Computer-Aided Sequential Assignment of Protein ^{1}H NMR Spectra

CHRISTIAN CIESLAR, G. MARIUS CLORE,* AND ANGELA M. GRONENBORN*

Max-Planck Institut für Biochemie, D-8033 Martinsried bei München, Federal Republic of Germany

Received November 24, 1987; revised February 17, 1988

A prerequisite for the determination of the three-dimensional structure of a protein in solution is the sequential assignment of its ^{1}H NMR spectrum (*1, 2*) using two-dimensional experiments to demonstrate through-bond (*3–8*) and through-space (*9–11*) connectivities. The assignment process is highly complex and time consuming so that the development of software to aid and automate it is clearly desirable. To date a few reports dealing with some aspects of this problem have appeared, in particular the automated analysis of cross-peak patterns in COSY spectra (*12–15*) and attempts to use pattern recognition in COSY spectra of a peptide (*16, 17*).

The present note describes general procedures for computer-aided sequential assignment of protein NMR spectra. The data used to develop and test our strategy are the two-dimensional pure-phase absorption NMR spectra of a 29-amino-acid-long peptide, growth hormone releasing factor (GHRF). The spectra comprised three HOHAHA spectra, two in D_2O with mixing times of 14 and 63 ms, one in H_2O with a mixing time of 61 ms, and one NOESY spectrum in H_2O with 300 ms mixing time. The ^{1}H NMR spectrum of GHRF has been completely assigned (*18*) and its three-dimensional structure determined (*19*).

A general outline of the program is illustrated in Fig. 1 and the various steps are described below. The main body of the program is written in PASCAL with some elements written in PROLOG using the York portable PROLOG interpreter PASCAL program (*20*). The program at present runs on a VAX 8600 computer and a modified version written in C runs on a CONVEX C1-XP computer. The program is available on request.

Peak recognition and sorting. The eigenvalues of the curvature matrix for each point in the spectrum are calculated from the eight adjacent points. If both eigenvalues are negative the point is marked, and if two marked points are neighbors then they belong to the same peak. Using this method, one avoids considering ridges as a single peak. The coordinates of a peak are defined by the center of gravity of the marked points belonging to a single peak. Subsequently all peaks are ordered according to intensity, and noise is excluded by rejecting all peaks below a defined threshold value (e.g., only the strongest 200 peaks are retained).

In contrast to the sharp HOHAHA cross peaks in the NH–C$^\alpha$H and NH–aliphatic regions, those connecting aliphatic protons (e.g., C$^\alpha$H–C$^\beta$H) frequently contain fine

* Present address: Laboratory of Chemical Physics, Building 2, Room 123, National Institute for Diabetes and Digestive and Kidney Disorders, National Institutes of Health, Bethesda, Maryland 20892.

