.		
Excess-peaks:	$P_{expected}$=1	r_l=2500	
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Cα	Hα	Cβ	Hβ
45.8→57.8	3.17→5.80	13.0→26.0	-0.40→1.92

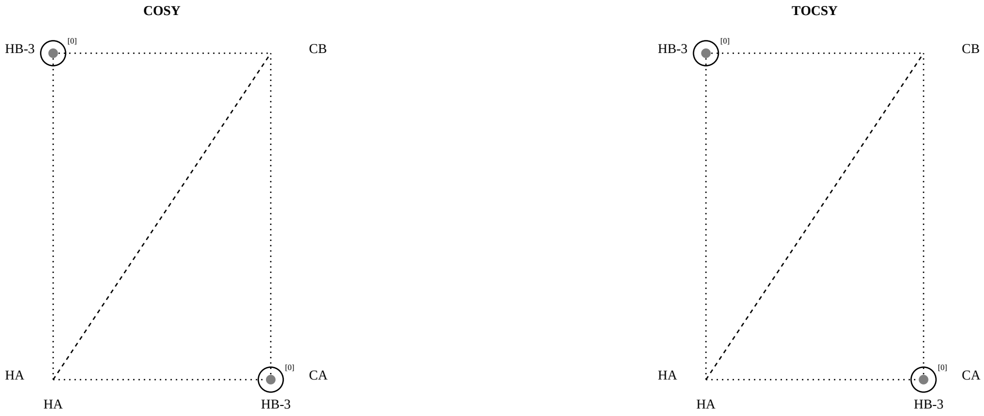


Figure 3.26: **Alanine pattern.**

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	256
Sub-block (data pts)	16	4	16
Sweep width (ppm)	8	80	8
Offset (ppm)	-1.48	4.5	-1.48

Results list processing parameters:

Ovrlp-clus:	i=0	j=4	Δ_t=0.15
Sym:	Active.		
Sec-clus:	N_l=3	Δ_l=0.30	Δ_u=70.00
Fract-score:	ϵ_{low}=0.001		
Too-far:	n_1=0	n_2=4	Δ_u=1.10
Thresh-count:	Active.		
Excess-peaks:	$P_{expected}$=3	r_l=1000	
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Cα	Hα	Cβ	Hβ
46.5→64.0	2.87→6.20	23.4→45.0	1.20→4.30

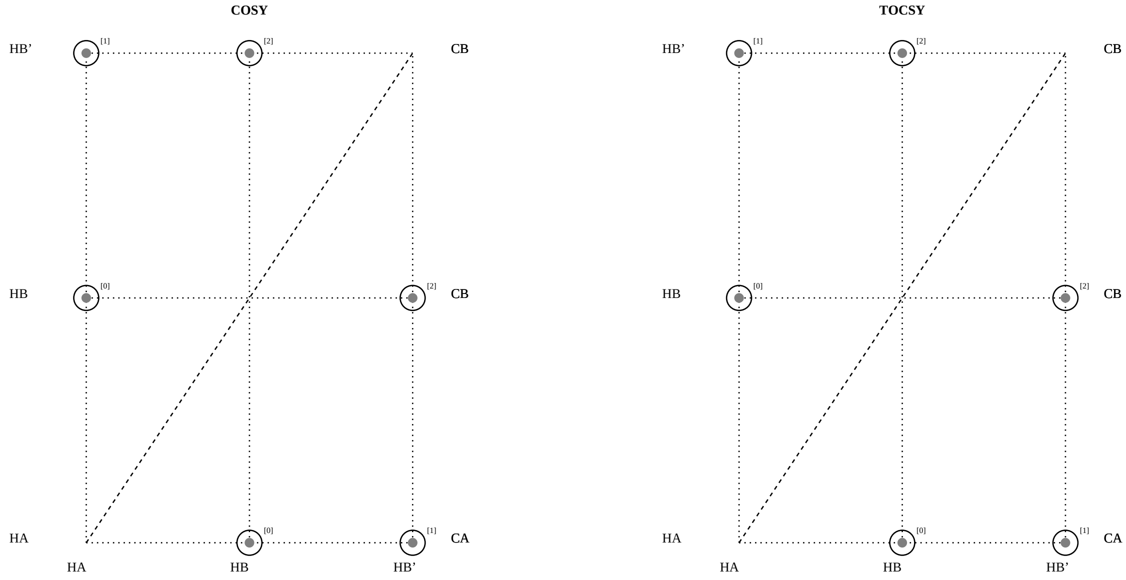


Figure 3.27: **AMX pattern.**

Number of correct results found: 3.

The program actually found the correct chemical shifts for all 4 serines, but because S71 has degenerate Hβ/Hβ' *and* exhibits extensive chemical shift overlap with *all* of the other serines, an extra, non-degenerate, Hβ' was added by *patt_recog*, even though it wasn't needed. All correct assignments are at the high-scoring end of the results list. Figure 3.28 shows the pattern and chemical shift ranges for serine, along with the results list filtering parameters.

Thr
Number of results expected (from the sequence): 5.
Actual results count: 31
Number of correct results found: 3.

The two results not found had either Hα or Hβ chemical shifts on the water line. Figure 3.29 shows the pattern and chemical shift ranges for threonine, along with the results list filtering parameters.

Val
Number of results expected (from the sequence): 6.
Actual results count: 6
Number of correct results found: 6.

The valine peak pattern is shown in Figure 3.30, along with the chemical shift ranges and results list filtering parameters used. The uniqueness of this pattern, plus the good dispersion of valine peaks in the spectra used, plus the fact that these peaks did not overlap significantly with peaks from other spin systems, combined to produce an exceptionally clean results list.

GLX
Number of results expected (from the sequence): 14.
Actual results count: 33
Number of correct results found: 7.

Four partially correct results were also found, where one or two chemical shifts were wrongly assigned. The problems with these spin systems are: i) Cβ and Cγ have overlapping ranges between about 32 and 34 ppm, and ii) many of the GLX spin systems had almost identical Cγ chemical shifts at about 36 ppm. Hence, the program had difficulties disentangling the peaks to construct coherent spin systems. Figure 3.31 shows the pattern and chemical shift ranges for the GLX spin systems, along with the results list filtering parameters.

Ile
Number of results expected (from the sequence): 3.

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	256
Sub-block (data pts)	16	4	16
Sweep width (ppm)	8	80	8
Offset (ppm)	-1.48	4.5	-1.48

Results list processing parameters:

Ovrlp-clus:	i=0	j=4	Δ_t=0.15
Sym:	Active.		
Sec-clus:	N_l=-1	Δ_l=0.15	Δ_u=1.00
Fract-score:	ϵ_{low}=0.001		
Exs-penal:	M_x=2		
Too-far:	n_1=0	n_2=4	Δ_u=1.10
Thresh-count:	Active.		
Excess-peaks:	$P_{expected}$=2	r_l=5000	
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Cα	Hα	Cβ	Hβ
47.0→63.0	3.50→6.10	55.0→70.0	2.87→4.60

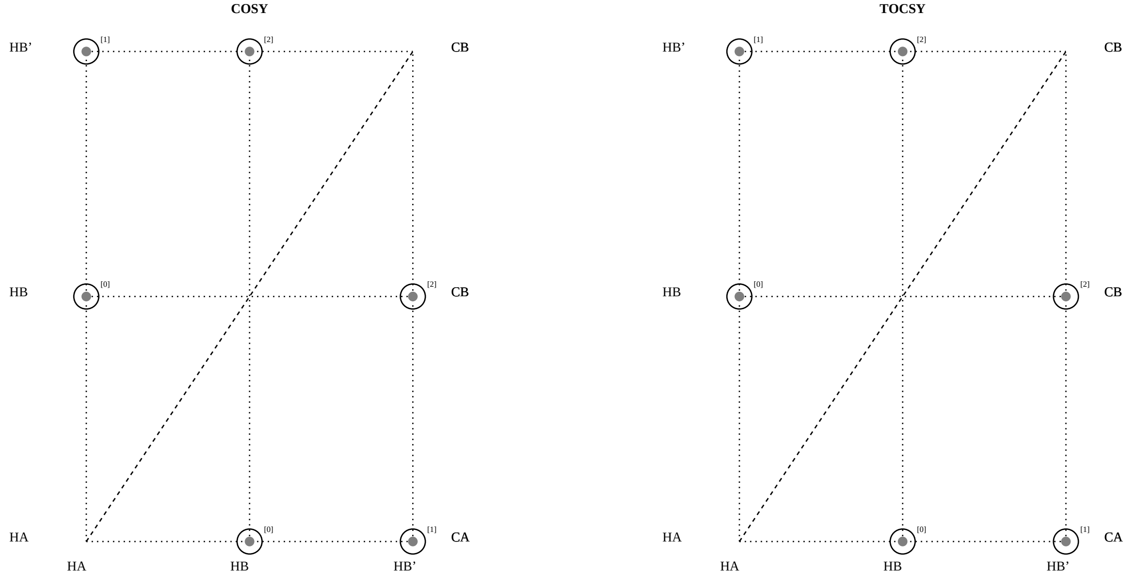


Figure 3.28: **Serine pattern.**

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	256
Sub-block (data pts)	16	4	16
Sweep width (ppm)	8	80	8
Offset (ppm)	-1.48	4.5	-1.48

Results list processing parameters:

Ovrlp-clus:	i=1	j=0	Δ_t=0.15
Sym:	Active.		
Sec-clus:	N_l=4	Δ_l=0.30	Δ_u=70.00
Fract-score:	ϵ_{low}=0.001		
Exs-penal:	M_x=2		
Thresh-count:	Active.		
Excess-peaks:	$P_{expected}$=3	r_l=1000	
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Cα	Hα	Cβ	Hβ
54.6→67.6	3.00→6.00	62.5→74.5	3.40→5.00

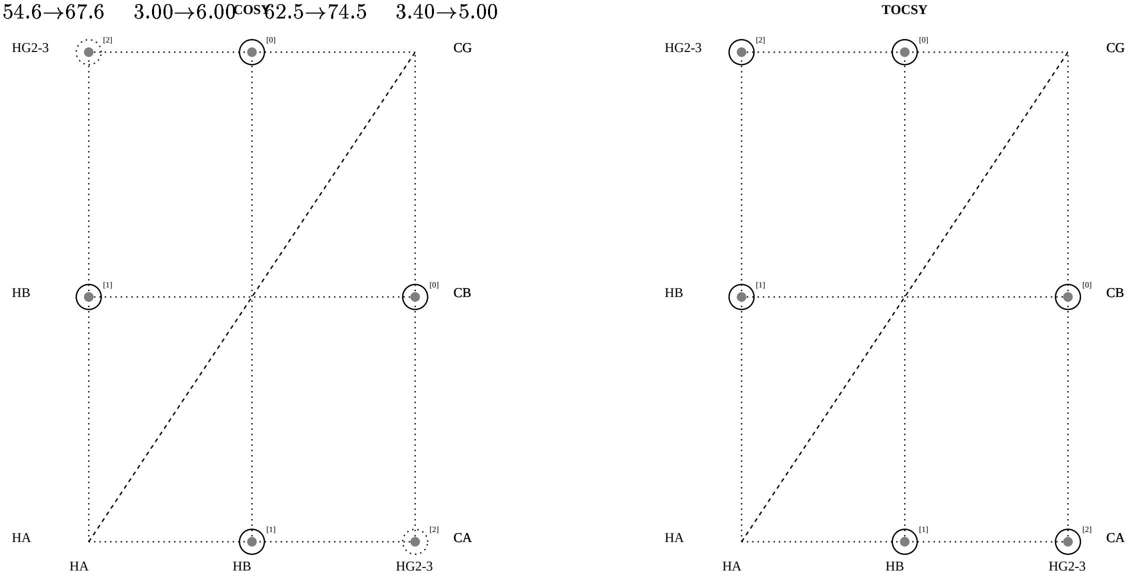


Figure 3.29: **Threonine pattern.**

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	256
Sub-block (data pts)	16	4	16
Sweep width (ppm)	8	80	8
Offset (ppm)	-1.48	4.5	-1.48

