A Constraint-based Assignment System for Automating Long Side Chain Assignments In Protein 2D NMR Spectra*

Scott Leishman[1,2], Peter MD Gray[1] and John E Fothergill[2]
[1]Department of Computing Science, [2]Department of Molecular & Cell Biology
University of Aberdeen, Aberdeen, Scotland, AB9 2UB
scott@csd.abdn.ac.uk

Abstract

The sequential assignment of protein 2D NMR data has been tackled by many automated and semi-automated systems. One area that these systems have not tackled is the searching of the TOCSY spectrum looking for cross peaks and chemical shift values for hydrogen nuclei that are at the end of long side chains. This paper describes our system for solving this problem using constraint logic programming and compares our constraint satisfaction algorithm to a standard backtracking version.

Introduction

Even with ten years of computational assistance, determining the structure of a protein by 2D NMR is still a time consuming task. There have been many systems that have concentrated on different aspects of the assignment and structure building process (e.g (Billeter, Basus, & Kuntz 1988), (Cieslar, Clore, & Gronenborn 1988), (Edwards *et al.* 1993)). While these systems have had varying degrees of success, one area of the sequential assignment stages has not been considered. This is the task of searching for cross peaks associated with long side chains.

The TOCSY Assignment Module (TAM) is a research prototype which carries out this searching. It makes use of a constraint manager implemented in CHIP and the P/FDM object-oriented database which stores the experimental spectra, preliminary assignments and the final results. TAM is an integrated part of ASSASSIN (Leishman, Gray, & Fothergill 1994) a constraint based assignment system for the structural assignment of 2D protein NMR.

Determining a protein structure by NMR (Wüthrich 1986) is a long and complicated task that is manually intensive. The process can be split into two stages. Firstly, the *backbone assignment* concentrates on identifying cross peaks associated with hydrogen nuclei attached to the appropriate amide, alpha and beta carbon atoms. The end result will identify elements of secondary structure, but not the overall fold. Secondly, the *structural assignment* stage builds on the results from the backbone assignment. It aims to carry out a detailed analysis of the NOESY (through space) spectrum and to discover distance constraints and torsion angle restraints in order to generate full 3-D structures that satisfy these requirements.

Background

2D Nuclear Magnetic Resonance (NMR) is an experimental method for determining the three-dimensional structure of small to medium sized proteins in solution. A 2D NMR experiment produces a large 2D data spectrum derived from the Fourier transform of experimental results, consisting of hundreds of small peaks, Figure 1 (Lian *et al.* 1992). In order to build the corresponding protein structure most of these peaks must be associated with a corresponding pair of hydrogen nuclei. This process of association, or *assignment*, is difficult because the spectra are susceptible to considerable noise. This noise is generated by spectrometer instabilities, thermal noise and instrumental limitations and many of the peaks come very close to one another.

Figure 2 shows there are two distinct patterns: the main diagonal running from top right to bottom left and the other peaks that surround the main diagonal, called cross peaks. Each hydrogen nucleus comes to magnetic resonance at some point along the *main diagonal.* The position of resonance is called the Chemical Shift and is normally given in parts per million (ppm) from a reference compound. The *cross peaks* show an interaction between two hydrogen nuclei. The hydrogen nuclei that are involved can be found by drawing lines parallel to the axes, from the cross peaks to the main diagonal. The cross peaks should be symmetri-

* This work has been supported by a Prize Studentship from The Wellcome Trust.