0022-2364/88 $3.00

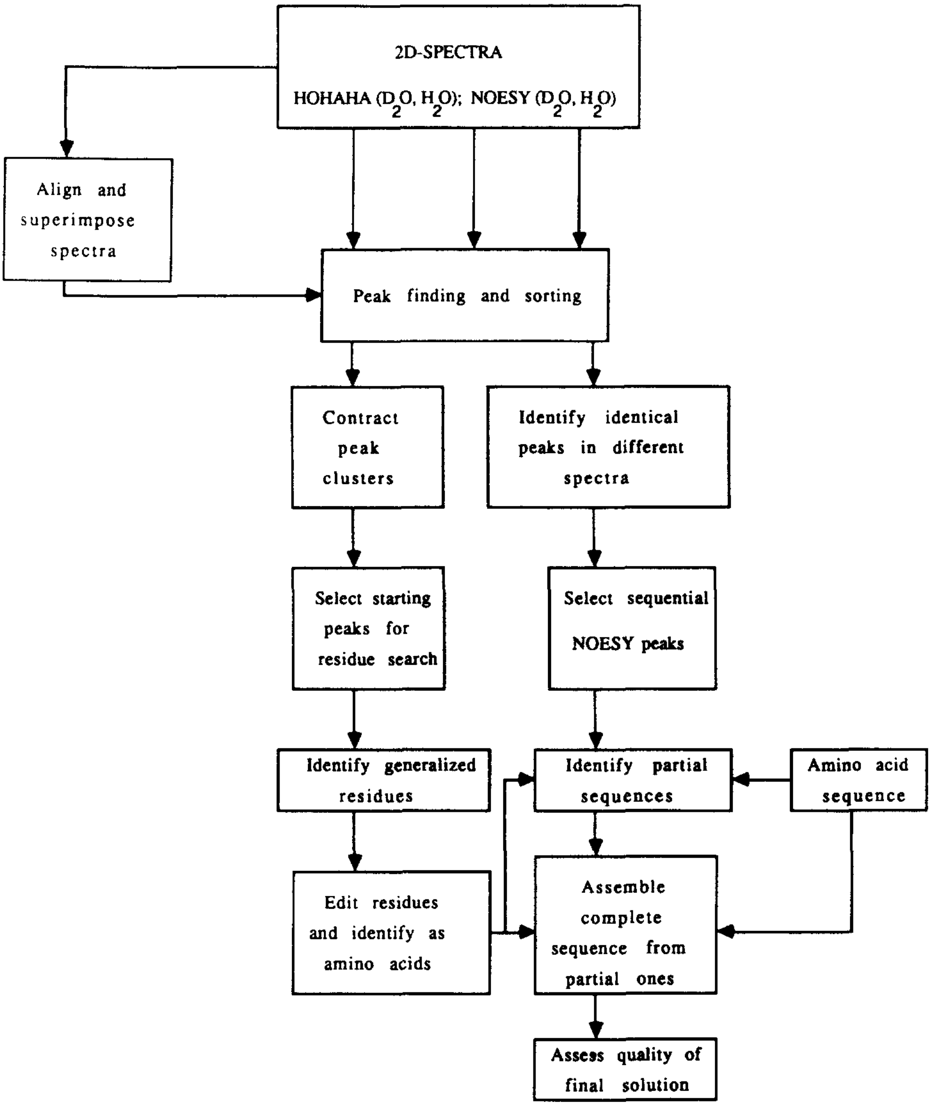

2D-SPECTRA
HOHAHA (D_2O, H_2O); NOESY (D_2O, H_2O)
Align and superimpose spectra
Peak finding and sorting
Contract peak clusters
Identify identical peaks in different spectra
Select starting peaks for residue search
Select sequential NOESY peaks
Identify generalized residues
Identify partial sequences
Amino acid sequence
Edit residues and identify as amino acids
Assemble complete sequence from partial ones
Assess quality of final solution

FIG. 1. Flowchart summarizing the program for automated assignment of protein ^{1}H NMR spectra.

structure and are thus fairly wide (see Fig. 2a). The peak-finding program may therefore find several peaks very close to each other. To utilize these cross peaks in the subsequent search for spin systems, these peak clusters are contracted into a single peak, using a minimum distance between their centers of gravity as the criterion for decision making. This minimum distance must be chosen very carefully, since, in the absence of additional information, the program may contract peaks which are close together by chance into a single peak. A typical choice would be $2J_{max}$, where J_{max} is the largest expected coupling constant. Figure 2 illustrates this procedure for the $C^{\alpha}H$–$C^{\beta}H$ region of the HOHAHA spectrum.