Results list processing params:

Ovrlp-clus:	i=4	j=6	Δ_t=0.15
Sym:	Active.		
Sec-clus:	N_l=3	Δ_l=0.15	Δ_u=1.00
Fract-score:	ϵ_{low}=0.001		
Exs-penal:	M_x=2		
Std-dev:	Active.		
Thresh-count:	Active.		
Excess-peaks:	$P_{expected}$=3	r_l=2500	
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Cα	Hα	Cβ	Hβ
49.0→68.2	1.50→6.00	26.8→36.4	0.81→2.75

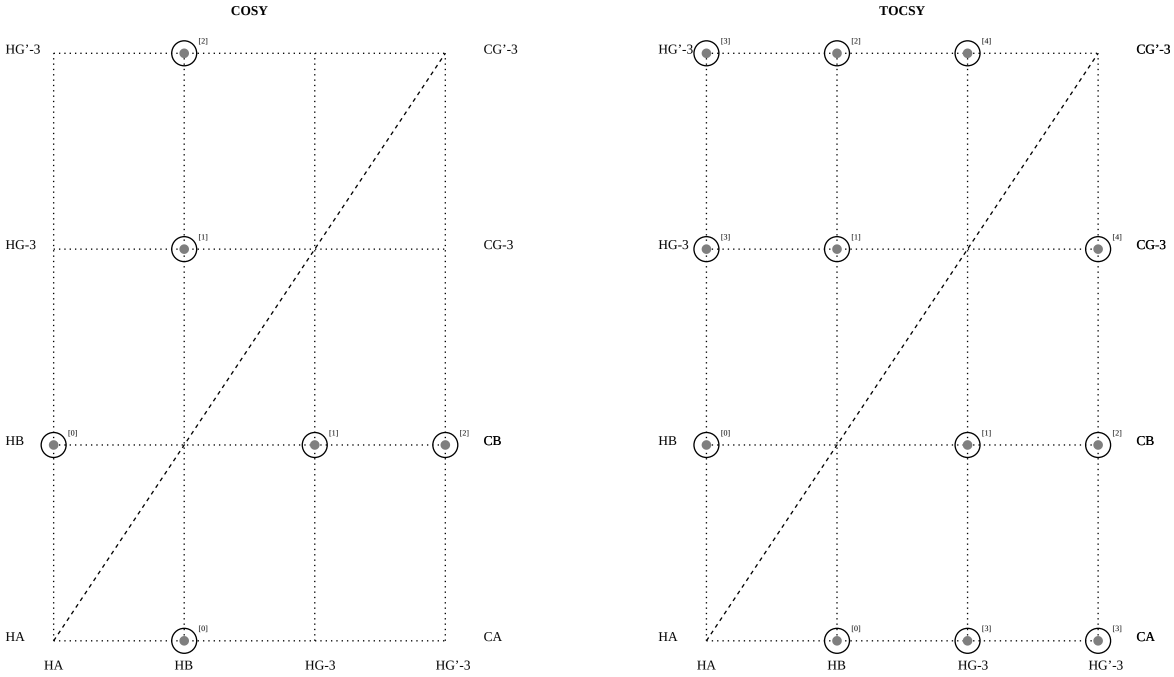


Figure 3.30: **Valine pattern.**

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	256
Sub-block (data pts)	16	4	16
Sweep width (ppm)	8	80	8
Offset (ppm)	-1.48	4.5	-1.48

Results list processing parameters:

Ovrlp-clus:	i=0	j=4	Δ_t=0.15
Sym:	Active.		
Sec-clus:	N_l=5	Δ_l=0.30	Δ_u=70.00
Fract-score:	ϵ_{low}=0.001		
Exs-penal:	M_x=2		
Too-far:	n_1=0	n_2=4	Δ_u=1.10
Thresh-count:	Active.		
Excess-peaks:	$P_{expected}$=5	r_l=1000	
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Cα	Hα	Cβ	Hβ
43.4→66.0	2.60→5.60	23.0→36.5	0.80→2.81

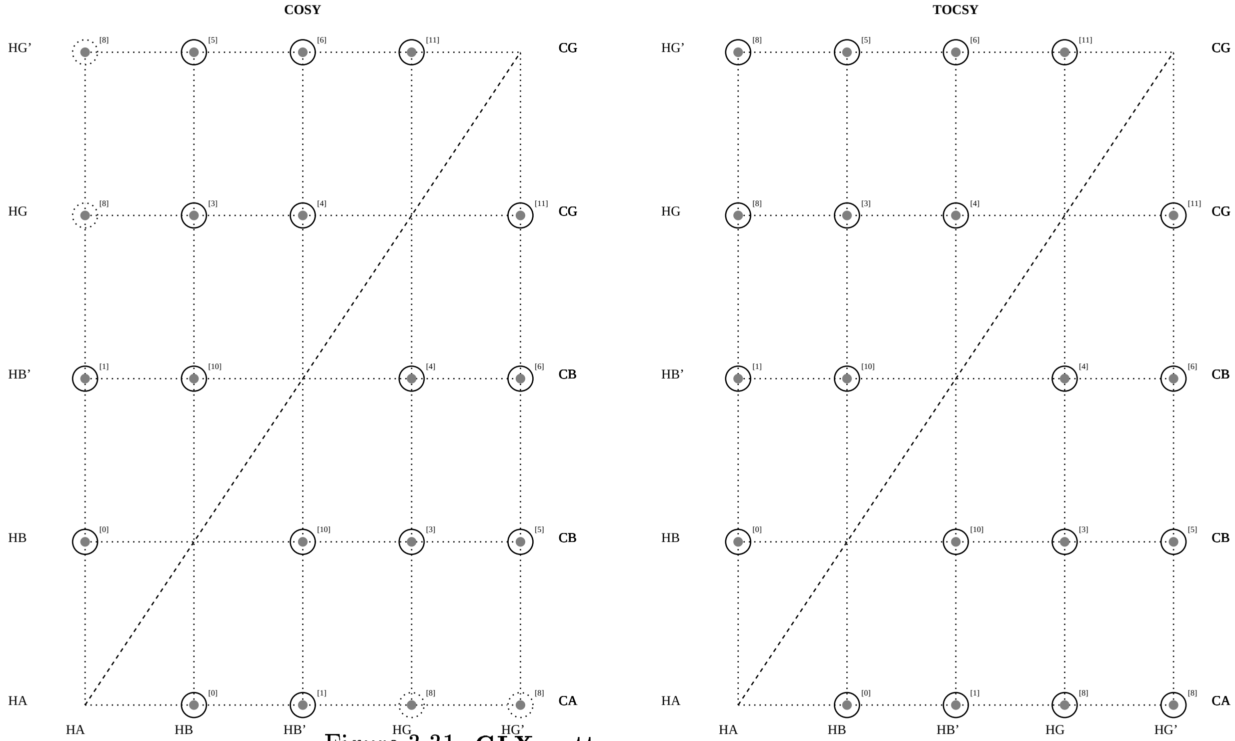


Figure 3.31: **GLX pattern.**

Actual results count: 4
Number of correct results found: 4.

Isoleucine has a rather unique pattern of peaks. However, identifying these patterns was not a trivial task, due to very weak peaks between Hβ and Hγ_1 protons in the TOCSY spectrum. The pattern used to tackle these problems is shown in Figure 3.32, along with the associated chemical shift ranges and results list filtering parameters. Although the sequence only contains three isoleucines, the program has found a fourth "ghost" spin system with no corresponding Cα or Cβ chemical shifts in the CBCANH or CBCA(CO)NH spectra. Manual inspection indicates that this is indeed a genuine isoleucine spin system. Careful inspection of the ^{13}C-3D-NOESY confirmed the presence of this additional spin system, which is probably also due to the existence of cis-trans isomeric prolines.

The pattern search also yielded a second proton at each Cγ_1, which were not found in the original manual assignment.

In Figures 3.33 and 3.34, one of the isoleucine results is shown graphically. This is interesting for a number of reasons. It displays near-degeneracy for both the Hβ/Hγ_1' and for the Hγ_1/Hγ_2 protons. The program has successfully found this pattern, but note how it has found two slightly different chemical shift values for Hβ and Hγ_1'. This is a side effect of the peak proximity penalisation, which leads to a calculation of two slightly different chemical shifts, even in cases where identical chemical shifts would be correct. Finally, this spin system shows extensive overlap with the "ghost" isoleucine, demonstrating the program's ability to distinguish between similar spin systems.

Leu
Number of results expected (from the sequence): 10.
Actual results count: 11
Number of correct results found: 10.

Leucine also produces a unique pattern of peaks, and the β-carbon chemical shift is unusually large. Figure 3.35 shows the pattern and chemical shift ranges used, along with the results list filtering parameters. As with the isoleucines however, the peaks between Hβ and Hγ were small, making this a challenging problem for the program. All ten leucine patterns were found, although in one case, an Hγ chemical shift did not agree with the value reported by Kemmink *et al* (assignments are available as supplementary information to [30]), and in another case, the Hβ was incorrectly assigned by the program as degenerate with the Hβ'. Also, one Cδ chemical shift found by the program was incorrect. The "extra" result found was composed almost entirely of peaks from a valine spin system.

In Figures 3.36 and 3.37, one of the leucine results is shown graphically. This is an interesting case, because the Cδ and Cδ' are very close together. Also, the Cα plane in the HCCH-COSY spectrum contains significant artifacts. In spite

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	256
Sub-block (data pts)	16	4	16
Sweep width (ppm)	8	80	8
Offset (ppm)	-1.48	4.5	-1.48

Results list processing params:

Ovrlp-clus:	i=6	j=8	Δ_t=0.15
Sym:	Active.		
Sec-clus:	N_l=5	Δ_l=0.15	Δ_u=25.00
Fract-score:	ϵ_{low}=0.001		
Exs-penal:	M_x=4		
Too-far:	n_1=6	n_2=8	Δ_u=1.20
Std-dev:	Active.		
Thresh-count:	Active.		
Excess-peaks:	$P_{expected}$=5	r_l=2500	
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Cα	Hα	Cβ	Hβ
49.0→66.6	3.06→6.00	30.1→42.6	0.73→2.70