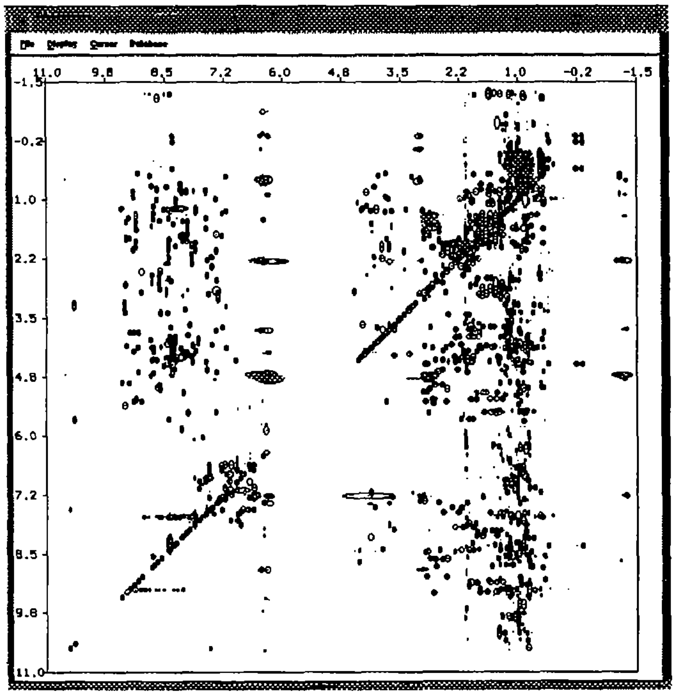

Figure 1: TOCSY spectrum of the IgG binding domain III from Protein G.

cal about the main diagonal, but in practice this is not true because of differences in resolution in the two axes and noise artifacts.

Different regions of the spectrum correspond to different types of side chain atom groups that contain hydrogen nuclei. Hydrogen nuclei along a side chain are classified according to the carbon atom they are attached to in the amino acid residue. Greek letters are used in sequence starting from the peptide carbonyl group, e.g α, β, γ, δ, ϵ, while the hydrogen nuclei attached to the amide nitrogen is denoted by an N (IUPAC-IUB 1970). Figure 2 shows along the axes of the spectrum the areas where the different hydrogen nuclei come to resonance. It is clear that all the hydrogen nuclei that are at the end of amino acid residues with long side chains come to resonance in the top right hand corner making the top right corner very crowded.

Through bond spectra, such as COSY, TOCSY and HOHAHA show interactions between the hydrogen nuclei in a single residue. Different residue types give rise to different patterns of cross peaks according to their chemical structure. The exact location of the cross peaks which form a residue pattern, or *spin-system*, cannot be predicted, but it is known that each cross peak should fall in a general region of the spectrum. Figure 3 shows the six general regions for a threonine residue on each side of the main diagonal. It is these spin-systems that are very important when trying to search for cross peaks associated with resonances at the end of side chains. In the rest of this paper the expected region where a cross peak that is part of a spin-system should appear will be called a Residue Peak (*rpeak*).

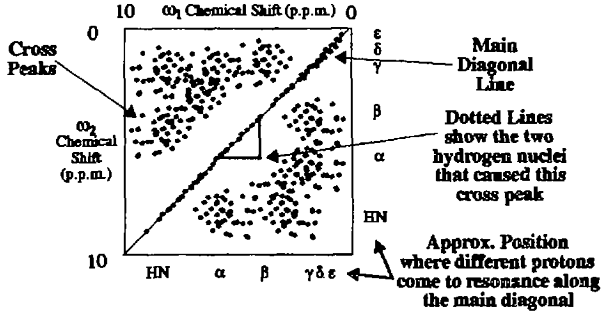
10
ω_1 Chemical Shift (p.p.m.)
0
0
Cross Peaks
ω_2 Chemical Shift (p.p.m.)
10
ε
δ
γ
β
α
HN
Main Diagonal Line
Dotted Lines show the two hydrogen nuclei that caused this cross peak
Approx. Position where different protons come to resonance along the main diagonal
HN
α
β
γδε

Figure 2: Schematic Representation of a 2D Spectrum showing the main diagonal and cross peaks. One specific cross peak and its corresponding hydrogen nuclei on the main diagonal have been identified. Along the bottom and right axis the approximate regions where each type of hydrogen nucleus comes to resonance are indicated.

Backbone Assignment

1. *Identification of amino-acid spin-systems*

As shown in Figure 3 there is a distinctive pattern of rpeaks for each residue type. The first task is to identify these spin-systems and associate them with the appropriate residue type. This is usually possible because the amide resonances associated with the backbone occur in a less crowded region of the spectrum. This involves COSY and TOCSY/HOHAHA (through bond) spectra and concentrates on the NH, $C^{\alpha}H$ and $C^{\beta}H$ regions.

Because of the number of cross peaks it is sometimes difficult to associate uniquely a spin-system to a residue type (especially in the top right of spectrum). This is because several residues have similar patterns in the NH, $C^{\alpha}H$ and C^{β}Hregions. To overcome this, similar spin-systems are grouped into J (Ser, Asp, Asn, Cys, Trp, Phe, Tyr and His) and U (Lys, Arg, Met, Gln, Glu and Pro) categories (Roberts 1993).