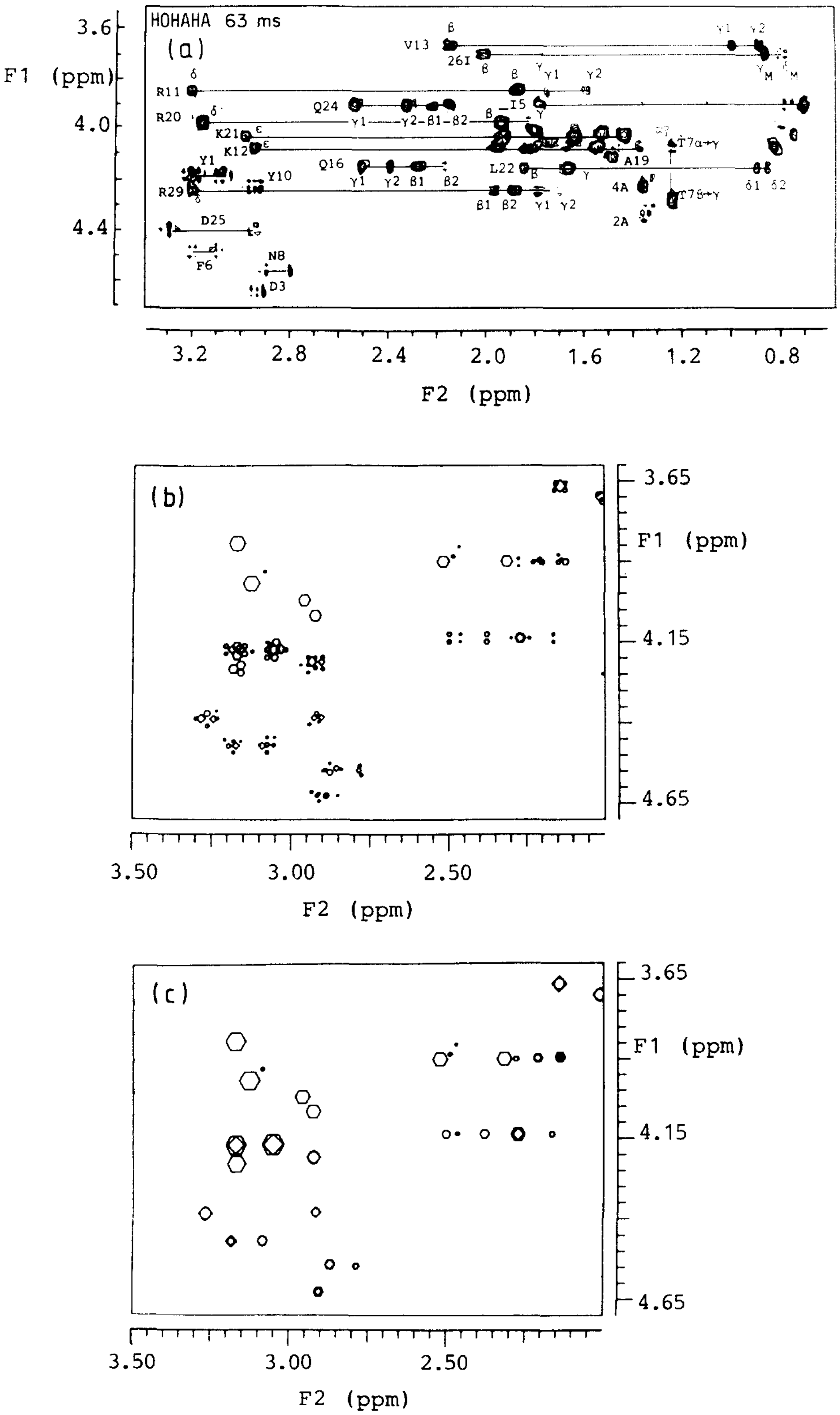
Delineation and identification of spin systems. All peaks in the spectra are represented as objects with two coordinates, the chemical shifts. For all peaks P, the identifiers u, d and the function C are defined as follows: $C(u, P)$ is the higher and $C(d, P)$ the lower chemical shift, respectively, of peak P. The function A operates on the set $[u, d]$ such that $A(u) = d$ and $A(d) = u$. For a pair of peaks P_1 and P_2 and a pair of elements, x_1 and x_2, of the set $[u, d]$, the relation Coupled(x_1, P_1, x_2, P_2) is defined as being equivalent to $C(x_1, P_1) \sim C(x_2, P_2)$. The reason for the approximate sign rather than an equals sign is the fact that, although the chemical shift of two peaks in one dimension may be the same, the computed chemical shifts of their corresponding centers of gravity in that dimension may not be identical, owing to limited digital resolution and peak overlaps. The basic procedure employed for finding spin systems involves constructing all pairs (x_i, P_i) for any given pair (x_i, P_1) for which the relationship Coupled(x_1, P_1, x_i, P_i) holds.