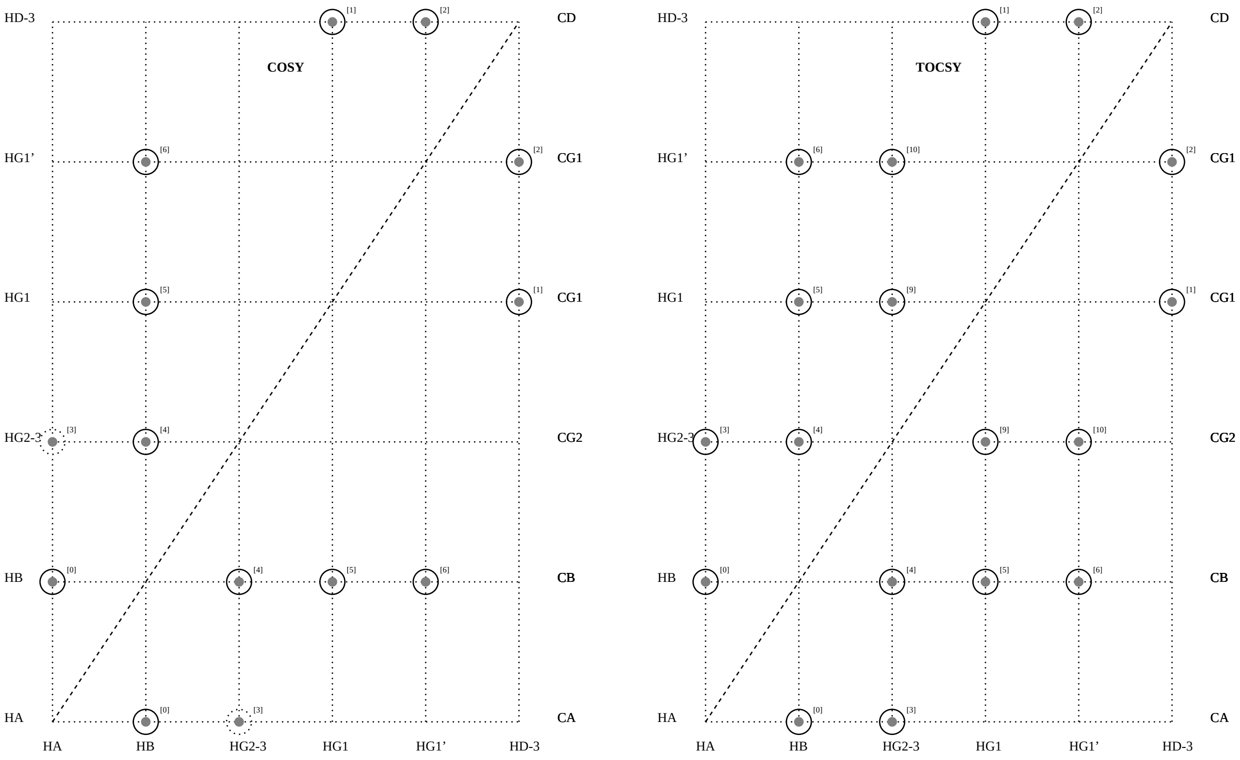


Figure 3.32: **Isoleucine pattern.**

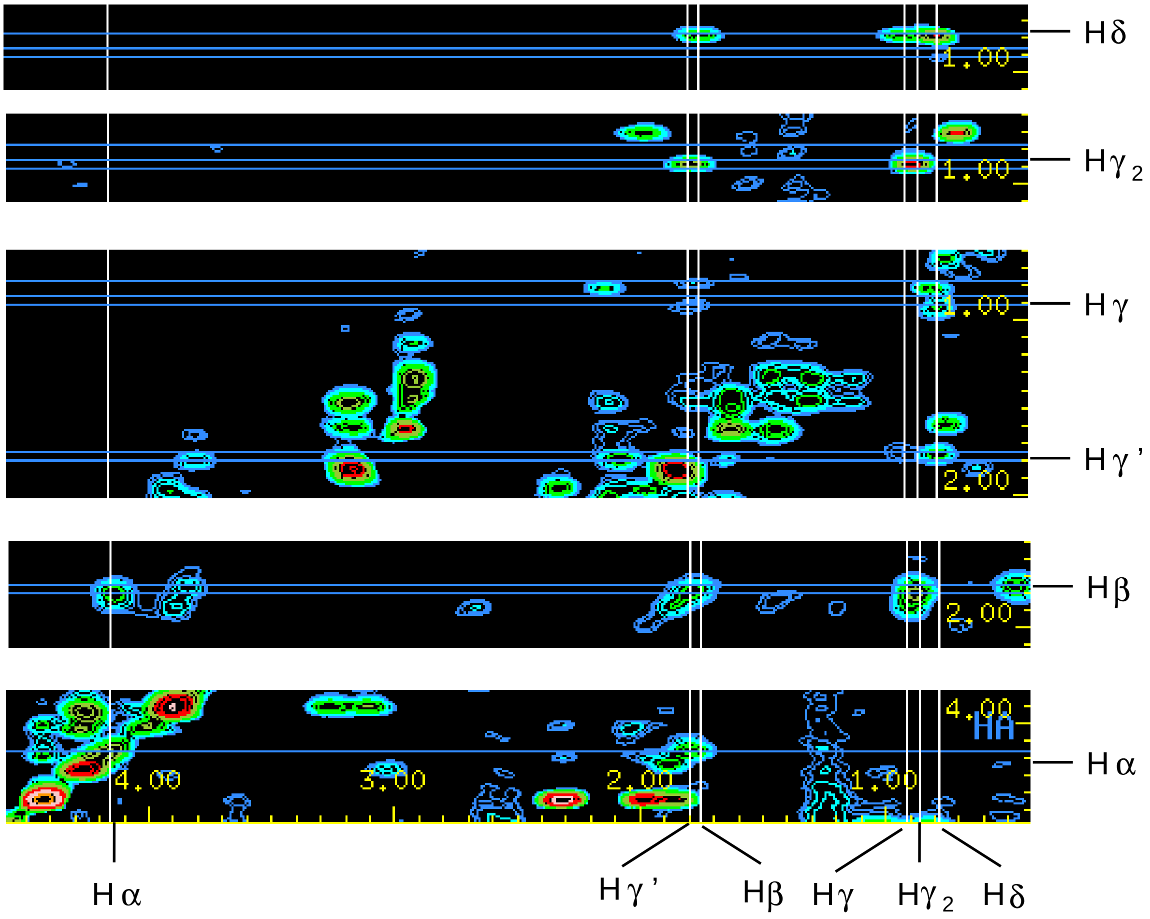


Figure 3.33: **Isoleucine result**
#1 from the list given in Appendix A, showing the peaks in the 3D COSY spectrum. The strips are taken from planes along the ^{13}C axis. In order, from bottom to top, they are: the Cα plane (62.9 ppm), the Cβ plane (38.3 ppm), the Cγ^1 plane (27.9 ppm), the Cγ^2 plane (17.9 ppm) and the Cδ plane (13.9 ppm). The lines indicate the F1 and F2 frequencies of the chemical shift positions of the individual spins, ie. Hα=4.18, Hβ=1.82, Hγ^1=0.93, H$\gamma^{1\prime}$=1.77, Hγ^2=0.88, Hδ=0.80.

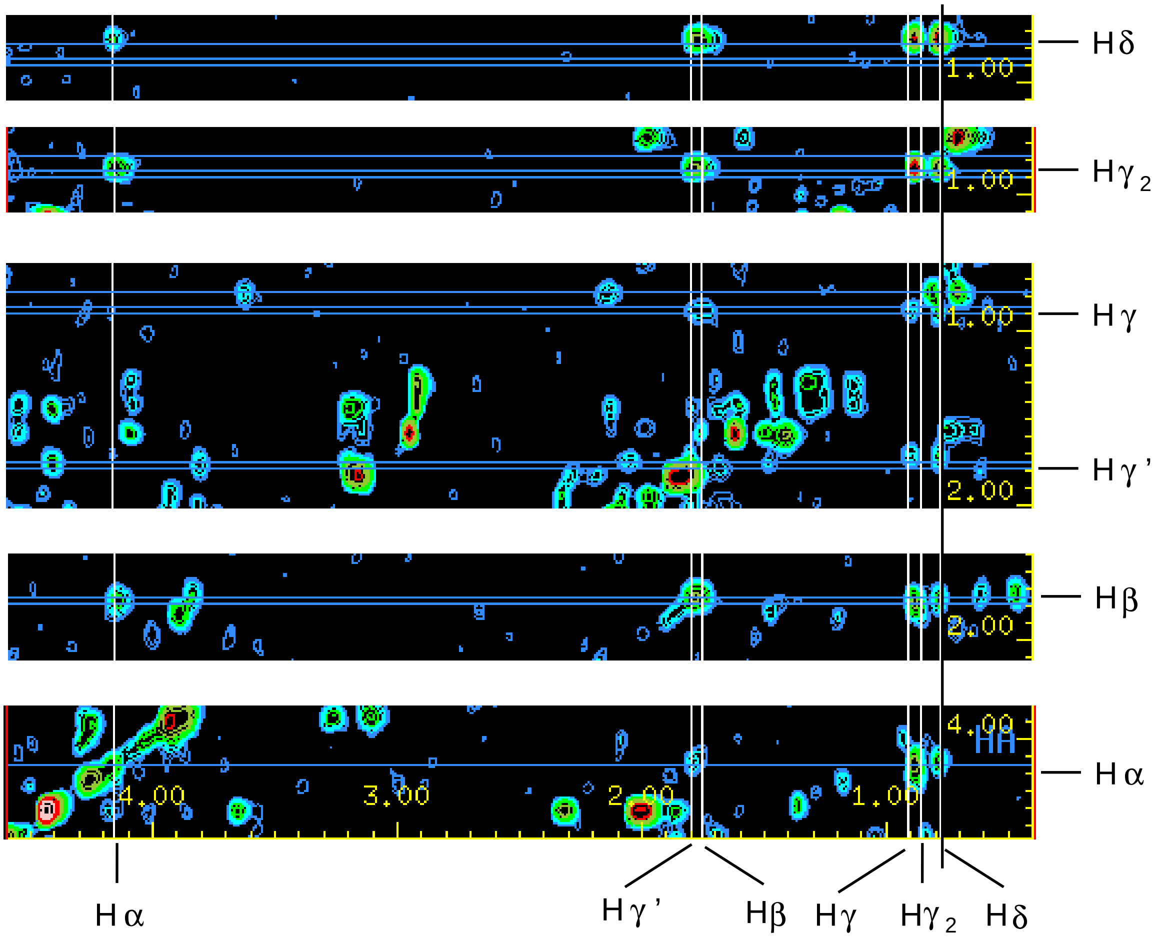


Figure 3.34: **Isoleucine result**

#1 from the list given in Appendix A, showing the peaks in the 3D TOCSY spectrum. The strips are taken from planes along the ^{13}C axis. In order, from bottom to top, they are: the Cα plane (62.9 ppm), the Cβ plane (38.3 ppm), the Cγ^1 plane (27.9 ppm), the Cγ^2 plane (17.9 ppm) and the Cδ plane (13.9 ppm). The lines indicate the F1 and F2 frequencies of the chemical shift positions of the individual spins, ie. Hα=4.18, Hβ=1.82, Hγ^1=0.93, H$\gamma^{1\prime}$=1.77, Hγ^2=0.88, Hδ=0.80.

of these complicating factors, the program has managed to locate the correct peaks.

3.5.2 Evaluating Backbone Spectra

The feasibility of pattern searches in backbone spectra was tested with a set of HNCO, HN(CA)CO, CBCANH, CBCA(CO)NH, HNCA and HN(CO)CA spectra. They are intended to constitute a minimal data set for reliable assignment; they may be augmented by other types of spectra if required. Our spectra were recorded under very different conditions, ie. on different NMR spectrometers with different field strengths and at slightly different temperatures, with gaps of several months between some of the measurements. For these reasons, the chemical shifts of the amide protons varied substantially in a non-systematic manner.

The emphasis in the searches presented in this section was therefore put on compensating for the large chemical shift differences between the spectra, with the unavoidable consequence that some of the results contained incorrect chemical shifts.

Search patterns containing the following chemical shifts were constructed:

- N_i, NH_i
- $C\alpha_i$, $C\beta_i$, CO_i,
- $C\alpha_{i-1}$, $C\beta_{i-1}$, CO_{i-1}

where i is the current residue in the sequence, and $i-1$ its predecessor. These patterns allow connected pairs of residues in the protein sequence to be recognised.

A special feature of glycine is that it shows very characteristic Cα chemical shifts. Also, Cβ peaks do *not* show up in the HNCA and HN(CO)CA spectra. Hence, glycine Cα peaks can be identified uniquely in these spectra.

