Journal of Biomolecular NMR, 3 (1993) 245–251
ESCOM

J-Bio NMR 114

Computer-assisted assignment of multidimensional NMR spectra of proteins: Application to 3D NOESY-HMQC and TOCSY-HMQC spectra

Reimond Bernstein, Christian Cieslar, Alfred Ross, Hartmut Oschkinat, Jens Freund and Tad A. Holak*

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

Received 23 December 1992
Accepted 26 January 1993

Keywords: Multidimensional NMR; Computer-assisted NMR assignment; Proteins

SUMMARY

An algorithm based on the technique of combinatorial minimization is used for the semi-automated assignment of multidimensional heteronuclear spectra. The program (ALFA) produces the best assignment compatible with the available input data. Even partially misleading or missing data do not seriously corrupt the final assignment. Ambiguous sequences of the possible assignment and all alternatives are indicated. The program can also use additional non-spectroscopic data to assist in the assignment procedure. For example, information from the X-ray structure of the protein and/or information about the secondary structure can be used. The assignment procedure was tested on spectra of mucous trypsin inhibitor, a protein of 107 residues.

The assignment of multidimensional spectra of macromolecules is still the bottleneck in structure determination by NMR (Wüthrich et al., 1982; Eccles et al., 1991; Kleywegt et al., 1991). There are a number of reasons why this procedure has resisted automation: all spectra contain noise and artifacts, and the list of spin systems and their peaks are often incomplete due to spectral overlap and internal motion. The characterization of the type of the spin system is far from unique (Neidig et al., 1984; Pfändler et al., 1985; Oh et al., 1988; Eads and Kuntz, 1989; Kleywegt et al., 1989, 1991; Kraulis, 1989; Weber et al., 1989; Van de Ven, 1990; Eccles et al., 1991). The use of interproton information to identify sequential neighbors is sometimes misleading, because of overlap and long range contacts (Billeter et al., 1988; Cieslar et al., 1988; Eads and Kuntz, 1989; Kleywegt et al., 1989, 1991; Catasti et al., 1990; Mitsuhiko et al., 1990; Van de Ven, 1990; Eccles et al., 1991; Ludvigsen et al., 1991). Therefore, a *deterministic* algorithm (e.g. the

* To whom correspondence should be addressed.

recursive depth-first search) for the assignment problem will fail when the data is incomplete and in part ambiguous or contradictory.

The algorithm described in this paper, called ALFA (Algorithm for Fast Assignment), works with probabilities and in all stages uses the entire information from the data to reach a sequential assignment that is in best agreement with the available input data (Fig. 1). The information about possible ambiguities is also supplied. The input data consists primarily of NOESY and TOCSY peak lists for observed spin systems. An additional feature of the program is that the information from an already existing structure (for example, the X-ray structure) or guesses about the secondary structure can be included in the assignment procedure. The performance of the program is illustrated with the assignment of proton and nitrogen resonances of mucous trypsin inhibitor, a protein of 107 amino acid residues. The input data for ALFA were derived from the 3D NOESY-HMQC and 3D TOCSY-HMQC spectra (Fesik and Zuiderweg, 1990).

The input data consists of: (1) the primary sequence of the protein; (2) a list of the observed spin systems classified according to the type of the amino acid (in a 3D TOCSY-HMQC spectrum, for example, a spin system is a column along the F_1 dimension at a $^{15}N(F_2)$-$NH(F_3)$ frequency); and (3) a list of all potential sequential neighbors and their number of interresidue contacts to all spin systems. The corresponding intraresidue cross peaks are identified from the 3D and/or 2D NOESY/TOCSY spectra. A potential sequential neighbor is a spin system that has a *significantly* high number of observed contacts to the given spin system. Stages 2 and 3 can be either performed manually or they can be done automatically by the program from a list of TOCSY and NOESY peaks; (4) optional information from structural data. For example, a PDB file of a similar structure (X-ray) or secondary structure elements.