For the automatic identification of generalized spin systems, the complete spectrum is first divided into the NH region (6.5–9 ppm) and the aliphatic region (<5.5 ppm). For every HOHAHA peak in the NH–C$^\alpha$H region the program then tries to find a corresponding spin system, namely a peak set comprising D_2O HOHAHA peaks in the aliphatic–aliphatic region, H_2O NH–aliphatic HOHAHA peaks, and H_2O NH–aliphatic intraresidue NOESY peaks. The peak set for a given starting peak P_s is constructed in three steps: (1) Find all pairs of peaks (P_1, P_2) with x_1, x_2 elements of the set $[u, d]$ that constitute a triangle with the starting peak P_s. The triangle is defined by the relations Coupled(u, P_s, u, P_1), Coupled(d, P_1, x_2, P_2), and Coupled$(A(x_2), P_2, d, P_s)$, where the chemical shift $C(d, P_1)$ of peak P_1 is in the aliphatic region. A set is then constructed consisting of peak P_s and all the peak pairs found. (2) Find all pairs of peaks P_1, P_2, none of which are elements of the set found in step (1) with chemical shifts $C(z, P_1)$ and $C(z, P_2)$ in the aliphatic region for all possible values of the element z of $[u, d]$. If there are x_1, x_2, and x_3 elements of $[u, d]$ for which the relations Coupled(d, P_s, x_1, P_1), Coupled$(A(x_1), P_1, x_2, P_2)$, and Coupled$(A(x_2), P_2, z, P_3)$ are true and if peak P_3 is a member of the set found in step (1), then peaks P_1 and P_2 are added to the set. (3) Add all peaks P_i to the set if there are peaks P_1 and P_2 already in the set and x_1, x_2 elements of $[u, d]$ for which the relations Coupled(u, P_i, x_1, P_1) and Coupled(d, P_i, x_2, P_2) are true. Each peak set found by this procedure corresponds to a different spin system.

A mapping function M is then introduced that maps every pair (x_i, P_i) in a given peak set on a number from 0 to n, where x_i is an element of $[u, d]$. If the peaks P_i, P_j satisfy the relation Coupled(x_i, P_i, x_j, P_j), then $M(x_i, P_i) = M(x_j, P_j)$ must be true. These numbers represent the different chemical shifts present in a given spin system and are referred to as chemical-shift numbers.

As an example, consider a valine with five different chemical shifts (corresponding to the NH, C$^\alpha$H, C$^\beta$H, C$^{\gamma 1}$H, and C$^{\gamma 2}$H protons) for which nine cross peaks (numbered from 1 to 9) are found in the HOHAHA and NOESY spectra:

	NH	C$^\alpha$H	C$^\beta$H	C$^{\gamma 1}$H
C$^\alpha$H	1			
C$^\beta$H	2	3		
C$^{\gamma 1}$H	9	4	5	
C$^{\gamma 2}$H		6	7	8


HOHAHA 63 ms
(a)
F1 (ppm)
3.6
4.0
4.4
F2 (ppm)
3.2
2.8
2.4
2.0
1.6
1.2
0.8
V13
26I
R11
Q24
I5
R20
K21
K12
Q16
L22
A19
Y1
Y10
R29
4A
2A
T7α→γ
T7β→γ
D25
F6
N8
D3
β
γ1
γ2
δ
ε
γ1—γ2–β1–β2
δ1
δ2
(b)
(c)
3.65
4.15
4.65
3.50
3.00
2.50

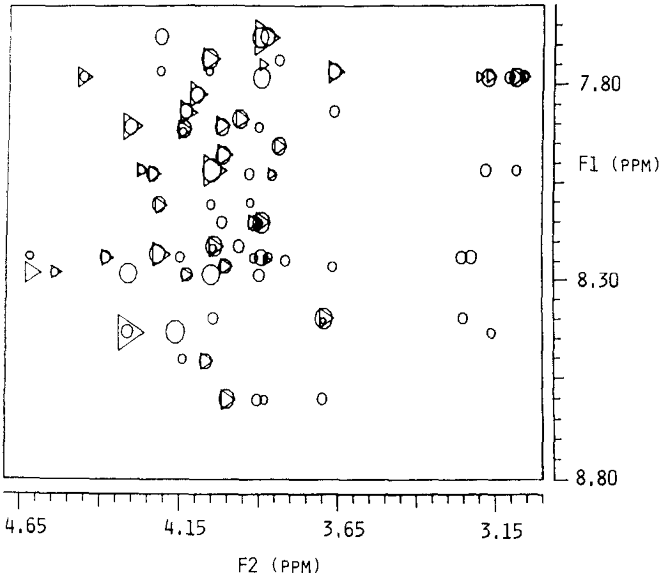


FIG. 3. NH ($F1$ axis)–C$^{\alpha}$H ($F2$ axis) region of a 200 ms NOESY spectrum (○) superimposed on a 63 ms (▷) HOHAHA spectrum recorded in 90% H_2O/10% D_2O. The size of the cross peaks is proportional to their intensity.