2. *Connection of amino-acid spin-systems*

   The corresponding $C^{\alpha}H$ of one residue is connected by a cross peak in the NOESY spectrum to the NH of the next residue. By following these connections, groups of spin-systems can be identified and matched into the primary sequence of the protein by using the Sequential Assignment Strategy (Wüthrich 1986). The J and U residue types match any one of their set of amino acid residues and are anchored in the sequence by neighbouring connectivities.

3. *Walking the Side-Chain*

   This step is only necessary for medium to long side-chains and aims to identify as many of the remaining cross peaks as possible in the TOCSY or HOHAHA spectrum. For this reason it is commonly called "walking the side-chain".

   Once the J and U residues have been placed in the sequence it is possible to look for cross peaks associated with the specific amino acids. For example, once a cross peak has been assigned specifically to a Met instead of a U, it is possible to search the TOCSY spectrum for NH-$C^{\gamma}H$ connectivities (Roberts 1993). This is achieved by using the known cross peaks and expected full spin-system to guide a search of the spectrum looking for other cross peaks that are members of the spin-system. At the end of this phase chemical shift values should be identified for a considerable number of hydrogen nuclei.

Overview of TAM

During the process of NMR structure determination it is important to note that the quality of a structure depends on the number of NOE distance constraints. These distance constraints come from the detailed structural assignment process where as many of the NOESY cross peaks as possible are assigned. This has resulted in an appreciation that the number of structural distance constraints is usually more important than their exact values (Clore, Robien, & Gronenborn 1993). Therefore an increase in the number of known chemical shift values will lead to more NOESY cross peaks being assigned, which in turn will lead to more distance constraints. This can be achieved by trying to assign chemical shift values to as many of the atom groups associated with the amino acid residues with long side chains as possible. Another advantage of detailed cross peak assignment in the TOCSY spectrum is that many of the cross peaks also appear in the NOESY spectrum. Deciding which atoms correspond to a single cross peak can become very difficult and human spectroscopists find it difficult to reason with all the possibilities at one time. Thus it would be useful to have a semi-automatic method to assist with searching for remaining chemical shift resonances.

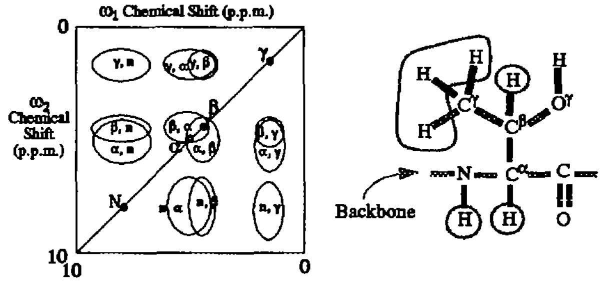

Figure 3: The Rpeak regions and structure of a threonine residue. In the structure, the hydrogens that contribute to the Rpeak regions are indicated.

The Tocsy Assignment Module (TAM) was designed to search the TOCSY spectrum looking for cross peaks that are part of the residue's spin system. The output from TAM then passes onto the structural assignment stage of ASSASSIN, which does the important structure determination. TAM's initial input is some chemical shift values and a partial set of assigned cross peaks for each residue. This input data can either be the result of a manual assignment or output from an automated system (e.g. PNA (Edwards *et al.* 1993)). Typically the search would be for the J and U residue types. This task is complicated by missing peaks and random noise.

This problem has been implemented as a Constraint Satisfaction Problem in the constraint logic programming language CHIP V4.0 (COSYTEC 1993). CHIP combines logic programming with integrated constraint solving. This means that problems can be stated declaratively, but results are obtained in an efficient way (Van Hentenryck 1989). Specifically, CHIP helps to solve complex combinatorial search problems especially when using finite domains. In CHIP terminology, constraints between special programming variables (called domain variables) are used to construct a model of the problem. Each domain variable must have a finite number of possible values and CHIP's task is to find values for all the domain variables that satisfy all the constraints. CHIP has been used to tackle a wide number of different problems such as car-sequencing (Dincbas & co workers 1988), protein β-strand topology prediction (Clark *et al.* 1993) and job-shop scheduling (Duncan 1994). Recently the connection of amino acids in 3D CA-TOCSY and CO-TOCSY experiments has been tackled as a constraint satisfaction problem (but not using constraint logic programming) by (Zimmerman, Kulikowski, & Montelione 1993) in the AUTOASSIGN system.