This fact was used in constructing three patterns: one which searched for GX (glycine/non-glycine) pairs, one which searched for for XG (non-glycine/glycine) pairs, and one which searched for XX (non-glycine/non-glycine) pairs. X is a pseudo amino acid, with very broad search ranges, but its Cα chemical shift range does not extend down into the glycine region. In these pairs, the first letter corresponds to residue $i-1$ in the above chemical shift terminology, the second letter to residue i.

The XX pattern (see Figure 3.38) used information from the HNCO, HN(CA)CO, CBCANH and CBCA(CO)NH spectra only. The GX (see Figure 3.40) and XG (see Figure 3.39) pattern used all these spectra and in addition, the HNCA spectrum. The negative "expected" peaks in the HNCO, CBCANH and HNCA spectra (shown as dashed circles in the figures) are used by the

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	256
Sub-block (data pts)	16	4	16
Sweep width (ppm)	8	80	8
Offset (ppm)	-1.48	4.5	-1.48

Results list processing params:

Ovrlp-clus:	i=0	j=4	Δ_t=0.15
Sym:	Active.		
Sec-clus:	N_l=5	Δ_l=0.15	Δ_u=25.00
Fract-score:	ϵ_{low}=0.001		
Exs-penal:	M_x=4		
Too-far:	n_1=0	n_2=4	Δ_u=1.10
Std-dev:	Active.		
Thresh-count:	Active.		
Excess-peaks:	$P_{expected}$=5	r_l=2500	
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Cα	Hα	Cβ	Hβ
47.6→59.9	3.00→6.00	34.5→47.7	0.15→2.45

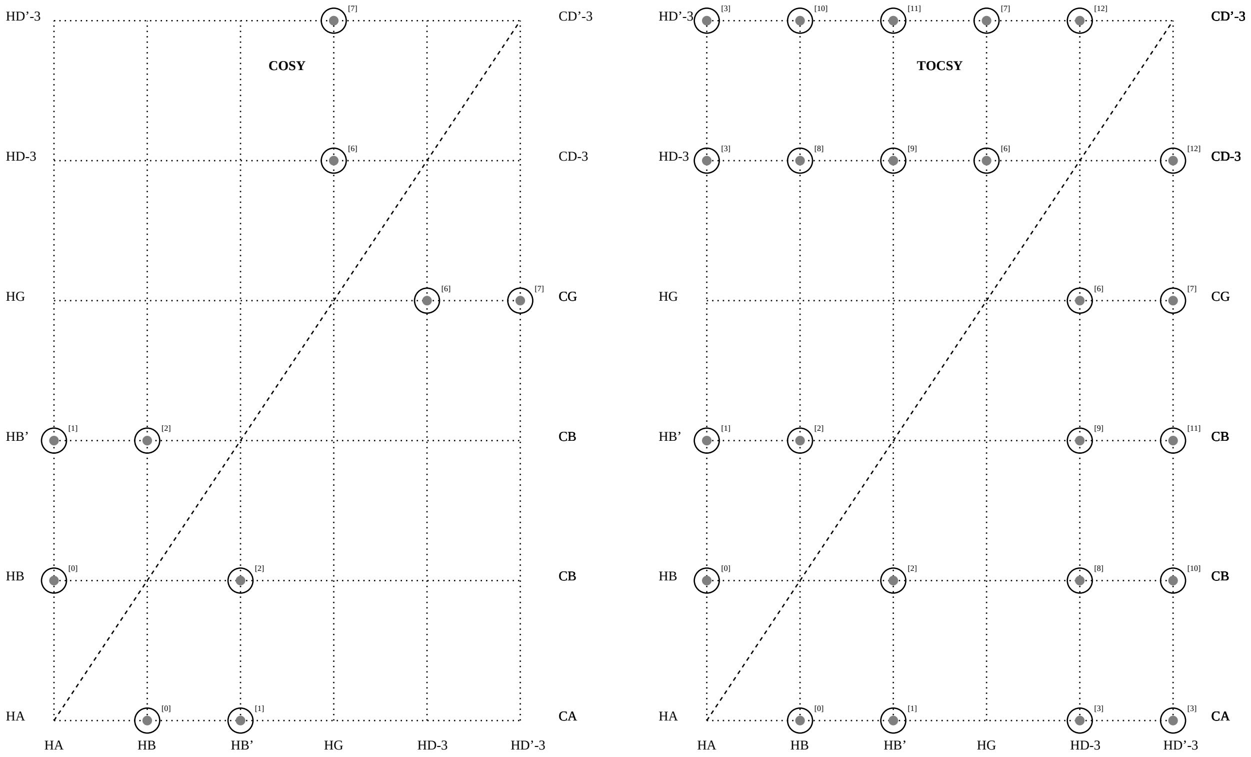


Figure 3.35: **Leucine pattern.**

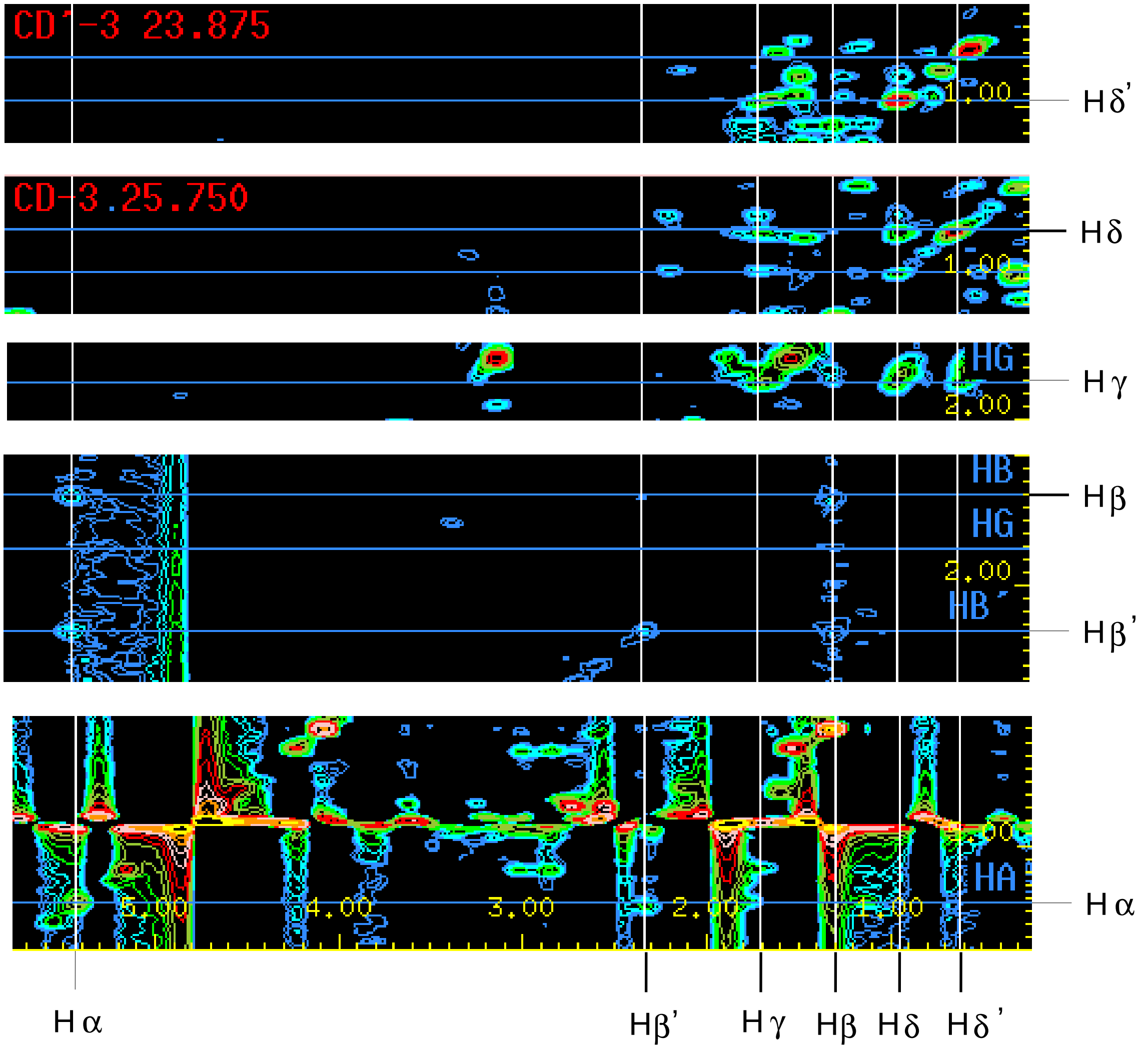


Figure 3.36: **Leucine result**
#8 from the list given in Appendix A, showing the peaks in the 3D COSY spectrum. The strips are taken from planes along the ^{13}C axis. In order, from bottom to top, they are: the Cα plane (53.3 ppm), the Cβ plane (46.1 ppm), the Cγ plane (26.7 ppm), the Cδ plane (25.8 ppm) and the Cδ' plane (23.9 ppm). The lines indicate the F1 and F2 frequencies of the chemical shift positions of the individual spins, ie. Hα=5.43, Hβ=1.16, Hβ'=1.80, Hγ=1.18, Hδ=0.29, Hδ'=0.55.