If peak 1 is the starting peak, step 1 will find two pairs (2, 3) and (9, 4); step 2 will find a further two pairs (6, 7) and (6, 8); and finally step 3 will find peak 5. This method relies on the existence of relayed HOHAHA peaks and intraresidue NOESY peaks; however, the latter are not assumed but identified automatically through the above relationships.

Once the various potential spin systems have been identified automatically, interactive procedures allow one to edit them. This is often necessary on account of chemical-shift degeneracy resulting in peaks of different amino acids being included in a single peak set. By this means certain chemical-shift numbers of a potential spin system can be eliminated manually and a new spin system reconstructed by the program according to the rules set out above. When this has been completed the program requires the user to supply a set of possible identifications for each spin system and the chemical-shift numbers corresponding to the C$^{\alpha}$H and C$^{\beta}$H protons. The spin system plus this additional information is then stored as a "prepared spin system" for the sequential assignment procedure described below.

At present we have not implemented any algorithm for the automatic identification of delineated spin systems with a particular amino acid or class of amino acids. This presents a more complex problem owing to chemical-shift degeneracy and must be based on both empirical rules for the expected patterns of HOHAHA cross peaks

FIG. 2. C$^{\alpha}$H($F1$ axis)–aliphatic ($F2$ axis) region of pure-phase absorption HOHAHA spectra of GHRF. (a) Original 63 ms HOHAHA spectrum as in Ref. (*18*); (b) 63 ms (○) and 14 ms (◇) spectra superimposed after processing with the peak-picking program; (c) the spectra shown in (b) after peak contraction. The size of the cross peaks in (b) and (c) is proportional to their intensity.

TABLE 1

Final Solution of the Automated Sequential Assignment

			Sequential NOEs used in assignment	
Residue	Partial sequence[a]	Chemical shift of NH proton	X (*i*)	NH (*i* + 1)
1 Tyr				
2 Ala	a	8.431	1.355	8.279
			4.315	8.281
3 Asp	a	8.280[b]	2.896	8.232
			2.925	8.236
4 Ala	a	8.229	8.228	7.679
			1.368	7.679
			4.221	7.678
5 Ile	a	7.679	0.728	7.783
			7.678	7.783
			3.902	7.783
			1.779	7.784
6 Phe	a	7.781	3.084	8.019
			7.783	8.018
			3.182	8.018
7 Thr	a	8.015	4.054	8.283
8 Asn	a	8.279[b]	2.868	8.022
			8.279	8.022
			2.791	8.025
9 Ser	a	8.026	8.025	8.107
10 Tyr	a	8.103	8.107	7.954
			2.913	7.952
11 Arg	a	7.952	3.845	7.737
			1.875	7.736
			7.954	7.733
12 Lys	a	7.736		
13 Val	d	7.766	7.766	8.260
			1.011	8.260
			0.902	8.260
			2.135	8.260
			3.669	8.260
14 Leu	d	8.259		
15 Gly				
16 Gln	b	7.866	4.142	8.501
			7.866	8.505
			2.259	8.502
			2.143	8.502
17 Leu	b	8.503	1.254	8.280
			8.506	8.284
			4.054	8.283
18 Ser	b	8.284	8.284	7.825
19 Ala	b	7.825	1.49	7.891
20 Arg	b	7.890	7.884	7.976
			1.932	7.974

TABLE 1—*Continued*

Residue	Partial sequence[a]	Chemical shift of NH proton	Sequential NOEs used in assignment	
			X (i)	NH ($i+1$)
21 Lys	b	7.976	4.026	7.907
22 Leu	b	7.912	7.912	8.210
			1.849	8.210
			0.873	8.209
23 Leu	b	8.208		
24 Gln	c	8.149	3.896	8.236
			8.146	8.243
25 Asp	c	8.236	3.252	8.397
			2.927	8.398
			2.896	8.397
			8.236	8.396
26 Ile	c	8.396	3.700	8.599
			2.005	8.596
			0.885	8.596
			8.397	8.599
27 NLeu	c	8.597	8.600	7.904
			4.026	7.907
28 Ser	c	7.907	7.902	7.560
29 Arg	c	7.566		

[a] a–d are the four partial sequences identified by the program.

[b] The incorrect solution did not use the NOE between the C$^{\alpha}$H of Asp 3 (4.630 ppm) and the NH of Ala 4 (8.230 ppm), and reversed the assignments of Asp 3 and Asn 8.

for the different spin systems and an extended data base of amino acid chemical shifts and spin patterns in proteins.