The algorithm is based on the technique of combinatorial minimization. Several energies are defined. The cornerstone of the program is contained in the energy term called $E_{topology}$ (Fig. 2). There are also optional energies: $E_{structure}$, E_{helix}, and E_{smooth}. The energy $E_{topology}$ is responsible for features of the assignment that involve the topology of the protein. It depends on whether an

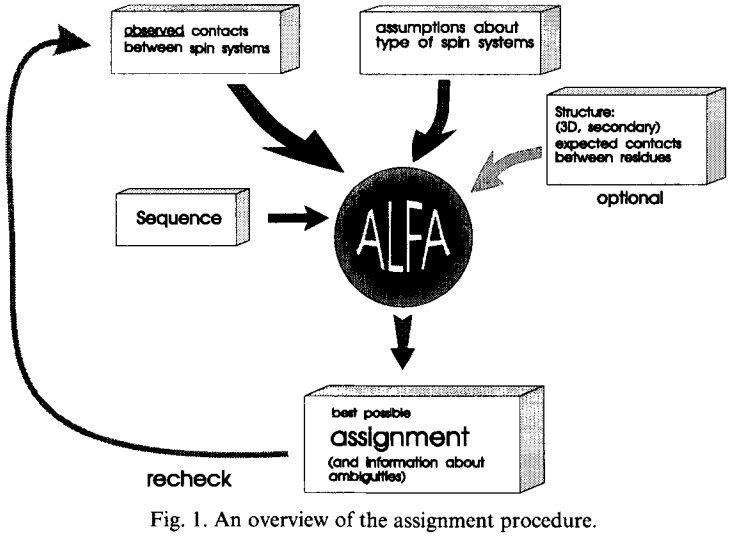

Fig. 1. An overview of the assignment procedure.

$$E_{topology} = -\sum_{\substack{i\text{ th residue}\\ \text{of sequence}}}\left(\begin{array}{c}\text{Fit of the spin system}\\ \text{assigned to i th residue}\end{array}\right) - \sum_{\substack{i\text{ th residue}\\ \text{of sequence}}}\left(\begin{array}{c}\text{contact between the spin systems}\\ \text{assigned to i th and i+1 th residue}\end{array}\right)$$

$$E_{structure} = -\sum_{\substack{i\text{ th residue}\\ \text{of sequence}}}\sum_{\substack{j\ \varepsilon\ \{\text{expected contacts}\\ \text{to the i th residue}\}}}\left(\begin{array}{c}\text{contact between the spin systems}\\ \text{assigned to i th and j th residue}\end{array}\right)$$

Fig. 2. Energy terms used in ALFA. In $E_{topology}$, the first term is 1 when the type of the observed spin system fits to the residue, and 0 otherwise. The second term is the number of contacts that are weighted by a factor of 0.5 in order to balance the contribution to the total $E_{topology}$ from the first term.

observed spin system fits to the residue it is assigned to in the sequence, and whether two spin systems that are assigned to neighboring residues have any observed interresidue contacts. The energy $E_{structure}$ is optional and is responsible for long-range contacts that could be derived from available 3D structures. An additional feature here is the use of the relaxation matrix approach to simulate observable contacts. The energy E_{helix} includes information about the secondary structure elements if available. The energy E_{smooth} takes into account the fact that the conditional probability is high for two sequential residues to have long-range contacts to two other residues that are also sequential to each other. This is especially true in β-sheets.