TAM uses the P/FDM Database Management System as a database to store the experimental NMR spectra, knowledge of the protein sequence and the results generated. P/FDM (Gray, Kulkarni, & Paton 1992) is an implementation, mostly in Prolog, of Shipman's (Shipman 1981) Functional Data Model. CHIP and P/FDM have been linked by ChipLink (Leishman 1995) which allows the passing of arbitrary prolog terms between the two systems using Remote Procedure Calls. Specifically this allows stubs of the P/FDM data retrieval primitives to be implemented on the CHIP side which transparently call the corresponding P/FDM primitives. This architecture makes it possible to program in CHIP as if one were programming in P/FDM, Figure 4.

User interaction with TAM is through a graphical interface written in X/Motif. This interface is designed to be familiar to spectroscopists and displays the current 2D NMR spectrum in a conventional format. The interface is directly coupled into TAM and it presents the progress of the assignment by colour coding the cross peaks. The cross peaks can be in one of three states: assigned, considered for assignment or not being considered at present. The cross peaks are updated after each assignment operation.

Constraint Representation

In constraint satisfaction problems it is helpful to identify four characteristics about the problem. These are: what *objects* are involved; what *attributes* do they have; what are the possible *values* of these attributes; and what are the *constraints* between the objects. Within TAM the objects are residues in the protein sequence; the attributes are rpeaks that correspond to the residue; the values of each rpeak are the cross peaks on the spectrum that correspond to the area where the rpeak is expected to appear; and the constraints determine that no cross peak be assigned twice and that the number of rpeak regions that are allowed to have a missing cross peak is set. The constraint describing the maximum number of unobserved cross peaks is very important when searching noisy data because it is very unlikely that all the rpeak regions will have an unique cross peak.

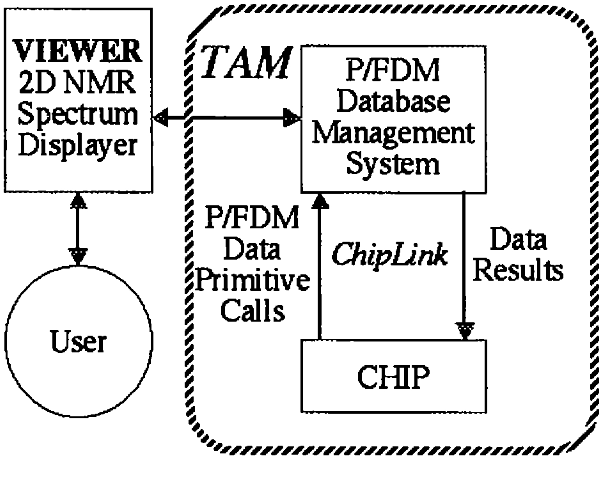

Figure 4: Architecture of TAM

Before the constraints among the rpeak regions can be posted the domains must be set. It is possible to divide the TOCSY spectrum into rpeak regions by using published statistical tables of mean and standard deviations of chemical shifts (Groß & Kalbitzer 1988). Therefore, it is possible to find each cross peak which lies in a particular rpeak region. Each cross peak is identified by a unique spectrum-wide integer, and the group of cross peaks within an rpeak region can easily be converted into an integer domain for the domain variable representing that rpeak region.

Figure 3 shows that even for relatively simple amino acids the rpeak regions overlap. If a cross peak is a member of multiple rpeak regions then when it has been assigned in one rpeak region, it should be removed from consideration in the remaining rpeak regions. To model this fact that a cross peak can only be assigned to one rpeak region we can use the *alldifferent* constraint in CHIP. This symbolic constraint implements forward checking because as soon as one domain variable becomes instantiated, the domain of all remaining domain variables is restricted. The grouping of cross peaks into potential rpeaks can take a considerable amount of time to calculate, so to remove this unnecessary overhead each time the program is run, the results are calculated once and stored in the database.

The constraints used to enforce the maximum number of rpeaks that can be left unassigned are set up when the CHIP data structures are created. Included in the domain of each rpeak is an integer value that is used to represent an unobserved cross peak. It is important to represent an unobserved cross peak because CHIP will generate a fail if a domain-variable's domain is empty. These values must be unique so they do not interact in the *alldifferent* constraints. The value that is assigned to it is calculated using:

```
MissingValue is 50000 + (Pos * 100) + Count,
```

where Pos is the amino acid sequence number of the current residue, Count is an integer incremented from one for each rpeak region associated with each residue and 50000 is a constant greater than any cross peak number.