Determination of sequential connectivities. A given sequence of N residues contains $N(N-1)/2$ partial sequences. For example, for four residues there will be six partial sequences made up of the following elements: [1, 2, 3, 4], [1, 2, 3], [2, 3, 4], [1, 2], [2, 3], and [3, 4]. The result of an assignment of spin systems to a given sequence will always consist of (a) assignments to partial sequences in which every spin system is connected to its successor via interresidue NOEs of the type NH(i)-NH($i+1$), C$^{\alpha}$H(i)-NH($i+1$), and C$^{\beta}$H(i)-NH($i+1$); (b) gaps, that is to say, regions in the sequence where no spin system has been assigned to; and (c) "isolated" spin systems that are not connected via NOEs to their $i-1$ and $i+1$ neighbors in the sequence.

The sequential assignment algorithm makes use of the prepared systems and interresidue cross peaks in the NOESY spectra, in particular, those between the NH, C$^{\alpha}$H, and C$^{\beta}$H protons of a given residue, on the one hand, and the NH proton of another residue, on the other. Candidates for such interresidue NOESY cross peaks are easily identified by excluding all NOESY cross peaks which have corresponding cross peaks in the HOHAHA spectra. Figure 3 shows the NH–C$^{\alpha}$H region of the spectra with HOHAHA peaks represented by triangles and NOESY peaks by ovals.

For each prepared spin system the following is carried out: (1) The algorithm tries to find interresidue NOESY cross peaks that link NH, $C^{\alpha}H$, and $C^{\beta}H$ protons to NH protons of the other prepared spin systems. The result is a set of spin systems that are *assumed* to be directly connected via (i, i + 1) interresidue NOESY peaks to the chosen prepared spin system. (2) From the amino acid sequence and the possible identifications of the spin system, the set of all positions that this spin system could occupy in the sequence is calculated. (3) From this information the program constructs all the possible assignments for all partial sequences that are consistent with the input data; i.e., the spin systems are at allowed positions in the sequence, NOESY connectivities connect each spin system to the directly following spin system, and no spin system is used more than once in any one assignment of a partial sequence.

If there is a sequential assignment that allows one to go from the first to the last residue via (i, i + 1) NOESY connectivities, this assignment will already be found at this stage. Normally this is not the case and the assignment of the spin systems to the whole sequence consists of some assigned partial sequences with gaps.

To generate assignments of the whole sequence the assignments to partial sequences are arranged in a linear data base ordered with respect to length. The longer ones are at the beginning and picked first in the search process which proceeds as follows: (1) pick the first assignment to a partial sequence in the list; (2) go down the data base from the first partial sequence until another partial sequence is found whose assignment is consistent with the first; (3) go down the data base from the second partial sequence until a third partial sequence is found whose assignment is consistent with that of the first two partial sequences. This is repeated until no more subsequences can be found and a solution is obtained. Backtracking to earlier steps then permits one to obtain alternative solutions.

For consistency, the partial sequences must not contain overlaps (i.e., there should be no sequential overlaps in the set of partial sequences used for any given whole sequence assignment), and no particular spin system should be used in more than one position.

Since it would be too time consuming to let the program run to the end of the data base for every solution, two restrictions are imposed on the search. First, a value is set for the minimum total number of residues that should be assigned at each stage of the search. For example, in the case of GHRF we imposed the restriction that the first three steps should assign a minimum of 24 spin systems. Second, two picked partial sequences which are direct successors without a gap are only allowed if there is no NOESY connectivity between the last spin system of the first partial sequence and the first spin system of the second. If this restriction were not imposed, identical solutions would be found many times over.

All solutions for the assignment of the complete sequence are then checked with respect to their quality. At present two criteria are used: the solutions are first examined with respect to the number of assigned residues; those with the highest number of assigned residues are then selected and examined for the number of $C^{\alpha}H(i)$–NH(i + 3) and unused (i, i + 1) NOEs consistent with the data and the assignment.