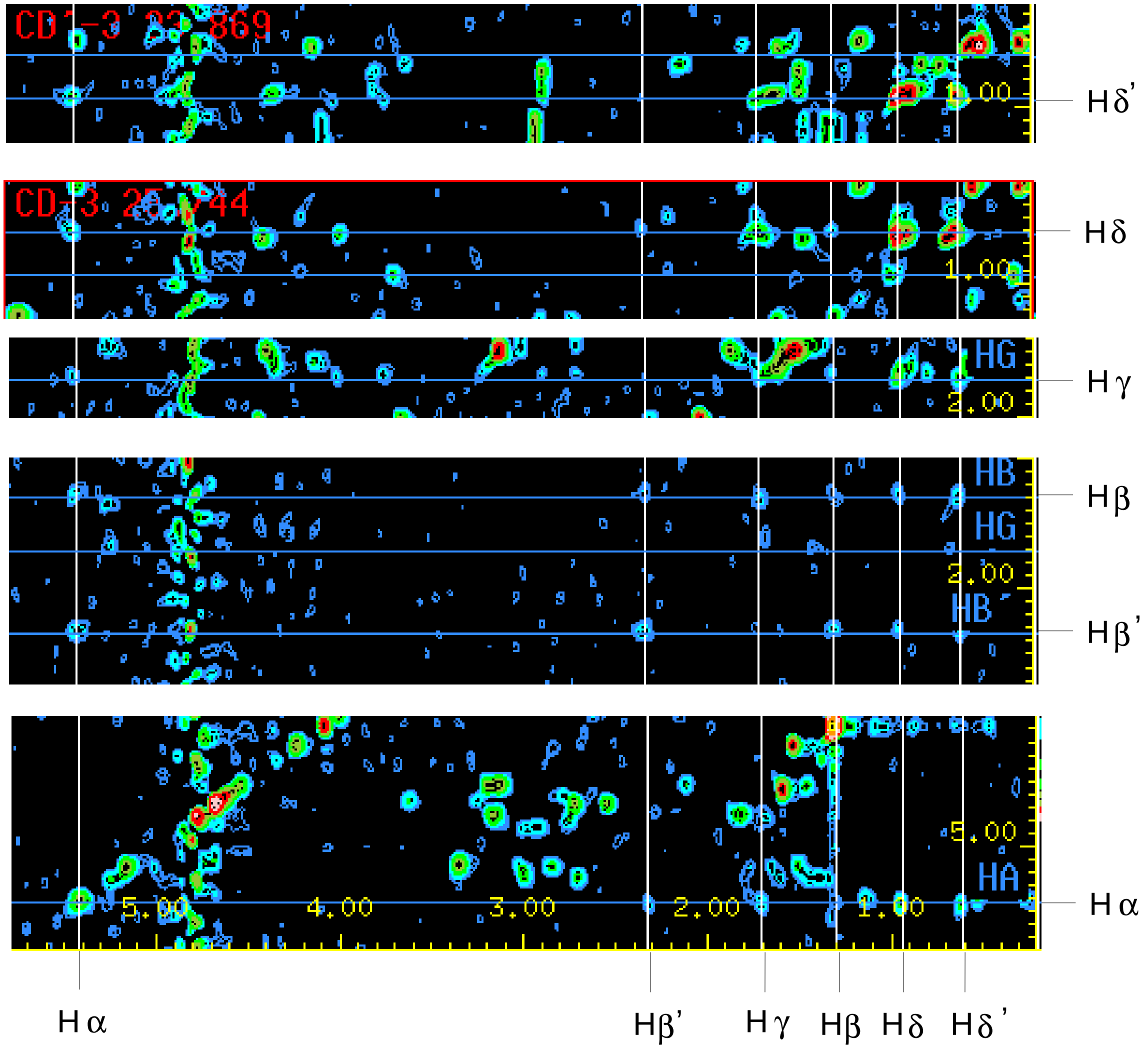


Figure 3.37: **Leucine result**
#8 from the list given in Appendix A, showing the peaks in the 3D TOCSY spectrum. The strips are taken from planes along the ^{13}C axis. In order, from bottom to top, they are: the Cα plane (53.3 ppm), the Cβ plane (46.1 ppm), the Cγ plane (26.7 ppm), the Cδ plane (25.8 ppm) and the Cδ' plane (23.9 ppm). The lines indicate the F1 and F2 frequencies of the chemical shift positions of the individual spins, ie. Hα=5.43, Hβ=1.16, Hβ'=1.80, Hγ=1.18, Hδ=0.29, Hδ'=0.55.

forbidden peaks algorithm to penalise results with peaks at those points. However, the negative "expected" peaks in the CBCA(CO)NH spectrum are there because these spectra really do exhibit negative peaks for Cα_i and Cβi.

The results lists produced by these patterns can be found in Appendix B; below is a summary of these results.

XG
Number of results expected (from the sequence): 8.
Actual results count: 7
Number of correct results found: 7.

The expected number of results of this kind is 8. Since the glycine missed by the program was not manually assignable anyway, this can be considered a very satisfactory outcome. Similar experiments without the HNCA spectrum (unpublished results) successfully found the same results set, but in addition, found several non-glycine containing spin systems, due to the poor discrimination between glycine and the other residues in the CBCANH and the CBCA(CO)NH spectra.

GX
Number of results expected (from the sequence): 8.
Actual results count: 5
Number of correct results found: 5.

The results list is good, in that all results are completely correct, but three results are missing. Of these, one could not be manually assigned. The others could not be found due to discrepancies between the spectra. Excluding the HNCA spectrum allows all seven manually assignable results to be found by the program (unpublished results), but many incorrect results are found also.

XX
Number of results expected (from the sequence): 98.
Actual results count: 99
Number of correct results found: 80.

The actual count of identifiably sensible results is 80, ie. 82% of the expected figure. Of the others, 8 involved a glycine, 8 were completely unidentifiable, and the remainder contained mixtures of spin systems which shared multiple common chemical shifts.

A "sensible" result in this context is one for which the N_i and NH_i chemical shifts are correct, as well as at least 50% of the remaining chemical shifts. In 25 of the results, the CO_i and CO_{i-1} chemical shifts could not be ascertained, due to differences in the conditions under which the HNCO and HN(CA)CO spectra were taken.

Spectrum size parameters:

	HNCO			CBCANH		
	w1	w2	w3	w1	w2	w3
Size (data pts)	128	128	512	128	128	512
Sub-block (data pts)	16	4	16	4	4	16
Sweep width (ppm)	39.460	19.878	6.847	39.456	80.813	6.847
Offset (ppm)	98.870	167.561	4.810	98.131	5.765	4.810

Results list processing parameters:

Ovrlp-clus:	i=3	j=2	Δ_t=0.50
Sym:	Active.		
Sec-clus:	N_l=5	Δ_l=0.20	Δ_u=25.00
Exs-penal:	M_x=2		
Std-dev:	Active.		
Thresh-count:	Active.		
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Resid.	N	H	CO	Cα	Cβ
X	106.5→133.5	6.00→10.55	167.6→187.4	45.0→66.6	13.0→72.3

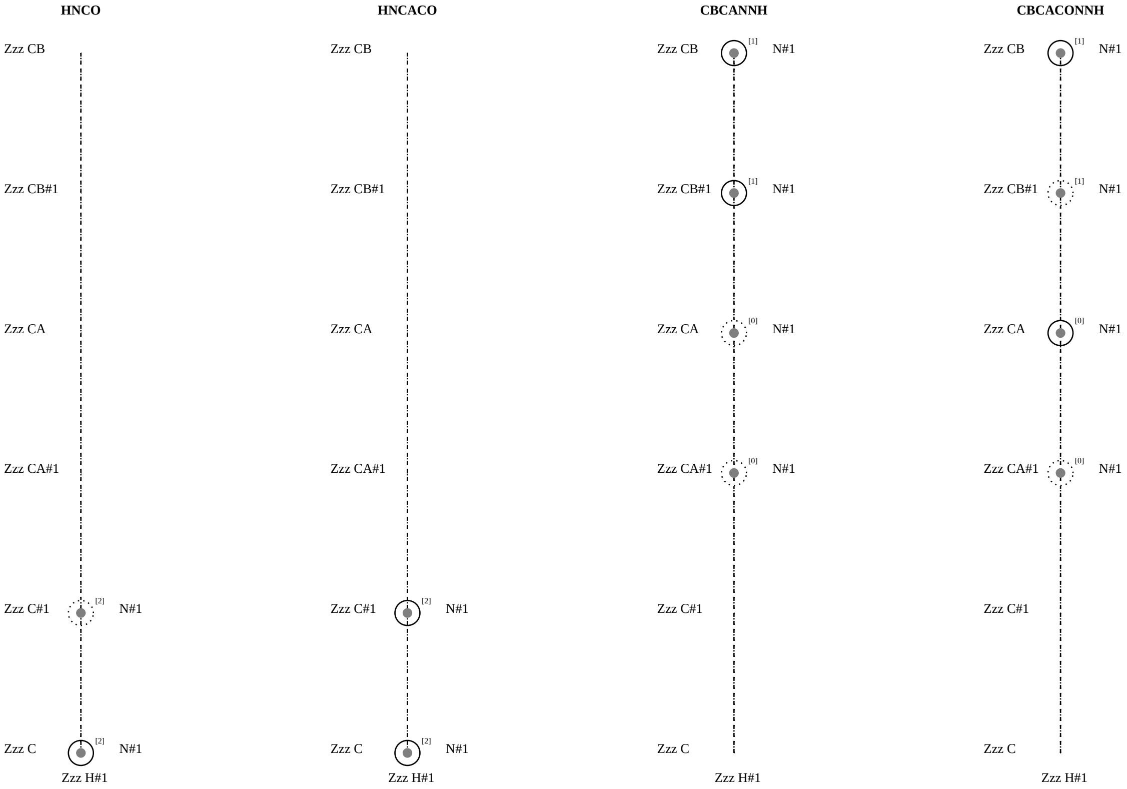


Figure 3.38: **XX pattern.**

Spectrum size parameters:

	HNCO			CBCANH			HNCA		
	w1	w2	w3	w1	w2	w3	w1	w2	w3
Size (data pts)	128	128	512	128	128	512	128	128	512
Sub-block (data pts)	16	4	16	4	4	16	16	4	16
Sweep width (ppm)	39.460	19.878	6.847	39.456	80.813	6.847	39.460	39.756	6.847
Offset (ppm)	98.870	167.561	4.810	98.131	5.765	4.810	100.170	36.322	4.810

Results list processing parameters:

Ovrlp-clus:	i=3	j=2	Δ_t=0.50
Sym:	Active.		
Sec-clus:	N_l=4	Δ_l=0.20	Δ_u=60.00
Exs-penal:	M_x=2		
Std-dev:	Active.		
Thresh-count:	Active.		
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Resid.	N	H	CO	Cα	Cβ
X	106.5→133.5	6.00→10.55	167.6→187.4	45.0→66.6	13.0→72.3
G	99.0→117.6	6.00→9.47	167.6→187.4	35.5→48.5	

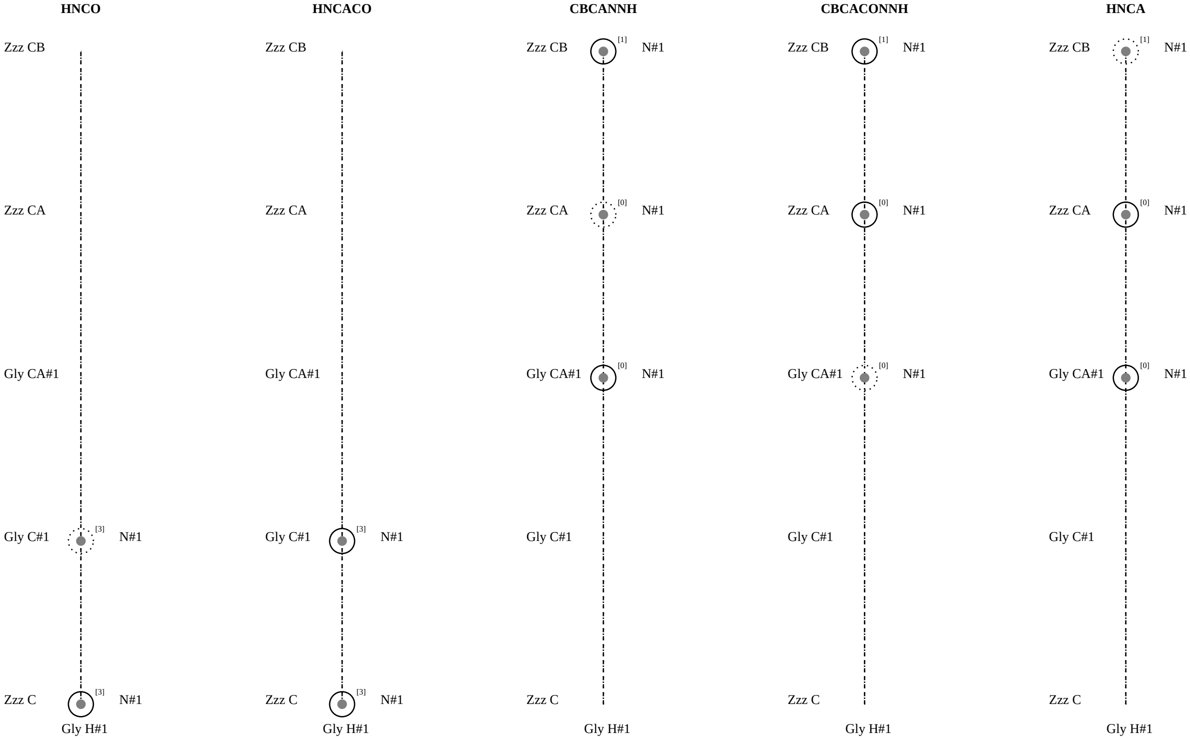


Figure 3.39: **XG pattern.**

Spectrum size parameters:

	HNCO			CBCANH			HNCA		
	w1	w2	w3	w1	w2	w3	w1	w2	w3
Size (data pts)	128	128	512	128	128	512	128	128	512
Sub-block (data pts)	16	4	16	4	4	16	16	4	16
Sweep width (ppm)	39.460	19.878	6.847	39.456	80.813	6.847	39.460	39.756	6.847
Offset (ppm)	98.870	167.561	4.810	98.131	5.765	4.810	100.170	36.322	4.810

Results list processing parameters:

Ovrlp-clus:	i=3	j=2	Δ_t=0.50
Sym:	Active.		
Sec-clus:	N_l=4	Δ_l=0.20	Δ_u=60.00
Exs-penal:	M_x=2		
Std-dev:	Active.		
Thresh-count:	Active.		
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Resid.	N	H	CO	Cα	Cβ
X	106.5→133.5	6.00→10.55	167.6→187.4	45.0→66.6	13.0→72.3
G	99.0→117.6	6.00→9.47	167.6→187.4	35.5→48.5	

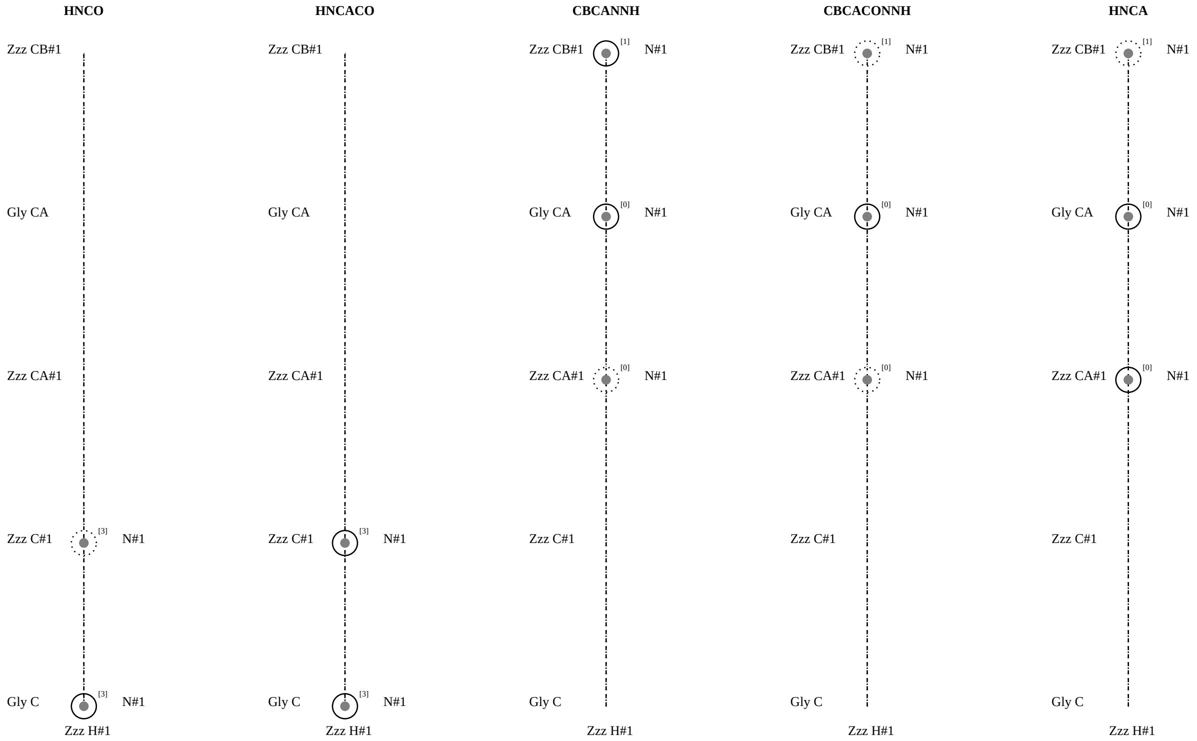


Figure 3.40: **GX pattern.**

3.6 Results Obtained with the RalGDS Ras Binding Domain, a Side Chain Case Study

The patterns used to find spin systems in the spectra of the N-terminal thioredoxin-like domain of protein disulphide isomerase should, in principal, work with other proteins as well. In order to test this hypothesis, the sidechain patterns for glycine, alanine, the AMX spin systems, serine, the GLX spin systems, threonine, valine and isoleucine were applied to a HCCH-COSY and -TOCSY spectrum pair from the RalGDS ras binding domain [6]. Both spectra were recorded at a temperature of 300K and at a pH of 6.5.

The domain consists of the last 127 residues at the c-terminal end of the RalGDS protein, a putative member of the MAP kinase signal transduction pathway ([63]).

Since there are no pre-existing assignments for comparison purposes, the judgements about the correctness of spin systems assigned by the program are more subjective than in Section 3.5. A summary of the full results set is shown in Figure 3.41. It can be seen that the centre of mass measure is always high, often 100% in fact, showing that the ordering of the results in the results list is a good indication of their reliability. Also, in the majority of cases, more than 60% of the expected residues were found. However, in a number of cases, the number of correct results found was small relative to the total number of results found. It would thus be reasonable to say that, as long as one uses only the highest scoring results, one has a very good chance of selecting valid assignments. The results of these pattern searches are discussed in detail below.

The results are tabulated in full in Appendix C.

Gly
Number of results expected (from the sequence): 3.
Actual results count: 3
Number of correct results found: 1.

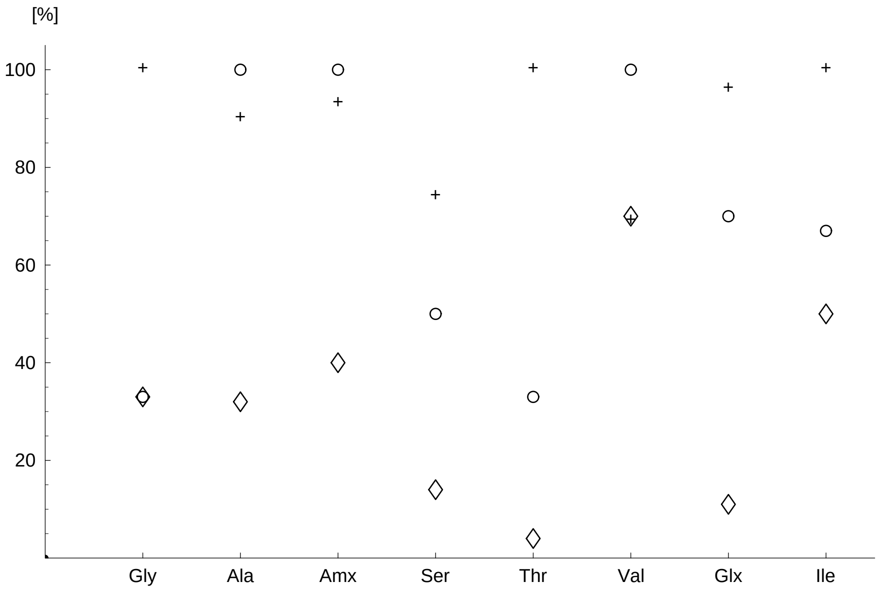
The pattern shown in Figure 3.25 (page 75) was not very successful with the new data, it actually found mainly $H\beta/H\beta'$ peaks from larger spin systems. It was modified slightly, by making the excess peak penalisation more vigorous.

Four excess penalisation lines were added, one for each expected peak, in each of the spectra. Excess peaks were searched for in the range 1 to 6.4 ppm, designed to find peaks coming from AMX, GLX, proline and leucine spin systems. Additional penalisation was applied in cases where peaks appeared on different lines with common chemical shifts.

This led to a final results set in which most of the incorrect results were so heavily penalised that the score sank to zero, and they were deleted. Of the

[6] Kindly provided by Peter Schmieder at the Forschungsinstitut für Molekularpharmakologie in Berlin.

Figure 3.41: **Summary of side chain results.**
Three measures are used to characterise the performance of the program with a given pattern, all normalised and expressed in percent, and indicated by three different symbols in the graph. The "o" symbol shows how many correctly found assignments there are compared to the number of assignments expected from the sequence. The "◇" symbol shows the number of correct assignments compared to the total number in the results list. The "+" symbol shows the centre of mass of the correct results, measured from the low-scoring end of the results list, and is hence a measure of the effectiveness of the results list penalisation algorithms. The larger this value, the more correct results appear at the high-scoring of the list.



results that remained, one looked reasonable, though it was not possible to be absolutely sure, since the peaks lay on the rather broad water line. The pattern for this experiment is shown in Figure 3.42.

Ala
Number of results expected (from the sequence): 5.
Actual results count: 19
Number of correct results found: 6.

The simplicity of the alanine pattern causes many false assignments, as seen in the N-terminal thioredoxin-like domain of protein disulphide isomerase example. Manual inspection of the results showed six promising candidate patterns. It is possible that some alanines produce more than one set of chemical shifts, due to minor conformations caused by proline cis-trans isomerisation. The pattern for this experiment is shown in Figure 3.43.

AMX
Number of results expected (from the sequence): 26.
Actual results count: 87
Number of correct results found: 35.

A manual inspection of the AMX results indicated that the correct results were the highest scoring ones, but that other spin systems, such as the GLX spin system, had also found their way into the results list. As with the alanine results, there were more plausible-looking patterns than would be expected from the sequence. The pattern for this experiment is shown in Figure 3.44.

Ser
Number of results expected (from the sequence): 6.
Actual results count: 22
Number of correct results found: 3.

Serine is difficult to identify from the data in the HCCH-COSY and -TOCSY experiments alone because the chemical shift ranges for Hα and Hβ, and for Cα and Cβ overlap significantly, and because the spin system also overlaps with the threonine spin system. In spite of this, the program managed to find three convincing patterns. It also found many patterns in which Hα and Hβ were degenerate; it is possible that some of these are correct. The pattern for this experiment is shown in Figure 3.45.

Thr
Number of results expected (from the sequence): 3.
Actual results count: 24
Number of correct results found: 1.

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	512
Sub-block (data pts)	32	16	128
Sweep width (ppm)	6.665	82.835	6.489
Offset (ppm)	-0.227	3.999	-0.094

Results list processing params:

Sym:	Active.		
Exs-penal:	M_x=1		
Too-far:	n_1=0	n_2=1	Δ_u=1.70
Thresh-count:	Active.		
Excess-peaks:	$P_{expected}$=2	r_l=300	
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Cα	Hα
38.5→48.5	3.00→4.36

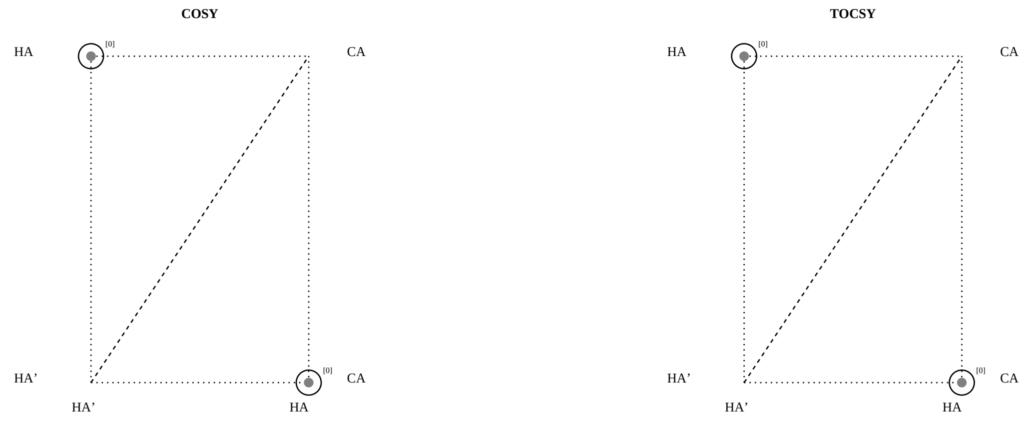


Figure 3.42: **Glycine pattern.**

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	512
Sub-block (data pts)	32	16	128
Sweep width (ppm)	6.665	82.835	6.489
Offset (ppm)	-0.227	3.999	-0.094