This representation makes it easy to impose constraints on a group of rpeak domain variables. For example, if we wanted to say that at most MaxMissing rpeaks could be missing in a single residue (Residue) we could represent it as a CHIP constraint over integers thus:

```
sum_rpeaks_domain_variables(Residue, Total),
Total #< (MaxMissing + 1)
    * (50000 + (Pos * 100)) - 1,
```

Rather than including all the rpeak domain variables in the list, we could exclude rpeak domain variables that must be present to identify the spin-system. Therefore, we have a flexible way of imposing which rpeak domain variables can and cannot be allowed to be missing.

Implementation

The high level tasks performed by TAM are described in the following pseudo code.

```
load spectrum and current chemical
shift values from P/FDM
create CHIP data structures
impose rpeak region constraints using
alldifferent
call propagate_all_known_cs_values for
each residue
REPEAT
    choose best remaining residue
    assign best remaining residue using
    residue_assign
UNTIL( all residues assigned )
write results back to P/FDM
```

Firstly, the NMR spectral data and previous assignment data must be loaded from P/FDM and the appropriate CHIP data structures created (for example see (Leishman, Gray, & Fothergill 1994)). Next the necessary CHIP constraints are posted and the domain of each rpeak is limited as much as possible by calling *propagate_all_known_cs_values* for each residue. Then, the most promising residue likely to cause fewest problems for assignment is selected and assignment proceeds. This allows concentration on the easier spin systems, before moving on to the more difficult ones, such as lysine, arginine and proline. The selection continues for the remaining residues until they have all been assigned. Finally the results of the assignments are sent back to P/FDM and stored for future use.

Searching for cross peaks in spin-systems has been implemented by a number of Prolog rules. These rules are specific for each residue type and are constructed from primitive operations carried out by 2D NMR spectroscopists. These primitive operations are described below:

1. *choose_smallest*

   Compares two regions and returns the region with the smallest number of unassigned cross peaks. This rule aims to reduce the search combinatorics by working with a smaller set of peaks first. On backtracking the alternative region is returned.

2. *assign*

   Assign chooses a cross peak in a rpeak region and assumes (subject to backtracking) that it is the correct assignment.

3. *propagate*

   Propagates an assignment into the relevant rpeak regions. It takes a known chemical shift value and collects all the rpeak regions that involve this atom group. For each rpeak region it locates all the cross peaks with the same chemical shift as the assigned hydrogen nucleus and removes the rest from the domain. A margin of error is allowed because cross peaks can often drift slightly within the spectrum.

4. *propagate_all_known_cs_values*

   Propagates all known assignments into all the rpeak regions. This rule aims to make full use of any constraints on a residue's rpeak regions and reduces their domains as much as possible. It takes each each known chemical shift value and calls *propagate.*

5. *confirm*

   Aims to backup an assignment by identifying a cross peak in a symmetrical rpeak region.

6. *additional_peak_missing*

   Adds extra constraints to enforce the maximum and minimum number of rpeak regions that are missing.

7. *lower_value*

   Acts as a filter after a cross peak has been assigned. It takes two hydrogen nuclei names and makes sure that the first hydrogen nuclei's chemical shift value is lower than the second. This is particularly useful in assigning an arbitrary ordering to a sterospecific assignment or as an extra help to an assignment rule if it is known that one hydrogen nuclei's chemical shift values are always lower than anothers.

