RUBIDIUM, A Program for Computer-Aided Assignment of Two-Dimensional NMR Spectra of Polypeptides

CHIN YU,*,† JAN-FU HWANG,† TUNG-BO CHEN,‡ and VON-WUN SOO‡

Department of Chemistry and Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan 30043, Republic of China

Received October 7, 1991

Taking advantage of the rule-based expert system technology, a program named RUBIDIUM (Rule-Based Identification In 2D NMR Spectrum) was developed to accomplish the automatic ^{1}H NMR resonance assignments of polypeptides. Besides noise elimination and peak selection capabilities, RUBIDIUM detects the cross-peak patterns of amino acid residues in the COSY spectrum, assigning these patterns to amino acid types, performing sequential assignments using combined COSY/NOESY spectra, and finally, achieving the total assignment of the ^{1}H NMR spectrum.

INTRODUCTION

For a small protein, 2D NMR[1] is a powerful technique to determine the 3D structure in a solution state, whereas the X-ray technique is applicable for the single crystal. Whenever proteins cannot be crystallized, or when the behavior of a protein in solution is to be studied, NMR spectroscopy becomes indispensible. The NMR determination of the 3D structure of a protein typically passes three stages: (a) acquisition of 2D NMR data; (b) peak assignment of the spectrum resonances, and consequent deduction of distance constraints; and (c) generation of a molecular structure which satisfies the constraints. Among them, the peak assignment process in stage b is highly complex and time-consuming; several months may be required to accomplish it.

It is obvious that computer programs to alleviate this task would be useful. Several methodologies appear promising,[2-9] but we decided to take advantage of the rule-based expert system to accomplish the ^{1}H NMR resonance automatic assignment for the polypeptides. In this paper, we present the program that we developed, named RUBIDIUM (Rule-Based Identification In 2D NMR Spectrum). Two polypeptide samples, oxytocin and vasopressin, were tested.

MATERIALS AND METHODS

Vasopressin and oxytocin were obtained commercially (Sigma). Each peptide (10 mg) was dissolved in a DMSO-d_6 (Aldrich) solution (500 μL) to make the concentration ~20 mmol L^{-1}. 4,4-Dimethyl-4-silapentane-1-sulfonate was used as an internal standard. The NMR tube was degassed and sealed. The NMR experiments were executed on a 400-MHz spectrometer (Bruker AM-400). The double-quantum filtered COSY[10] (DQF-COSY) experiment was carried out in the phase-sensitive mode to obtain J-connectivities. The 2D NOE (NOESY)[11-13] experiment was also performed in the phase-sensitive mode with a mixing period of 120 ms.

Data Preprocessing. 2D NMR data were acquired in Bruker Aspect-3000 and transferred to a μVaxIII (MV 3600) computer. The data were processed with a FTNMR program (Hare Research) in μVaxIII to generate a SMX file and then transferred to a SUN 386i computer for noise elimination and peak selection.

Noise Elimination. The 2D data matrix was divided into 16 areas along the x-axis (ω_2 dimension). The average value was calculated for each column and set to be the threshold. Any point along the column with a value smaller the threshold is set to 0. The t_1 noise is further eliminated according to the approach expressed in the following two equations:

$$\text{If } ||I_i| - |I_j|| \leq \min(I_i,I_j), \text{ then } I_i = I_j = (I_i + I_j)/2 \quad (1)$$

$$\text{If } ||I_i| - |I_j|| > \min(I_i,I_j), \text{ then } I_i = I_j = \min(I_i,I_j) \quad (2)$$

Thus for any point i with intensity I_i (positive or negative value) in the 2D spectrum, there is a symmetrical point j (with respect to the diagonal axis) with intensity I_j. The intensities of these two points i and j were set to be the lesser of i and j if the difference between $|I_i|$ and $|I_j|$ was greater than I_i or I_j (eq 2). Otherwise, the intensities of these two points were set to be the average value of I_i and I_j (eq 1). Because the signals in the upper left and lower right regions along the diagonal axis in 2D spectra should be symmetric,[14] the asymmetric signals were regarded as noise and thus eliminated.

After this process, the unsymmetrical cross peaks (with respect to the diagonal) were suppressed, and the cross peaks masked by the t_1 noise were enhanced.

Peak Selection. After noise elimination, a peak was listed if its intensity exceeded the threshold. For the NOESY spectrum, the selected peaks were directly listed. For DQF-COSY, the peak was not listed until there was a minimum of two-thirds of the anti-phase fine structure[10] component. Figure 1 shows the COSY spectrum after peak selection was done.