Results list processing params:

Sym:	Active.		
Sec-clus:	N_l=3	Δ_l=0.10	Δ_u=70.00
Fract-score:	ϵ_{low}=0.001		
Thresh-count:	Active.		
Excess-peaks:	$P_{expected}$=2	r_l=500	
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Cα	Hα	Cβ	Hβ
45.8→57.8	3.17→5.80	13.0→26.0	-0.40→1.92

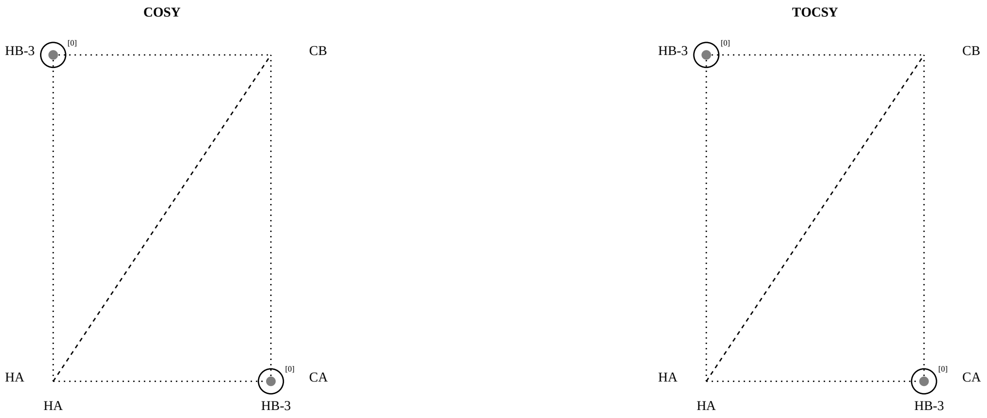


Figure 3.43: **Alanine pattern.**

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	512
Sub-block (data pts)	32	16	128
Sweep width (ppm)	6.665	82.835	6.489
Offset (ppm)	-0.227	3.999	-0.094

Results list processing parameters:

Ovrlp-clus:	i=1	j=4	Δ_t=0.15
Sym:	Active.		
Sec-clus:	N_l=3	Δ_l=0.20	Δ_u=70.00
Fract-score:	ϵ_{low}=0.001		
Too-far:	n_1=1	n_2=4	Δ_u=1.10
Thresh-count:	Active.		
Excess-peaks:	$P_{expected}$=3	r_l=500	
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Cα	Hα	Cβ	Hβ
46.5→64.0	2.87→6.20	23.4→45.0	1.20→4.30

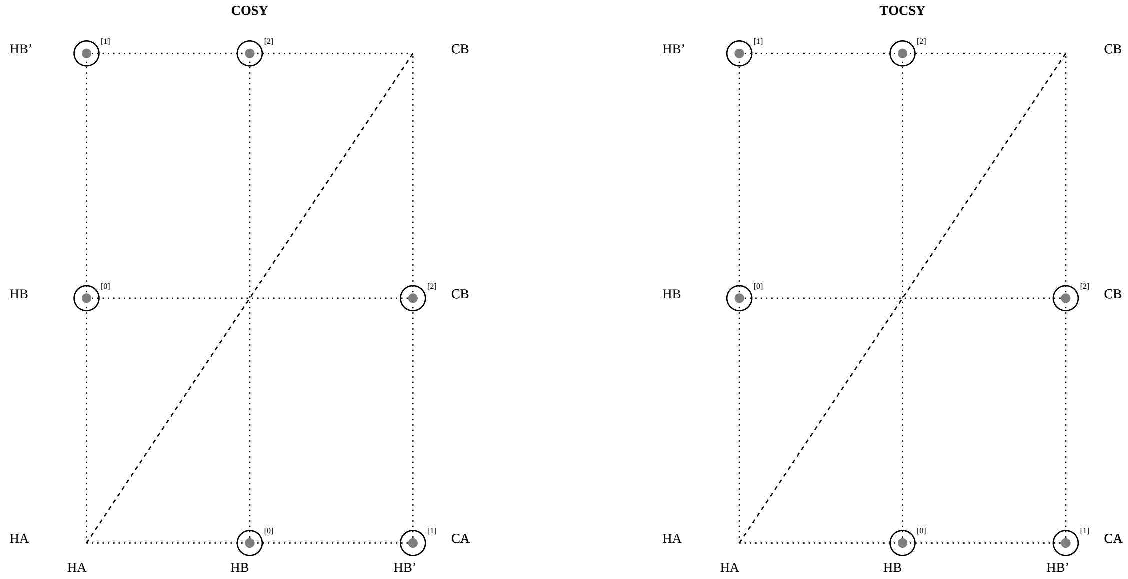


Figure 3.44: **AMX pattern.**

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	512
Sub-block (data pts)	32	16	128
Sweep width (ppm)	6.665	82.835	6.489
Offset (ppm)	-0.227	3.999	-0.094

Results list processing parameters:

Ovrlp-clus:	i=1	j=4	Δ_t=0.15
Sym:	Active.		
Sec-clus:	N_l=3	Δ_l=0.20	Δ_u=70.00
Fract-score:	ϵ_{low}=0.001		
Exs-penal:	M_x=2		
Too-far:	n_1=1	n_2=4	Δ_u=1.10
Thresh-count:	Active.		
Excess-peaks:	$P_{expected}$=3	r_l=500	
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Cα	Hα	Cβ	Hβ
47.0→63.0	3.50→6.10	55.0→70.0	2.87→4.60

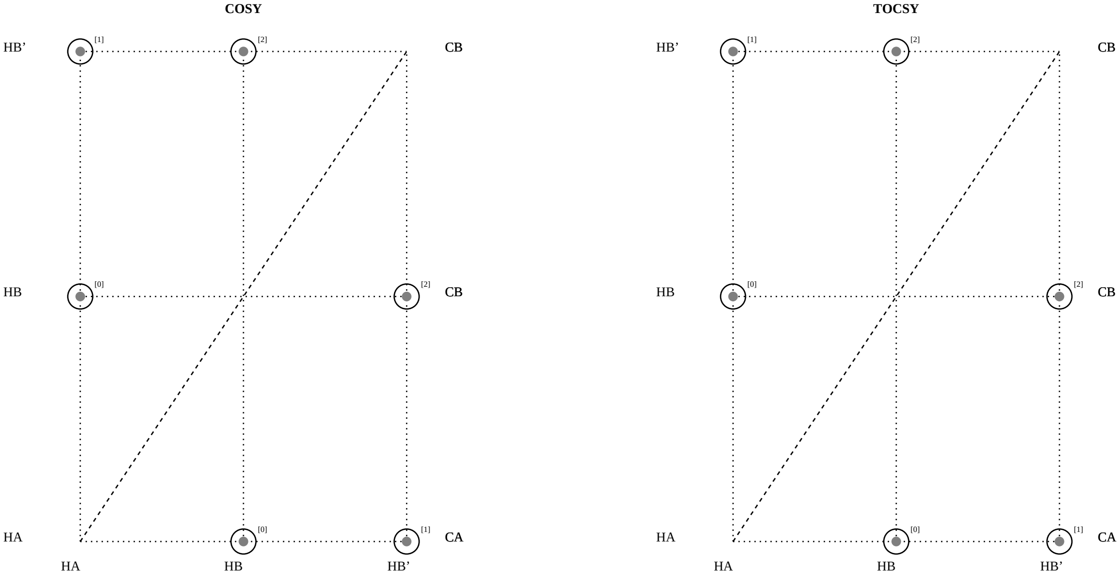


Figure 3.45: **Serine pattern.**

Only one of the threonine patterns found by the program looked sensible; this was the top-scoring result. Some of the remaining patterns *may* have been correct, because they contained very small peaks in the correct places, but these peaks were right down in the noise, so these results could not be classified as correct with any degree of confidence. The pattern for this experiment is shown in Figure 3.46.

Val
Number of results expected (from the sequence): 6.
Actual results count: 10
Number of correct results found: 7.

In the HCCH-COSY and -TOCSY spectra, the valine patterns showed up especially clearly, and they were easily identified by the program. However, it seems that one of the patterns found was a false positive. The pattern for this experiment is shown in Figure 3.47.

GLX
Number of results expected (from the sequence): 10.
Actual results count: 66
Number of correct results found: 6.

This pattern is difficult for the program to identify accurately, because so many of the chemical shift ranges overlap. Nevertheless, the sensible-looking results appeared at the high-scoring end of the results list, indicating that the pattern design was correct. In addition to the six fully correct patterns, the program also found two other patterns which appear substantially correct. #63 (see Appendix C, page 145) may even be completely correct; it seems in this case that Cβ and Cγ are degenerate. #59 (Appendix C) shows a typical problem for the GLX pattern - the program has confused Hβ and the Hγ peaks - it has correctly assigned the peaks to a GLX spin system, but it has assigned them to the wrong nuclei. This is an especially difficult case, because from a visual inspection, it is apparent that Hβ' and Hγ' are degenerate. The pattern for this experiment is shown in Figure 3.48.

Ile
Number of results expected (from the sequence): 6.
Actual results count: 8
Number of correct results found: 4.

In spite of weak peaks between Hβ and Hγ, the program was able to find four of the expected eight isoleucines very convincingly; it is likely that some of the other patterns found were partial assignments for isoleucines. The pattern for

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	512
Sub-block (data pts)	32	16	128
Sweep width (ppm)	6.665	82.835	6.489
Offset (ppm)	-0.227	3.999	-0.094

Results list processing parameters:

Ovrlp-clus:	i=1	j=0	Δ_t=0.15
Sym:	Active.		
Sec-clus:	N_l=4	Δ_l=0.30	Δ_u=70.00
Fract-score:	ϵ_{low}=0.001		
Exs-penal:	M_x=2		
Thresh-count:	Active.		
Excess-peaks:	$P_{expected}$=3	r_l=1000	
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Cα	Hα	Cβ	Hβ
54.6→67.6	3.00→6.00	62.5→74.5	3.40→5.00

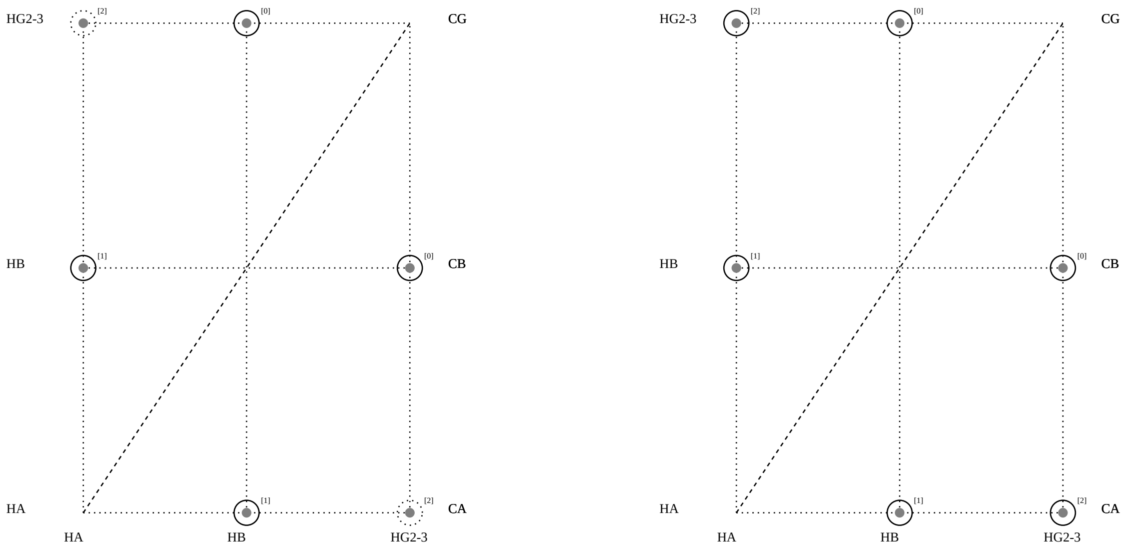


Figure 3.46: **Threonine pattern.**

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	512
Sub-block (data pts)	32	16	128
Sweep width (ppm)	6.665	82.835	6.489
Offset (ppm)	-0.227	3.999	-0.094