The exact representation of the assignment of a single residue type is covered in detail below. The following example shows how the primitive operations can be combined for threonine residues (Figure 3). While threonine does not have a long side chain, it is used here as an example because of its simple structure. Similar rules are being implemented for all amino acid residues.

```
% -- Assumes Threonine amide and alpha
% -- chemical shift values known
residue_assign(thr(n,a), Residue, Residue2):-
    % -- Section 1
    additional_peak_missing(thr, Residue),
    % -- Section 2
    choose_smallest(
        [region('g*' ,n), region(n, 'g*')],
        Residue, Smallest),
    assign('g*', Smallest, Residue,Residue1),
    propagate('g*',
        [region('g*',n), region('g*', a),
         region('g*',b), region(b, 'g*'),
         region(a, 'g*'), region(n, 'g*')],
         Residue1),
    % -- Section 3
    choose_smallest(
        [region('g*' ,b), region(b,' g*')],
        Residue1, Smallest1),
    assign(b, Smallest1, Residue1, Residue2),
    lower_value('g*', b, Residue2),
    propagate(b,
        [region('g*', b), region(b, 'g*'),
         region(n, b), region(b, n),
         region(a, b), region(b, a)],
         Residue2),
    % -- Section 4
    assign(region('g*', n), Residue2),
    assign(region(n, 'g*'), Residue2),
    assign(region('g*', a), Residue2),
    assign(region(a, 'g*'), Residue2),
    assign(region('g*', b), Residue2),
    assign(region(b, 'g*'), Residue2),
    assign(region(n, b), Residue2),
    assign(region(b, n), Residue2),
    assign(region(a, b), Residue2),
    assign(region(b, a), Residue2).
```

The above Prolog rule has been separated into four sections. The first section imposes the minimum number of rpeak regions to be left unassigned; sections two and three are responsible for assigning chemical shift values for g*, b atom groups respectively; and section four gives specific assignments to remaining rpeak regions. These three general tasks will be discussed in more detail.

1. *Set Number of rpeak regions left unassigned*

   When the residue data structure is created the maximum number of rpeaks that are allowed to be missing is set. When we come to search for the cross peaks we want to start off looking for a solution that has all the rpeak regions of a spin-system assigned to a cross peak. This is done by imposing a more restrictive constraint on the rpeak region's domain variables which says only N rpeaks are allowed to be missing. If this is not possible the rule backtracks and the primitive then allows one more rpeak region to be missing. This backtracking continues until either the residue has been assigned or until the rule reaches the maximum number of missing rpeak regions and the whole assignment rule backtracks. As an efficiency consideration to prevent thrashing, the primitive sets the minimum number of rpeak regions that are allowed to be missing. Without this constraint the search space for one missing rpeak also includes the search space for zero missing rpeak regions, and the search space for two missing rpeaks includes the search space for zero and one missing rpeaks. This extra constraint (illustrated below for threonine residues) overcomes the problem very cleanly.

```
additional_peak_missing(thr,Pos,Residue):-
    sum_rpeaks_domain_variables(Residue,
        Total),
    member(Missing, [0, 1, 2, 3, 4, 5, 6]),
    Max is (Missing+1)*(50000+(Pos*100))-1,
    Total #< Max,
    Min is Missing*(50000+(Pos*100)),
    Total #> Min.
```

2. *Single atom group assignment*

   For each atom group we enter a standard set of operations. Firstly a number of regions are examined to find the smallest region. This is an attempt at controlling the combinatorics by always working with as small a domain of cross peaks as possible. From this smaller region a single cross peak is chosen by *assign*. This effectively calls CHIP's *indomain* to return a viable cross peak from the domain. Notice that in section 3 a call to *lower_value* is made to filter any assignments to make sure that the $C^{\beta}H$ chemical shift is less than the $C^{\gamma}H$ chemical shift value.

   Next, *propagate* is called to apply extra constraints as a consequence of the selection. Finally, if the regions in *choose_smallest* are symmetrical and the size of the rpeak region has not been reduced by two chemical shift values intersecting, *confirm* is called to search for a cross peak in the alternative region.

   Failure is likely in *propagate* or *lower_value* when the number of allowable missing rpeaks is exceeded. This causes classic Prolog backtracking into *assign* and another cross peak is chosen. There are also a number of different assignment rules for each residue type that follow alternative assignment strategies. The advantage of CHIP is that it is being used to detect failures early on, and thus avoid the extra time a pure backtracking search would spend in fruitless assignments and then have to backtrack much later.

3. *Specific Assignments*

   Once all the assignments have been propagated, there could still be a number of rpeak regions that do not have an unique assignment. This happens when then there are a few peaks very close or overlapping. This final *assign* stage explicitly chooses one of these assignments.

An example of tracing through the assignment rule for alanine is shown in (Leishman, Gray, & Fothergill 1994).