The Main Program: RUBIDIUM. RUBIDIUM is written in CLIPS[15,16] and C languages. It can be run on any computer that supports the C language. It is important to have enough memory (RAM) for the computer to run RUBIDIUM. On the personal computer (IBM PC-type), for example, if there are 80 cross peaks for COSY and 150 for NOESY spectra, RUBIDIUM requires 640K RAM (including CLIPS interpreter). Other information about RUBIDIUM are the following:

Data Input. To execute RUBIDIUM, the following data input are needed:

(a) amino acid sequence of peptide
(b) table of cross peaks from the COSY spectrum
(c) table of cross peaks from the NOESY spectrum

* To whom correspondence should be addressed.
† Department of Chemistry.
‡ Department of Computer Science.

0095-2338/92/1632-0183$03.00/0

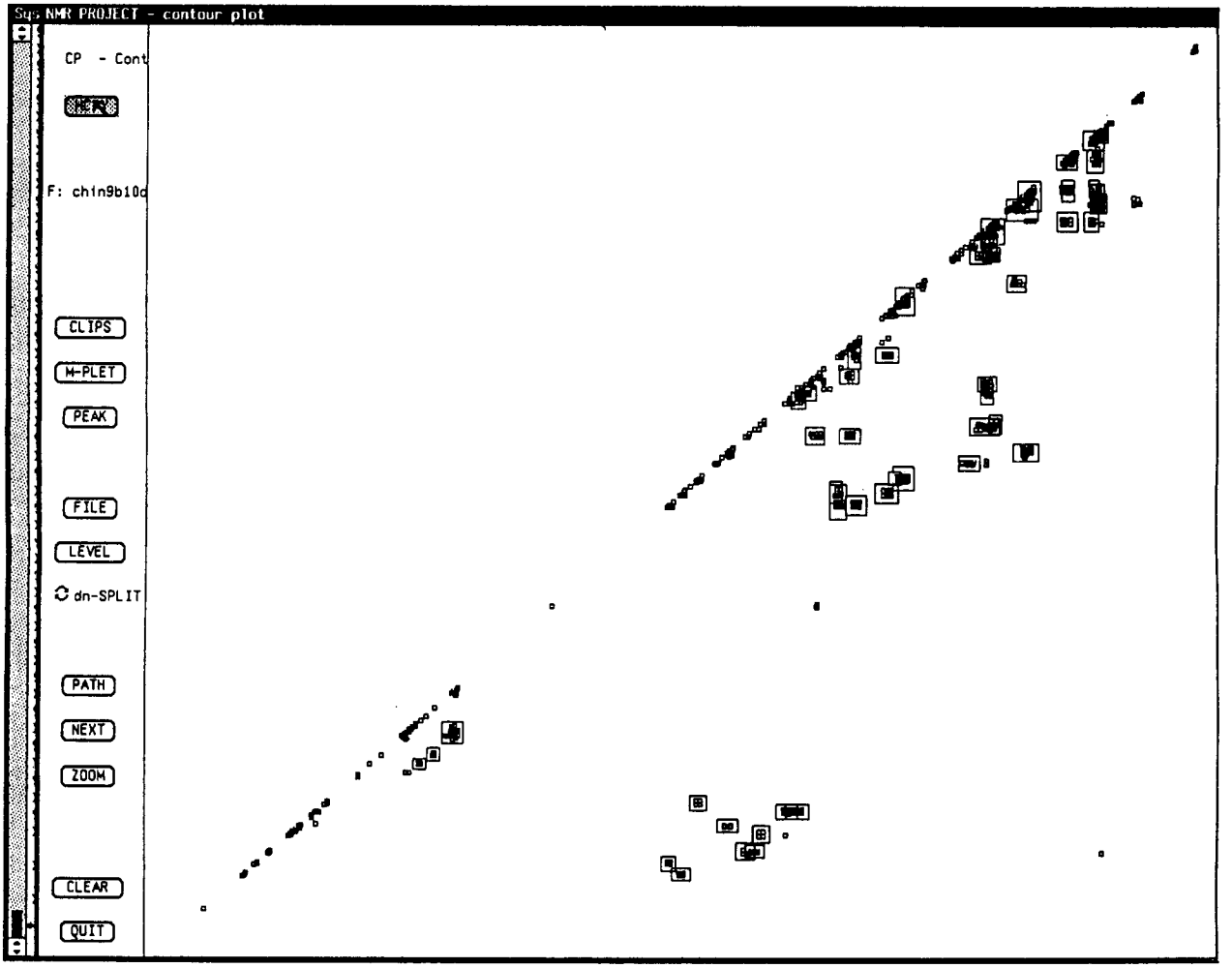

Figure 1. Result of oxytocin COSY spectrum after peak selection. Because the spectrum is symmetric about the diagonal axis, only the lower right region is shown.