The program minimizes the total energy, i.e., it finds the assignment that is in best agreement with the given input data. Because the space of all possible solutions is not smooth, minimization cannot be carried out with gradient methods used, for example, in molecular dynamics calculations. However, minimizing only the energy of small parts of the protein sequence (a subspace of the whole sequence) is analogous to a force on an atom (a subspace of the whole molecule) in molecular dynamics calculations. We start with an initial assignment that is chosen arbitrarily, i.e., every experimental spin system is assigned arbitrarily to a residue in the primary sequence of a protein (Fig. 3). The program selects randomly a pair of segments of a random length (two to seven sequential residues each). The assignments of the selected residues anywhere in the two segments are rearranged systematically to optimize the total energy (Fig. 3). A spin system will be assigned to a residue if one of the following conditions is fulfilled in a descending priority: (1) the spin system matches an anticipated residue type and there is a sequential contact to an already assigned spin system; (2) only the spin system type fits; or (3) there is only a sequential contact to the neighboring residue. If none of the above conditions is fulfilled, a randomly chosen spin system is assigned to the residue. The two (modified) segments are inserted back into the sequence and a new pair of segments is chosen. This procedure is repeated until the energy cannot be minimized further. The whole procedure, which consumes 100 000 cycles, takes about 30 min on a Silicon Graphics Personal IRIS 4D workstation. The method of evaluating pairs of segments mainly avoids the problem of local minima, i.e., when correct sequential assignment of spin systems is placed at incorrect locations in the sequence. The program produces the optimal assignment together with statistics of the various other intermediate assignments once the lowest

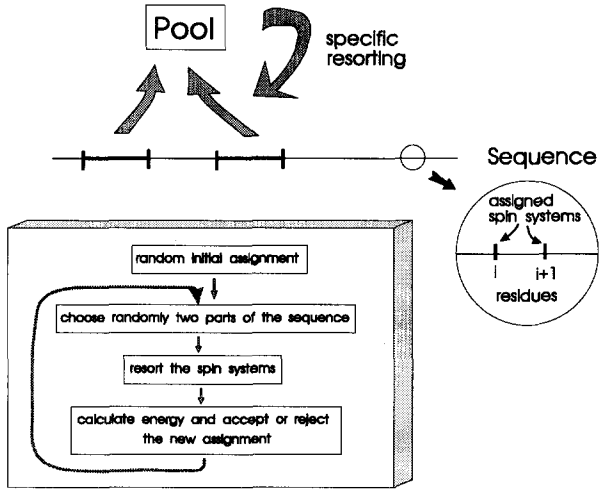

Fig. 3. Schematic illustration of the procedure for energy minimization.

energy is reached. At this point the user is provided with the list of ambiguous assignments and can recheck the input data. In all stages of the assignment the program uses all the input information so partially incorrect assignments are suppressed early in the assignment loop by conflicting with the total assignment. The algorithm is thus not easily misled either by partially incorrect or ambiguous data.

The program was tested on the mucous trypsin inhibitor (MPI), a protein of 107 amino acid residues (Heinzel et al., 1986). The protein has 13 prolines. Prolines were removed from the sequence (as they do not give any NH signals in the 3D NOESY(TOCSY)-HMQC) and no sequential contact was expected between residues X and Y that were separated by a proline (X-Pro-Y). The input data were derived from the 3D NOESY-HMQC and 3D TOCSY-HMQC spectra. The complete assignment was also performed manually using these and other (2D) spectra. After subtracting 13 prolines, 94 residues were left. As two spin systems could not be found in the spectra, the data for 92 spin systems were available for the assignment. For methodological purposes the spin systems were labeled according to their sequential assignment obtained by the manual procedure. The two spin systems that were missing corresponded to the two first residues in the protein sequence. Thus the correct assignment is: 0, 0, 1, 2.....92; where 0 means that no spin system fitted to the residue. As can be seen from Fig. 4, a substantial amount of ambiguous and misleading information is present in the input data. Fig. 4A shows the number of potential sequential neighbors for each spin system. Bars that are above the axis indicate that the correct sequential spin system was present within these potential sequential neighbors. A bar below the axis indicates that the correct sequential spin system was not supplied by the input data as a possible candidate. This situation would lead to a serious distortion in the assignment if the whole input information was not taken into account in the assignment procedure. The high number of 13 prolines plays a special role. As already mentioned, no sequential contacts were expected between residues X and Y separated by a proline (X-Pro-Y). If there was a contact X-Y, such an assignment was not punished but also it was not taken into account in the energy calculation. In general, the misassignments were most frequently near proline sites as there was