System Performance

We are currently working with two simulated sets of test data for a 16 residue chain. One is an idealised data set that has all the necessary cross peaks in the correct rpeak regions, the other is a subset of an actual TOCSY spectrum (Figure 1). This second data set was constructed by extracting the corresponding cross peaks that were present in the actual TOCSY spectrum and adding an extra 10% random noise peaks. While the second data set is not a "full" spectrum it does represent the problems of a real data set with noise and missing peaks. The number of residues represented in this data set are Thr (7), Ala (5), Val (2), Lys (2) giving about 400 cross peaks.

The assignment tests we have run compare the time and number of backtracks of the CHIP version of TAM versus a standard backtrack Prolog version. This is similar to the comparison protocol adopted by Van Hentenryck (Van Hentenryck 1989). The standard backtracking version is identical, except that the constraints have to be used in a checking rather than an active way. This effectively means that after each *residue_assign*, the constraints to prevent a cross peaks appearing in more than one rpeak are checked; and after each assignment primitive the number of rpeaks assigned to be missing is checked.

It should be noted that we have not made a "strawman" out of the standard backtracking version in order to show the superiority of CHIP. The major difference is that assignments in one rpeak are not propagated through all the other rpeaks. This is a form of constraint propagation and would break away from our aim to compare a constraint propagation version with a standard backtracking version.

For Thr, Ala and Val, the chemical shifts and cross peaks according to the NH and $C^{\alpha}H$ hydrogens were given, and for Lys the $C^{\beta}H$ chemical shifts were also given. Typically, the system would always be started with at least the NH, $C^{\alpha}H$ and $C^{\beta}H$ chemical shifts and appropriately assigned cross peaks. We are starting from $C^{\alpha}H$ to compensate for the relative shortness of the residues in this simulated data set and to show proof of concept.

With this relatively small sample size exact numbers of backtracks and timings are susceptible to peak ordering and the position of noise peaks. Our general conclusions show that the standard backtracking version, carries out considerably more backtracks than the CHIP version and exhibits thrashing.

This is primarily due to the inability of the standard backtracking version to actively make use of the minimum number of rpeaks that are allowed to be missing. It is easy enough to detect when the maximum number of rpeaks allowed to be missing has been exceeded, but it is only possible to check if the minimum number of rpeaks has been exceeded once all the rpeaks have been given a value. As stated before, without this ability the search space expands exponentially. Further tests with a number of different spectra are being planned and should produce a quantitative analysis of the two versions.

Discussion

The CHIP system easily outperforms the standard backtracking version, but it still does takes a considerable amount of time. One reason for this is that the solution space is not drastically reduced once the constraints are initially applied. Instead, the domain variables are known but the extra constraints emerge as the search progresses. A similar problem was noted in AUTOASSIGN (Zimmerman, Kulikowski, & Montelione 1993).

NMR spectroscopists commonly use multiple spectra in their analysis and this is an area we are considering. Spectra collected at different experimental conditions or in different solutions (e.g. H_2O and D_2O) give a clearer overall view of the spectra and can overcome systematic noise effects. Our current spectra were collected in H_2O and suffer from a water resonance band ($\omega2$=4.5-6.0 ppm) which removes all cross peaks in that area.

Currently our test data only has two examples of long lysine side chains. The assignment rules for the lysine residue are considerably longer than the threonine example because it can have over 50 rpeak regions in its spin-system. The generation of alternative assignment rules seems relatively straightforward though perhaps time consuming. We are investigating the use of a graphical way in which the rules can be constructed in a visual programming manner.

TAM is the first stage of ASSASSIN (Leishman, Gray, & Fothergill 1994) and was designed to convert the results from the backbone assignment to input into the Structural Assignment Module (SAM). SAM, along with the Structural Refinement Module (SRM), cooperatively tackle the structural assignment loop by analysing the NOESY spectrum, generating 3D models (using X-PLOR (Brünger 1992)) and then reviewing the structures to identify possible mistakes and errors. Thus the two parts (TAM and SAM) have been implemented as different constraint satisfaction problems using CHIP. Implementation of SAM showed that CHIP was very successful with handling the strong geometric constraints from the NOESY spectrum and structural models. However, it needs a fairly well assigned TOCSY spectrum to get started. TAM helps significantly in this process, once again proving the value of CLP, but the constraints available are not as strong as the geometric ones. Thus we have to be satisfied with only a partial solution, but it is nevertheless crucial for the overall working of the system.