The input format for the amino acid sequence of oxytocin, for example, should be typed as follows:

CYS TYR ILE GLN ASN CYS PRO LEU GLY

in which the three-letter code represents the amino acid from the N-terminus (CYS) on the left to the C-terminus (GLY) on the right. Each amino acid is represented by a three-letter code.[17]

Pattern Recognition. Each amino acid residue has a different chemical shift in the ^{1}H NMR spectrum; this effect causes the cross peak in the COSY spectrum to have a specific pattern. From each cross-peak pattern in the COSY spectrum, the amino acid residue was recognized.

The cross-peak pattern of 20 amino acid residues were generally classified into three categories:[18]

(1) Unique spin system; amino acid residues include Gly, Ala, Val, Ile, Leu, and Thr; the cross-peak pattern for these amino acid residues is unique

(2) Three-spin system; in this category, there are two $C^{\beta}H$ protons that are not degenerated and shifted relatively upfield; the amino acids such as Asn, Asp, Cys, Ser, Phe, Tyr, His, and Trp belong to this class

(3) Long side-chain system; the amino acids with a long side chain such as Glu, Gln, Met, Pro, Arg, and Lys belong to this family.

The NOESY spectrum was also taken into consideration by RUBIDIUM to enhance the accuracy for pattern recognition. This effect is important for Gln, Glu, Phe, and Tyr; for these amino acids, there are NOESY cross peaks for ($C^{\beta}H$, $C^{\delta}H$) and ($C^{\gamma}H$, $C^{\epsilon}H$). Pro is the only amino acid that has no amide proton; however, this case can be recognized by the (NH, $C^{\alpha}H$) cross peak in the NOESY spectrum.

The pattern-matching rule for the amino acid glycine (Gly) to be recognized, for example, is written as a CLIPS[16] rule:

```
(defrule mark_GLY
  (exist GLY ?n)
  (peak cosy ?NH&:  (< (abs(-?NH 8.1)) 1)
    ?aH&:  (< (abs(- ?aH 3.8)) 1) ?x ?y)
  (peak cosy ?N1&:  (<(abs (- ?N1 ?NH)) 0.05)
    ?al&: (and (<(-?al ?aH) 0.2) (!=?a1 ?aG)) ?x1 ?y1)
  (or bypass &:  (<(abs(-?aH ?al)) 0.16)
  (peak cosy ?a2&:  (<abs(- ?a2 ?aH)) 0.05)
      ?a3&:  (<(abs(-?al ?a3)) 0.05) ?x2 ?y2))
→ (assert (GLY ?NH ?aH Na ?x ?y)))
```

In this mark_GLY rule about spin pattern for glycine, the term starting with a "?" represents a variable. For example, "?NH" and "?aH" are variables that represent the first and second chemical shifts for cross peak 1, respectively. "?N1" and "?al" are variables that represent the first and second chemical shifts for cross peak 2, respectively. The chemical shifts are generally specified in terms of ppm (parts per million). The notation "bypass &:(exp)" is an externally defined function which stands for the "don't care" situation if the condition "exp" is satisfied. The variables ?x and ?y correspond to the x- and y-coordinates of the singal, respectively. The term "abs" represents the absolute value function. The notation "&:" in CLIPS represents a condition that follows which must be satisfied by the variable preceding it. For example, the pattern "v&:(exp)" signifies that the variable ?v must satisfy the condition specified in exp; exp can be a

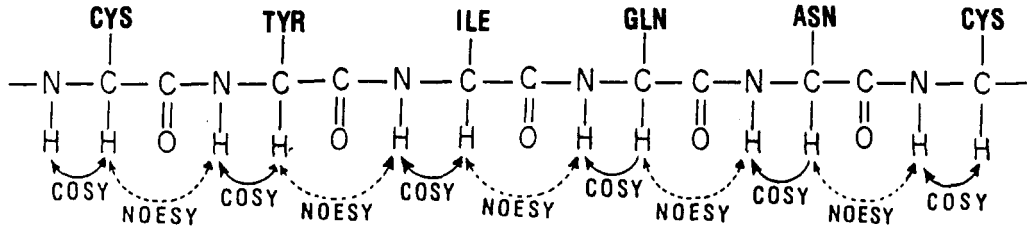

Figure 2. Sequential trace for the oxytocin backbone. Only the first six amino residues are shown. The solid lines (COSY) indicate through-bond connectivities by *J*-coupling between amide and $C^{\alpha}H$ protons in the same residue. The broken lines (NOESY) indicate through-space connectivities between amide (in the *i*th amino acid residue) and its neighboring $C^{\alpha}H$ protons (in the $i + 1$ residue).