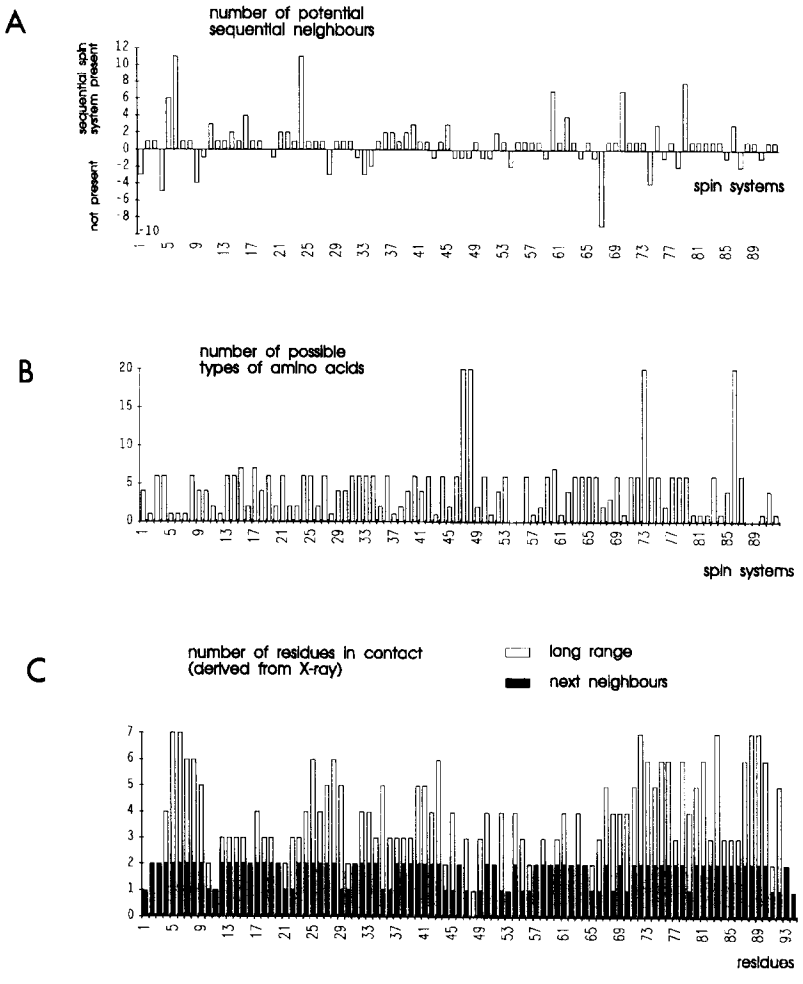

Fig. 4. (A) Number of potential sequential neighbors for each spin system found in the spectra of MPI. Bars above the axis indicate that the correct sequential spin system was present. Bars below the axis indicate that the correct sequential spin system was not supplied by the input data as a possible candidate. (B) Number of possible types of amino acid for each spin system of MPI. (C) Number of residues in contact to a given residue as determined from the X-ray structure.

a tendency for spin systems with no contacts to be placed at these X and Y positions. Fig. 4B shows the number of possible types of amino acids for each spin system. Four spin systems could not be categorized at all. However, four other spin systems could be attached to their residues uniquely. Figure 4C gives a number of long-range and sequential residues that had contacts to a given residue. These contacts were calculated from the X-ray structure.