Acknowledgments

Special thanks to D. Norman (Biochemistry Department, University of Dundee) for his discussions and assistance. We are grateful to L.-Y. Lian, G.C.K. Roberts (Biological NMR Centre and Department of Biochemistry, University of Leicester) and E. Laue and A. Raine (Cambridge Centre for Molecular Recognition, Department of Biochemistry, University of Cambridge) for our 2D NMR experimental data.

References

Billeter, M.; Basus, V.; and Kuntz, I. 1988. A Program for Semi-automatic Sequential Resonance Assignments in Protein 1H NMR Nuclear Magnetic Resonance. *J. Magn. Reson.* 76:400–415.

Brünger, A. 1992. *X-PLOR Manual version 3.0.* Yale University, New Haven, CT.

Cieslar, C.; Clore, G.; and Gronenborn, A. 1988. Computer-Aided Sequential Assignment of Protein 1H NMR Spectra. *J. Magn. Reson.* 80:119–127.

Clark, D.; Rawlings, C.; Shirazi, J.; Veron, A.; and Reeve, M. 1993. Protein Topology Prediction through Parallel Constraint Logic Programming. In Hunter et al. (1993), 83–91.

Clore, G.; Robien, M.; and Gronenborn, A. 1993. Exploring the Limits of Precision and Accuracy of Protein Structures Determined by Nuclear Magnetic Resonance Spectroscopy. *J. Mol. Biol.* 231:82–102.

COSYTEC. 1993. CHIP V4 User's Guide. Technical report, COSYTEC SA, Parc Club Orsay Universite, France.

Dincbas, M., and co workers. 1988. The Constraint Logic Programming Language CHIP. In *Proceedings of the International Conference on Fifth-Generation Computing Systems.*

Duncan, T. 1994. Intelligent Vehicle Scheduling: Experiences with a Constraint-based Approach. In Milne, R., and Montgomery, A., eds., *Applications and Innovations in Expert System — Proc. ES'94 Volume II*, 281–291. Cambridge: SGES Publications.

Edwards, P.; Sleeman, D.; Roberts, G.; and Lian, L.-Y. 1993. An AI Approach to the Interpretation of the NMR Spectra of Proteins. In Hunter, L., ed., *AI and Molecular Biology.* AAAI/MIT Press. 259–288.

Gray, P.; Kulkarni, K.; and Paton, N. 1992. *Object-Oriented Databases: a Semantic Data Model Approach.* Prentice Hall Series in Computer Science. Prentice Hall International Ltd.

Groß, K.-H., and Kalbitzer, H. 1988. Distribution of Chemical Shifts in 1H Nuclear Magnetic Resonance Spectra of Protein. *J. Magn. Reson.* 76:87–99.

Hunter, L.; Searls, D.; and Shavlik, J., eds. 1993. *First International Conference on Intelligent Systems for Molecular Biology.* AAAI Press.

IUPAC-IUB. 1970. Abbreviations and Symbols for the Description of the Conformation of Polypeptide Chains. *Biochmistry* 9:3475.

Leishman, S.; Gray, P.; and Fothergill, J. 1994. ASSASSIN: A Constraint Based Assignment System for Protein 2D Nuclear Magnetic Resonance. In Milne, R., and Montgomery, A., eds., *Applications and Innovations in Expert System — Proc. ES'94 Volume II*, 263–280. Cambridge: SGES Publications.

Leishman, S. 1995. ChipLink User Manual. Technical report, University of Aberdeen, Dept. of Computing Science, King's College, Aberdeen, U.K.

Lian, L.-Y.; Derrick, J.; Sutcliffe, M.; Yang, J.; and Roberts, G. 1992. Determination of the Solution Structures of Domains II and III of Protein G from Streptococcus by 1H Nuclear Magnetic Resonance. *J. Mol. Biol.* 288:1219–1234.

Roberts, G., ed. 1993. *NMR of Macromolecules.* Oxford University Press.

Shipman, D. 1981. The Functional Data Model and the Data Language DAPLEX. *ACM Transactions on Database Systems* 6(1):140–173.

Van Hentenryck, P. 1989. *Constraint Satisfaction in Logic Programming.* MIT Press.

Wüthrich, K. 1986. *NMR of Proteins and Nucleic Acids.* Wiley, New York.

Zimmerman, D.; Kulikowski, C.; and Montelione, G. 1993. A Constraint Reasoning System for Automated Sequence-Specific Resonance Assignment from Multidimensional Protein NMR Spectra. In Hunter et al. (1993), 447–455.