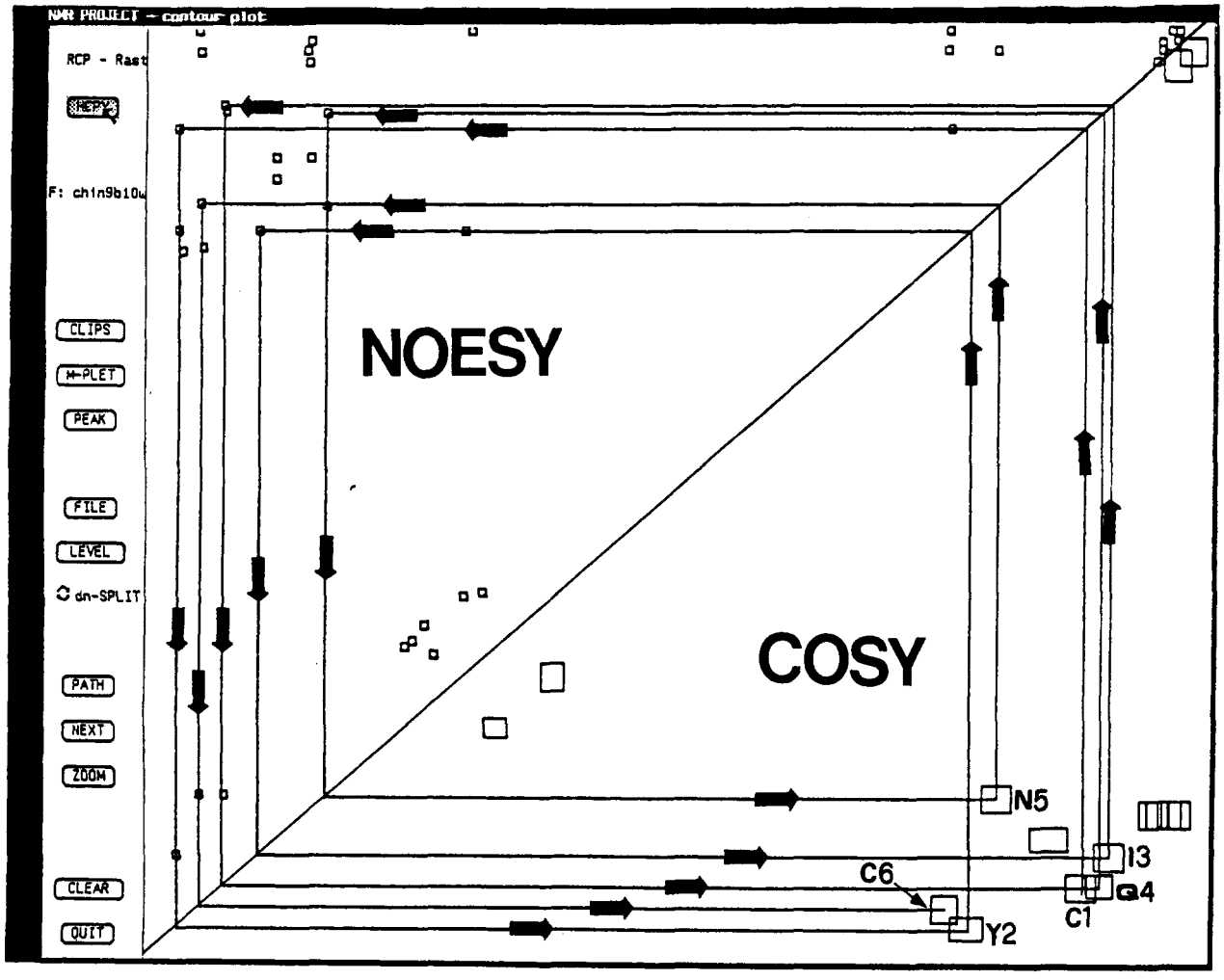

Figure 3. Result of RUBIDIUM after sequential assignment trace for the polypeptide backbone; only amide and $C^{\alpha}H$ proton regions are shown. The upper left region contains the NOESY data, and the lower right is COSY.