The results of the assignment using the MPI input data are shown in Fig. 5. The upper diagrams in Fig. 5A show residues that were assigned incorrectly. The lower diagrams indicate the number of possible spin systems for each residue that also fit the assignment at the final energy. When only minimization of the energy $E_{topology}$ was carried out, the value for $E_{topology}$ reached −156 units in an arbitrary scale (Fig. 5A). The energy $E_{structure}$ calculated at this point (but

not minimized in the course of the assignment) was −339 units. The total energy is equivalent to a correct assignment for 83% of the spin systems. The 17% level of incorrectly assigned residues originated from input data that were simply incorrect or inaccurate. Nevertheless, the deviation from the correct assignment is not large because the influence of some of the misleading data was compensated by the total assignment. It can also be seen from Fig. 5 that the residues in the sequence for which the assignment was incorrect were mostly the residues with a highest variability in the assignment at the final energy. It is therefore easy to recheck the input data for such spin systems.

With the data from an X-ray structure (Grütter et al., 1988) (the X-ray structure was very similar to the NMR structure), i.e., by minimizing both the energies $E_{topology}$ and $E_{structure}$, the final assignment was 90% correct (Fig. 5B). The energies for this best assignment were: $E_{topology}$ −153 units and $E_{structure}$ −368 units. Some erroneous information used in the energy $E_{topology}$ was thus compensated by the expected long-range contacts. For comparison, we calculated the energies for the correct assignment carried out manually; these were $E_{topology}$ −153 units and $E_{structure}$ −369 units. The relationship between the different energies is shown in Fig. 5C. The energy E_{helix} was

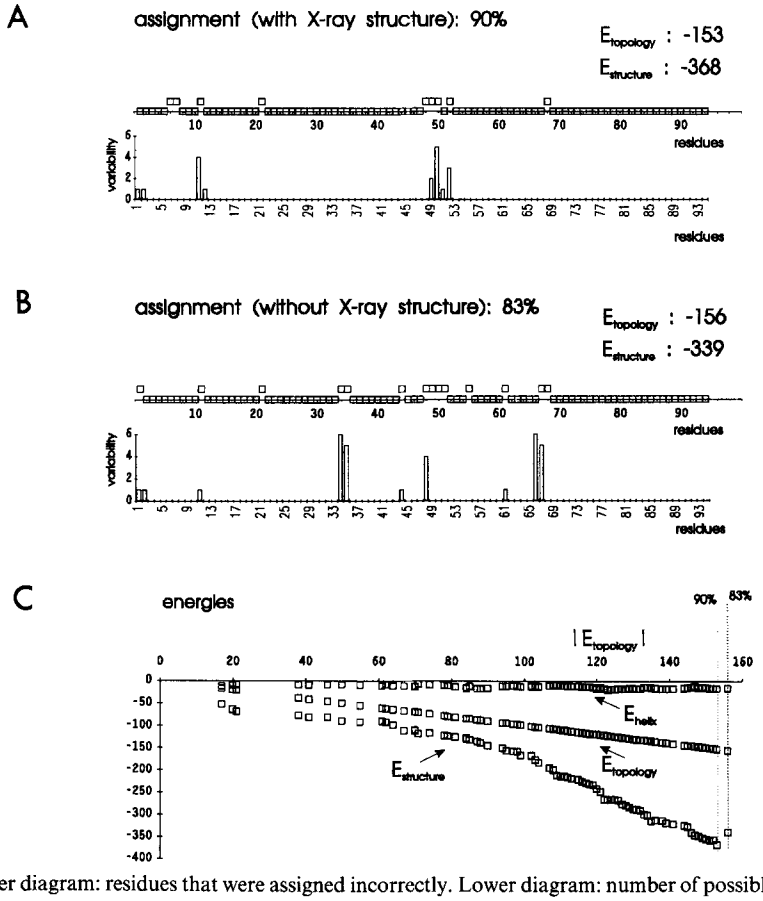

Fig. 5. (A–B) upper diagram: residues that were assigned incorrectly. Lower diagram: number of possible spin systems that also fit to a residue. (A) The assignment without the use of the X-ray data. (B) The assignment with the use of the X-ray data. (C) Plot of energies (the y-axis) versus the absolute of $E_{topology}$ (the x-axis).