Results list processing params:

Ovrlp-clus:	i=4	j=6	Δ_t=0.15
Sym:	Active.		
Sec-clus:	N_l=5	Δ_l=0.30	Δ_u=70.00
Fract-score:	ϵ_{low}=0.001		
Exs-penal:	M_x=2		
Thresh-count:	Active.		
Excess-peaks:	$P_{expected}$=3	r_l=2500	
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Cα	Hα	Cβ	Hβ
49.0→68.2	1.50→6.00	26.8→36.4	0.81→2.75

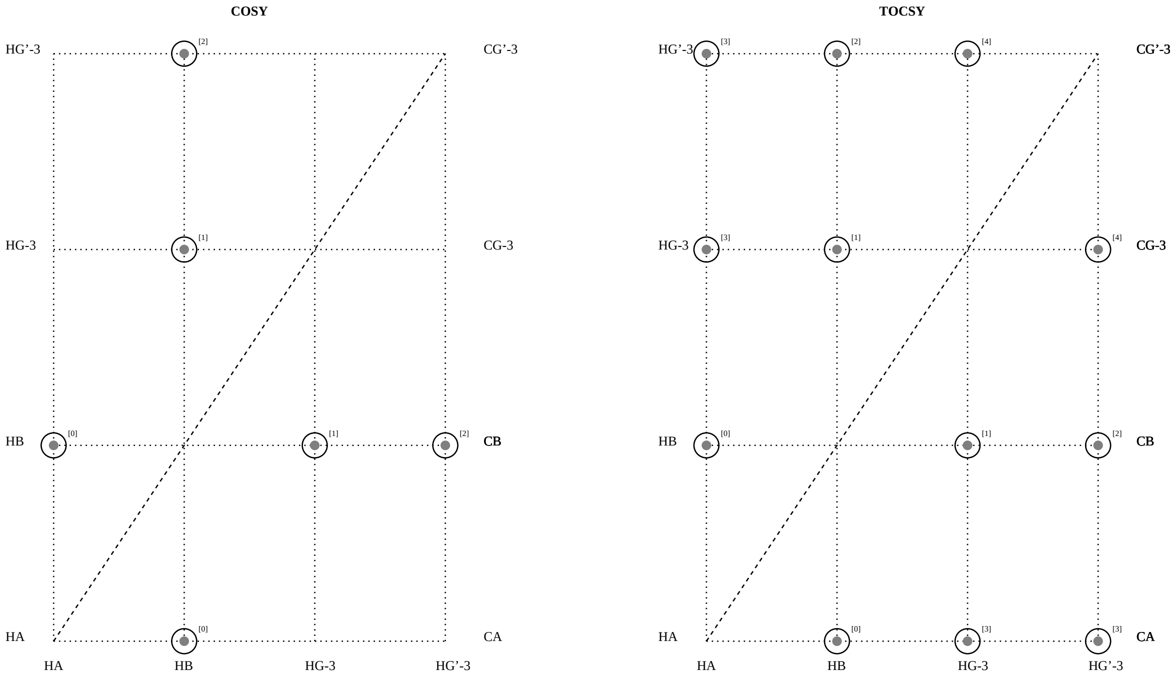


Figure 3.47: **Valine pattern.**

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	512
Sub-block (data pts)	32	16	128
Sweep width (ppm)	6.665	82.835	6.489
Offset (ppm)	-0.227	3.999	-0.094

Results list processing parameters:

Ovrlp-clus:	i=1	j=4	Δ_t=0.15
Sym:	Active.		
Sec-clus:	N_l=5	Δ_l=0.30	Δ_u=70.00
Fract-score:	ϵ_{low}=0.001		
Exs-penal:	M_x=2		
Too-far:	n_1=1	n_2=4	Δ_u=1.10
Thresh-count:	Active.		
Excess-peaks:	$P_{expected}$=5	r_l=500	
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Cα	Hα	Cβ	Hβ
43.4→66.0	2.60→5.60	23.0→36.5	0.80→2.81

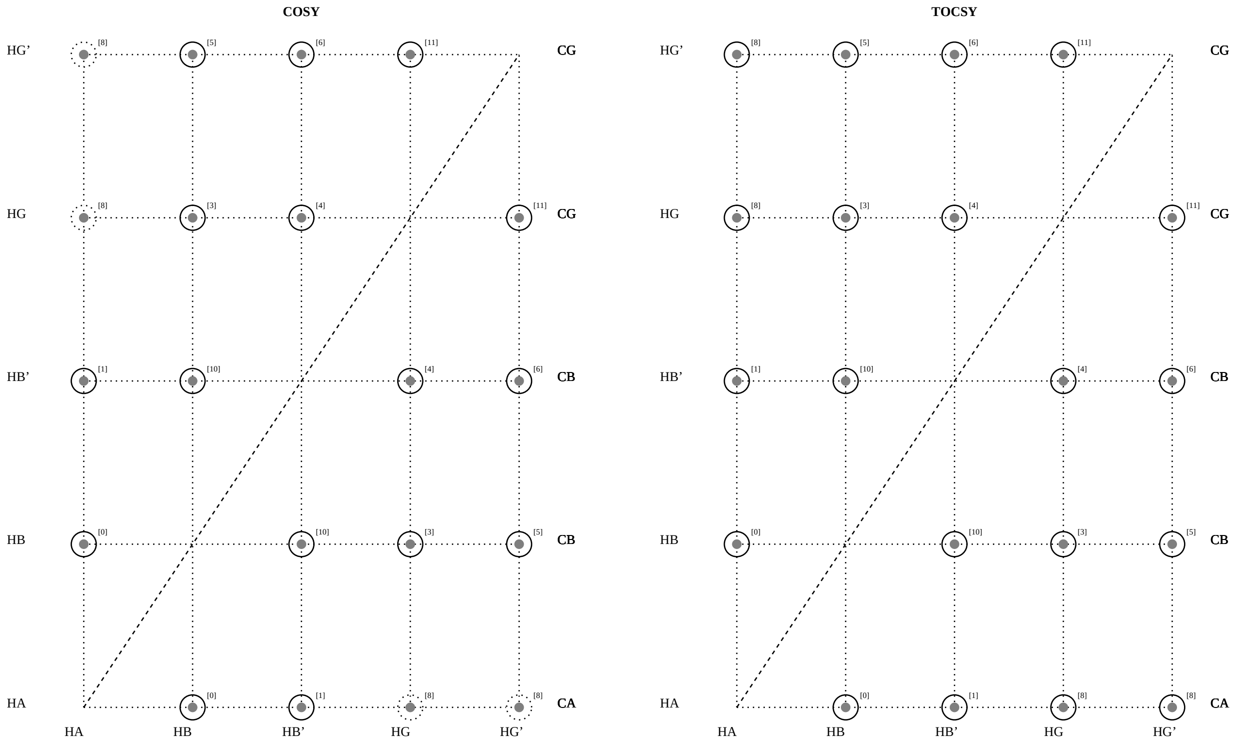


Figure 3.48: **GLX pattern.**

this experiment is shown in Figure 3.49.

Spectrum size parameters for both HCCH-COSY and -TOCSY:

	w1	w2	w3
Size (data pts)	512	256	512
Sub-block (data pts)	32	16	128
Sweep width (ppm)	6.665	82.835	6.489
Offset (ppm)	-0.227	3.999	-0.094

Results list processing params:

Sym:	Active.		
Sec-clus:	N_l=6	Δ_l=0.30	Δ_u=70.00
Fract-score:	ϵ_{low}=0.001		
Exs-penal:	M_x=4		
Too-far:	n_1=6	n_2=8	Δ_u=1.20
Thresh-count:	Active.		
Excess-peaks:	$P_{expected}$=5	r_l=2500	
Uninst-chm-shft:	Active.		
Uninst-resp:	Active.		

Chemical shift ranges:

Cα	Hα	Cβ	Hβ
49.0→66.6	3.06→6.00	30.1→42.6	0.73→2.70

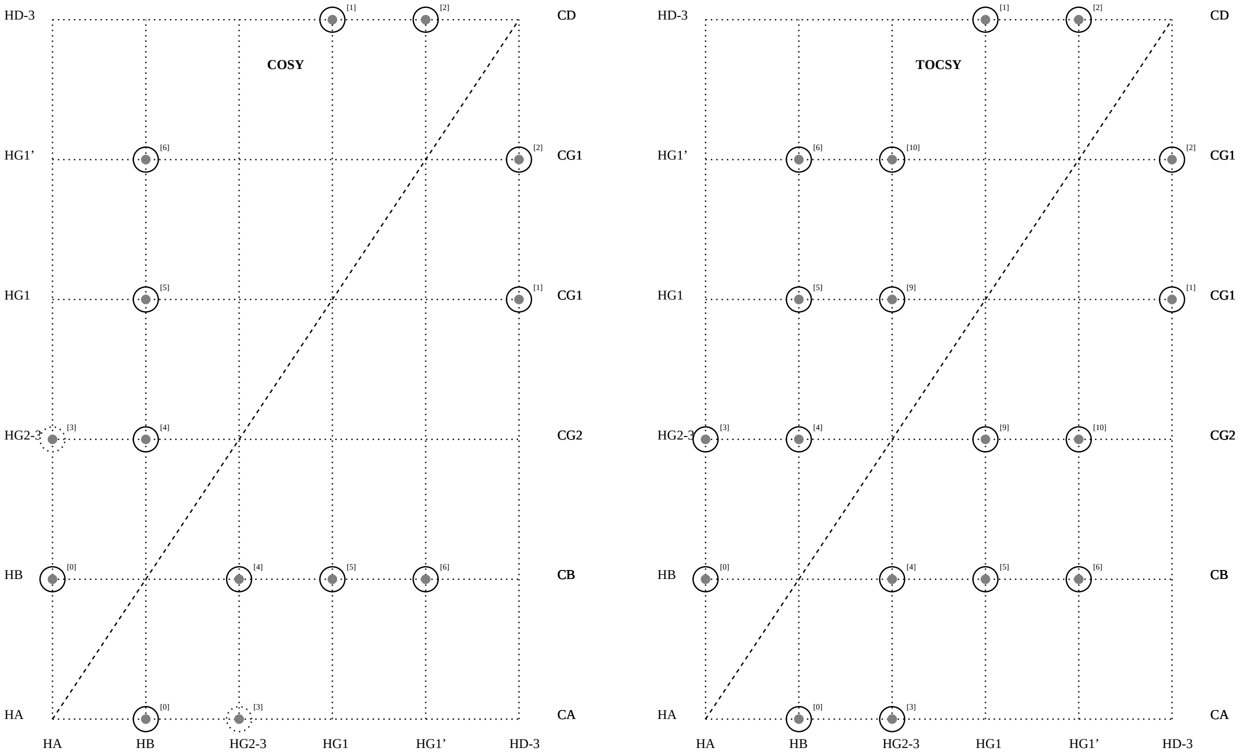


Figure 3.49: **Isoleucine pattern.**