complex expression. The pattern-matching rule for glycine (mark_Glycine) may be read in English-like phrases as follows:

IF glycine exists in the current amino acid sequence of the polypeptide,

AND there is a cross peak in COSY which falls within the range of chemical shifts

$$(8.1 \pm 1, 3.8 \pm 1)$$ (call it cross peak 1)

AND there must be another cross peak (call it cross peak 2) of which the first chemical shift falls within the 0.05 ppm error range of that of cross peak 1 and of which the second chemical shift is within the 0.2 ppm range with that of cross peak 1,

AND if both the second chemical shifts of cross peak 1 and cross peak 2 are not within 0.16 ppm, then there must be a cross peak 3 of which the first chemical shift resides within the 0.05 ppm error range of the second chemical shift of cross peak 1 and of which the second chemical shift resides within the 0.05 ppm error range of the second chemical shift of cross peak 2,

THEN conclude that there is a glycine pattern of which the first and second chemical shifts are ?NH and ?aH, respectively, which are characterized by cross peak 1.

Sequential Assignment for Polypeptide Backbone. In the COSY spectrum, the cross peak stands for through-bond coupling, that is, the *J* interaction between $C^{\alpha}H$ and an amide proton in the same amino acid residue, whereas in the NOESY spectrum, the $C^{\alpha}H$ interacts through space with the amide proton in the subsequent amino acid residue. Combining the COSY and NOESY spectra makes it possible to assign all protons in the polypeptide backbone chain, as is illustrated in Figure 2. Utilizing this property, RUBIDIUM finds the first possible (NH, $C^{\alpha}H$) cross peak in the COSY spectrum; for through-bond connectivity a sequence-specific assignment can thus be traced.[19]

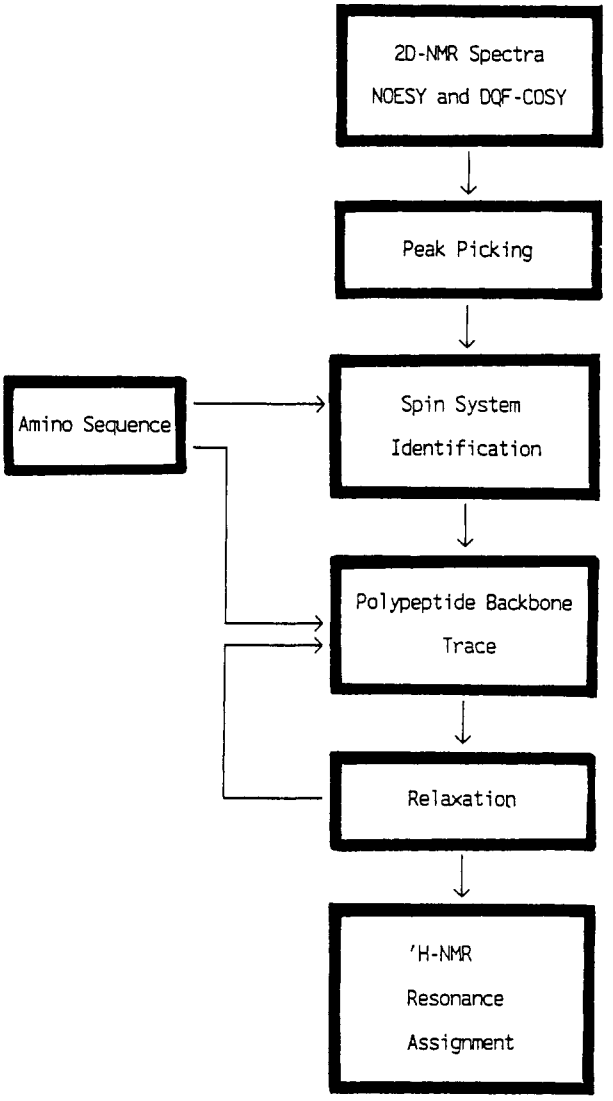

Figure 4. Simplified flowchart of RUBIDIUM protocol.

The sequential trace is stopped by having the Pro residue in a polypeptide chain, for lack of an amide proton in the Pro residue. In this case, RUBIDIUM divides the polypeptide into two segments at the Pro position and traces the sequential assignment in each segment accordingly. Figure 3 shows the sequential assignment in the polypeptide chain from amino acid residues Cys^1 to Cys^6 of oxytocin.