calculated but was not used in the minimization. The fact that it does not change from the random value while the assignment gets better indicates, in agreement with the X-ray structure, that there is no α-helix in the protein. For the energy $E_{structure}$, we used the full relaxation matrix to predict the long-range contacts. Only the residues in contacts and not specific protons were used in the present calculation of $E_{structure}$. ALFA has also an option for classifying such contacts according to the ranges of chemical shifts for the NH, α, β, and high-field-aliphatic protons (Wüthrich, 1986).

In conclusion, ALFA is an attractive alternative to the tedious process of manual assignment. Its main strength is its capability to come to the final assignment even when a partially corrupted input is used. Work is in progress on a more interactive system for collecting the input data from the 3D NOESY-HMQC and 3D TOCSY-HMQC spectra (automated peak picking with recheck from a neuronal network). Once the sequential assignment is achieved, ALFA also provides a list of possible long-range contacts.

ACKNOWLEDGEMENT

This work was supported by research grants from the Bundesministerium für Forschung und Technologie.

REFERENCES

Billeter, M., Basus, V.J. and Kuntz, I.D. (1988) *J. Magn. Reson.*, **76**, 400–415.
Catasti, P., Carrara, E. and Nicolini, C. (1990) *J. Comput. Chem.*, **11**, 805–818.
Cieslar, C., Clore, G.M. and Gronenborn, A.M. (1988) *J. Magn. Reson.*, **80**, 119–127.
Eads, C.D. and Kuntz, I.D. (1989) *J. Magn. Reson.*, **82**, 467–482.
Eccles, C., Güntert, P., Billeter, M. and Wüthrich, K. (1991) *J. Biomol. NMR.*, **1**, 111–130.
Fesik, S.W. and Zuiderweg, E.R.P.Q. (1990) *Rev. Biophys.*, **23**, 97–131.
Grüttner, M.G., Fendrich, G., Huber, R. and Bode, W. (1988) *The EMBO J.*, **7**, 345–351.
Heinzel, R., Appelhans, H., Gassen, G., Seemuller, U., Machleid, W. and Steffens, G. (1986) *Eur. J. Biochem.*, **106**, 61–67.
Kleywegt, G.J., Lammerichs, R.M.J.N., Boelens, R. and Kaptein, R. (1989) *J. Magn. Reson.*, **85**, 186–197.
Kleywegt, G.J., Boelens, R., Cox, M., Llinas, M. and Kaptein, R. (1991) *J. Biolmol. NMR*, **1**, 23–47.
Kraulis, P.J. (1989) *J. Magn. Reson.*, **84**, 627–633.
Ludvigsen, S.L., Andersen, K.V. and Poulsen, F.M. (1991) *J. Mol. Biol.*, **217**, 731–736.
Mitsuhiko, I., Lewis, E.K. and Bax, A. (1990) *Biochemistry*, **29**, 4659–4667.
Neidig, K.P., Bodenmüller, H. and Kalbitzer, H.R. (1984) *Biochem. Biophys. Res. Commun.*, **125**, 1143–1150.
Oh, B.H., Westler, W.M., Darba, P. and Markley, J.L. (1988) *Science*, **240**, 908–911.
Pfändler, P., Bodenhausen, G., Meier, B.U. and Ernst, R.R. (1985) *Anal. Chem.*, **57**, 2510–2516.
Van de Ven, F.J.M. (1990) *J. Magn. Reson.*, **86**, 633–644.
Weber, P.L., Malikayil, J.A. and Müller, L. (1989) *J. Magn. Reson.*, **82**, 419–426.
Wüthrich, K., Wider, G., Wagner, G. and Braun, W. (1982) *J. Mol. Biol.*, **155**, 311–319.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York.