The final solution of the automated sequential assignment of GHRF is given in Table 1 and agrees with that obtained manually (*18*). The spin system identifier identified 32 potential spin systems initially. Of these 5 did not correspond to any known amino acid spin pattern. After exclusion of some chemical-shift numbers followed by the spin system reconstruction procedure, 27 spin systems were identified. Of these 27, 21 could easily be identified with unique amino acids. On the basis of these 27 "prepared" spin systems, the sequential assignment program obtained two solutions for which 27 residues were assigned (residues 2 to 14 and 16 to 28). The two residues (Tyr-1 and Gly-15) were not assigned with the present version of the program as no NH-C^{α}H HOHAHA peaks were identified for them. These two solutions differed only in the assignment of residues 3 (Asp) and 8 (Asn) which were interchanged. This was due to the fact that these two residues have the same chemical shifts for their NH and C^{β}H protons and neither solution used a $C^{\alpha}H(i)$–$NH(i + 1)$ NOE between residues 3 and 4 or 8 and 9. The correct solution, however, could easily be identified as it was found to be consistent with an unused $C^{\alpha}H(i)$–$NH(i + 1)$ NOE between residues 3 and 4.

Although GHRF is only 29 residues long, this assignment is by no means trivial as the chemical-shift dispersion of the NH (7.5–8.6 ppm) and C^{α}H (3.6–4.6 ppm) resonances is small owing to the fact that under the experimental conditions employed (30% TFE, v/v) it adopts a mainly helical structure.

ACKNOWLEDGMENTS

This work was supported by the Max-Planck Gesellschaft Grant Gr 658/4-1 from the Deutsche Forschungsgemeinschaft and Grant No. 321/4003/0318909A from the Bundesministerium für Forschung und Technologie (G.M.C. and A.M.G.)

REFERENCES

1. K. WÜTHRICH, "NMR of Proteins and Nucleic Acids," Wiley, New York, 1986.
2. G. M. CLORE AND A. M. GRONENBORN, *Protein Eng.* **1,** 275 (1987).
3. R. R. ERNST, G. BODENHAUSEN, AND A. WOKAUN, "Principles of Nuclear Magnetic Resonance in One and Two Dimensions," Clarendon, Oxford, 1986.
4. W. P. AUE, E. BARTHOLDI, AND R. R. ERNST, *J. Chem. Phys.* **64,** 2229 (1976).
5. M. RANCE, O. W. SØRENSEN, G. BODENHAUSEN, G. WAGNER, R. R. ERNST, AND K. WÜTHRICH, *Biochem. Biophys. Res. Commun.* **117,** 479 (1983).
6. L. BRAUNSCHWEILER AND R. R. ERNST, *J. Magn. Reson.* **53,** 521 (1983).
7. D. G. DAVIS AND A. BAX, *J. Am. Chem. Soc.* **107,** 2821 (1985).
8. A. BAX, V. SKLENAR, G. M. CLORE, AND A. M. GRONENBORN, *J. Am. Chem. Soc.* **109,** 6511 (1987).
9. J. JEENER, B. H. MEIER, P. BACHMANN, AND R. R. ERNST, *J. Chem. Phys.* **71,** 4546 (1979).
10. A. A. BOTHNER-BY, R. L. STEPHENS, J. T. LEE, C. D. WARREN, AND R. W. JEANLOZ, *J. Am. Chem. Soc.* **106,** 811 (1984).
11. A. BAX AND D. G. DAVIS, *J. Magn. Reson.* **63,** 207 (1985).
12. B. U. MEIER, G. BODENHAUSEN, AND R. R. ERNST, *J. Magn. Reson.* **60,** 161 (1984).
13. P. PFÄNDLER, G. BODENHAUSEN, B. U. MEIER, AND R. R. ERNST, *Anal. Chem.* **57,** 2510 (1985).
14. P. PFÄNDLER AND G. BODENHAUSEN, *J. Magn. Reson.* **70,** 71 (1986).
15. S. GLASER AND H. R. KALBITZER, *J. Magn. Reson.* **74,** 450 (1987).
16. B. U. MEIER, Z. L. MADI, AND R. R. ERNST, *J. Magn. Reson.* **74,** 565 (1987).
17. K. P. NEIDIG, H. BODENMUELLER, AND H. R. KALBITZER, *Biochem. Biophys. Res. Commun.* **125,** 1143 (1984).
18. G. M. CLORE, S. R. MARTIN, AND A. M. GRONENBORN, *J. Mol. Biol.* **191,** 553 (1986).
19. A. T. BRÜNGER, G. M. CLORE, A. M. GRONENBORN, AND M. KARPLUS, *Protein Eng.* **1,** 399 (1987).
20. J. M. SPIVEY, Portable Prolog Interpreter Release 2.1, University of York, United Kingdom (1984).