Relaxation Technique. The sequence-specific trace along the polypeptide backbone might fail if (1) there are too weak signals that are missed by the peak-picking process or (2) the pattern-matching rule fails to match the patterns for the given amino acid residue due to degeneracy or overlapping of the resonances. In these situations, RUBIDIUM uses a relaxation technique to solve the problems. It first relaxes the conditions for pattern matching and then performs a sequence-specific trace in both forward and reverse directions along the polypeptide backbone. In order to do relaxation for the residue, RUBIDIUM finds the cross peak in the COSY spectrum which is within the chemical shift range but is not the standard signal pattern for the residue. This new selected cross peak might significantly alter the subsequent sequence-specific trace on the peptide backbone and hence provide a solution. If the single-residue relaxation method still fails to find a complete trace, RUBIDIUM then tries to increase the number of residues to be relaxed. In this situation, the credibility of the interpretation results is decreased. The flowchart diagram of RUBIDIUM is shown in Figure 4. When the relaxation technique is applied, a sequential assignment can be made even if some cross peaks are overlapped or do not fit the pattern exactly. This is very important for the complicated system such as protein spectrum where the overlapping situation is more popular.

Table I. Automatic Assignments of Oxytocin Chemical Shift (in ppm) by RUBIDIUM

amino acid residue	NH	αH	βH	others
Cys^1	8.36	4.01	3.14	
			3.43	
Tyr^2	8.60	4.64	2.79	δH, 7.17
			3.22	ϵH, 6.72
Tle^3	8.18	3.93	1.84	γMe, 1.18
				γH, 1.50
				δM, 0.91
Gln^4	8.36	3.97	1.90	γH, 2.19
			a	ϵH, 7.36
Asn^5	7.86	4.50	2.63	
			2.69	
Cys^6	8.48	4.76	3.07	
			3.23	
Pro^7		4.34	2.04	γH, *a*
			1.87	δH, *a*
Leu^8	8.09	4.22	1.53	γH, 1.50
				δMe, 1.15
Gly^9	7.94	3.61		
		3.68		

[a] Not able to be assigned by RUBIDIUM.

Total Assignment. The searching for sequence-specific assignments done by RUBIDIUM, via a combination of the COSY and NOESY spectra, gives us the answers for the assignment of polypeptide chain. If there is only one solution of the sequence assignment, RUBIDIUM then assigns all the ^{1}H NMR resonances for both the backbone and the side chain of each amino acid residue accordingly. If there are multiple solutions of the sequence-specific assignment, RUBIDIUM keeps the consistent tracing part and neglects those of inconsistent assignment.

RESULTS AND DISCUSSION

Table I shows the results of the RUBIDIUM assignments for the 2D spectra of oxytocin. Due to the reasonable quality of the 2D spectra, all the ^{1}H NMR resonances of oxytocin were correctly assigned by RUBIDIUM. In the case of vasopressin, the poorer quality of COSY spectra makes the (NH, C^αH) cross peak undetectable. Moreover, the chemical shifts for two C^βH protons in Asn^5 are degenerate. This causes RUBIDIUM to have difficulty in its pattern-matching search. After using the relaxation technique to relax the conditions for performing pattern matching, the (NH, C^αH) COSY cross peak of Cys^1 was found. There were two cross peaks assigned as the (NH, C^αH) COSY cross peak for Tyr^2 after relaxation was applied. The reason is that an impurity in the vasopressin sample has a cross peak accidentally falling in the C^αH region of the proton chemical shift. Table II shows the total ^{1}H NMR resonance assignment by RUBIDIUM for the vasopressin 2D spectrum.

Some reasons for RUBIDIUM to fail in assigning the ^{1}H NMR resonances are the following:

(1) the (NH, C^αH) cross peaks are not detectable in COSY spectrum

(2) the degeneracy of the geminal proton chemical shift, which makes the pattern difficult to match

(3) the chemical shift is not in the normal range, leading to a cross peak in an abnormal region.

Condition 2 is overcome by applying relaxation techniques in RUBIDIUM as discussed above. For the first condition, the problem is solved by sequential tracing in reverse direction. Condition 3 could be solved by the enlargement of the chemical shift range.

Table II. Automatic Assignments of Vasopressin Chemical Shift (in ppm) by RUBIDIUM

amino acid residue	NH	αH	βH	others
Cys[1]	*c*	*c*		
Tyr[2]	8.63	4.44	2.87	δH, 7.00
			2.73	εH, 6.69
Phe[3]	8.41	4.30	2.98	δH, 1.18
			2.93	
Gln[4]	8.40	4.00	1.90	γH, 2.17
			a	εH, 7.35
Asn[5]	8.03[b]	4.57[b]	*a*	
			a	
Cys[6]	8.35	4.78	3.16	
			3.00	
Pro[7]		*a*	*a*	
			a	
Lys[8]	8.20	4.23	1.77	γH, 1.39
Gly[9]	8.01	3.66		
		3.66		

[a] Not able to be assigned by RUBIDIUM. [b] Pattern not found but could be assigned by relaxation. [c] No cross peak found in COSY spectrum.

From the results of oxytocin and vasopressin, we conclude that reasonable quality 2D spectra are required for RUBIDIUM. In general, the signal-to-noise ratio should be better than 20:1 for RUBIDIUM to perform the peak-picking process. Not only the sensitivity but also the resolution is important. The resolution may be enhanced by having a stronger magnetic field or more numerous FIDs accumulated in order to have a greater digital resolution in both dimensions of the 2D spectrum.

Several recent attempts toward automation of this task were reported such as PEPTO,[6] PROSPECT,[9] ANSIG,[8] some other systems,[2,3] and commercial packages such as NMRi.[22] Most of these programs were written in PASCAL or FORTRAN, but the algorithms were not fully automatic to assign all ^{1}H NMR resonances for the polypeptide mainly because of overlapping of the cross peaks. RUBIDIUM, on the other hand, was constructed in an expert system architecture with the relaxation routine. It could have the fully automatic assignment done for both the backbone and the side chains for polypeptides. Moreover, it is more modulized and easier to maintain.[21] PEPTO is the only exception that was an expert system implemented in PROLOG and which operated on a personal computer. However, PEPTO only performed the subtask of peak selection. Our objective is to use a real NMR spectrum as test data, which is much more difficult than the situations that use artificial or simulated spectra. Some other approaches for computer analysis of 2D NMR spectra by using either graph theory[23] or least-squares fitting[24] were also proposed. They are different from RUBIDIUM in not having the artificial intelligent structure. Besides COSY and NOESY spectra, the HOHAHA[20] spectrum could also provide valuable information for RUBIDIUM to obtain sequential assignment. Recently, 3D NMR techniques[25] are widely applied for proteins and peptides. It could disperse the chemical shifts information such as ^{1}H, ^{1}H, and ^{15}N along three orthogonal dimensions. In this case, the sequential trace is operated in a cube instead of in a plane. The project to incorporate HOHAHA and 3D NMR spectrum into RUBIDIUM is now in progress.

ACKNOWLEDGMENT

We acknowledge the support from the CAD/CAM Center of the Metal Industrial Development Centre and the National Science Council of Republic of China (NSC 81-0208-M-007-81); the former allowed us to use CLIPS, and the latter provided a SUN 386i workstation. We thank Dr. D. Hare for providing FTNMR software. The computer program RUBIDIUM described in this paper is available upon request from the authors.

Registry No. Oxytocin, 50-56-6; vasopressin, 11000-17-2.

REFERENCES AND NOTES

(1) Wüthrich, K. *NMR of Protein and Nucleic Acids*; Wiley: New York, 1986.
(2) Cieslar, C.; Clore, M.; Gronenborn, A. M. Computer-Aided Sequential Assignment of Protein ^{1}H-NMR Spectra. *J. Magn. Reson.* **1988**, *80*, 119–127.
(3) Billeter, M.; Basus, V. J.; Kuntz, I. D. A Program for Semi-automatic Sequential Resonance Assignments in Protein ^{1}H NMR Nuclear Magnetic Resonance Spectra. *J. Magn. Reson.* **1988**, *76*, 400–415.
(4) Eads, C.; Kuntz, I. D. Programs for Computer-Assisted Sequential Assignment of Proteins. *J. Magn. Reson.* **1989**, *82*, 467–482.
(5) Webre, P. L.; Malikayil, J. A.; Mueller, L. Automated Elucidation of J-Connectivities in ^{1}H NMR spectra. *J. Magn. Reson.* **1989**, *82*, 419–426.
(6) Catasti, P.; Carrara, E.; Nicolini, C. Pepto: An Expert System for Automatic Peak Assignment of Two-Dimensional Nuclear Magnetic Resonance Spectra of Proteins. *J. Comput. Chem.* **1990**, *11*, 805–818.
(7) van de Ven, F. Prospect, a Program for Automated Interpretation of 2D NMR Spectra of Proteins. *J. Magn. Reson.* **1990**, *86*, 633–644.
(8) Kraulis, P. ANSIG: A Program for Assignment of Protein ^{1}H 2D NMR Spectra by Interactive Computer Graphics. *J. Magn. Reson.* **1989**, *84*, 627–633.
(9) Van de Van, F.; Lycksell, P.; Kammen, A.; Hilbers, C. W. Computer-aided Assignment of ^{1}H-NMR Spectrum of Viral-protein-genome-linked Polypeptide from Cowpea Mosaic Virus. *Eur. J. Biochem.* **1990**, *190*, 583–591.
(10) Marion, D.; Wüthrich, K. Application of Phase-Sensitive Two-Dimensional Correlated Spectroscopy (COSY) for Measurements of ^{1}H–^{1}H Spin–Spin Coupling Constants in Proteins. *Biochem. Biophys. Res. Commun.* **1983**, *113*, 967–974.
(11) Macura, S.; Ernst, R. R. Elucidation of Cross Relaxation in Liquids by Two-dimensional NMR Spectroscopy. *Mol. Phys.* **1980**, *41*, 95–117.
(12) Jeener, J.; Meier, B. H.; Bachmann, P.; Ernst, R. R. Investigation of Exchange Processes by Two-Dimensional NMR Spectroscopy. *J. Chem. Phys.* **1979**, *71*, 4546–4553.
(13) Kumar, A.; Wagner, G.; Ernst, R. R.; Wüthrich, K. Buildup Rates of the Nuclear Overhauser Effect Measured by Two-Dimensional Proton Magnetic Resonance Spectroscopy: Implications for Studies of Protein Conformation. *J. Am. Chem. Soc.* **1981**, *103*, 3654–3658.
(14) Bauman, R.; Wider, G.; Ernst, R. R.; Wüthrich, K. Improvement of 2D NOE and 2D Correlated Spectra by Symmetrization. *J. Magn. Reson.* **1987**, *44*, 402–406.
(15) Giarratano, J. *Expert System, Principles and Programming*; PWS-Kent: Boston, 1989.
(16) Giarratano, J. *CLIPS User's Guide, Version 4.2*; Artificial Intelligence Section, Lyndon B. Johnson Space Center; 1988.
(17) IUPAC-IUB Commission on Biochemical Nomenclature. Abbreviations and Symbols for the Description of the Conformation of Polypeptide Chains. *Biochemistry* **1970**, *9*, 3471–3479.
(18) Chazin, W. J.; Hugli, T. E.; Wright, P. E. ^{1}H NMR Studies of Human C3a Anaphylatoxin in Solution: Sequential Resonance Assignments, Secondary Structure, and Global Fold. *Biochemistry* **1988**, *27*, 9139–9148.
(19) Wagner, G.; Kumar, A.; Wüthrich, K. Systematic Application of Two-Dimensional ^{1}H Nuclear Magnetic Resonance Techniques for Studies of Proteins: A Combined Use of Correlated Spectroscopy and Nuclear Overhauser Spectroscopy for Sequential Assignments of Backbone Resonances and Elucidation of Polypeptide Secondary Structures. *Eur. J. Biochem.* **1981**, *114*, 375–384.
(20) Davis, D. G.; Bax, A. Assignment of Complex ^{1}H NMR Spectra via Two-Dimensional Homonuclear Hartman–Hahn Spectroscopy. *J. Am. Chem. Soc.* **1985**, *107*, 2820–2821.
(21) Soo, V. W.; Hwang, J. F.; Chen, T. B.; Yu, C. Divide-and-Conquer, Pattern Matching and Relaxation Methods in Interpretation of 2D NMR Spectra of Polypeptides. *Proc. Natl. Comput. Symp. R.O.C.* **1991**, *2*, 620–625.
(22) New Methods Research, Inc., 6035 Corporate Dr., East Syracuse, NY.
(23) Pfändler, P.; Bodenhausen, G. Automated Analysis of Two-Dimensional Correlation Spectra. Assembly of Fragments into Networks of Coupled Spins. *J. Magn. Reson.* **1990**, *87*, 26–45.
(24) Mádi, Z. L.; Ernst, R. R. Computer Analysis of Two-Dimensional NMR Spectras. Estimation of Spectral Parameters by Least-Squares Approximation. *J. Magn. Reson.* **1988**, *79*, 513–527.
(25) Clore, G. M.; Gronenborn, A. M. Applications of Three- and Four-Dimensional Heteronuclear NMR Spectroscopy to Protein Structure Determination. *Prog. Nucl. Magn. Reson. Spectrosc.* **1991**, *23*, 43–92.