An automated procedure for the assignment of protein $^1$HN, $^{15}$N, $^{13}$C$_\alpha$, $^1$H$_\alpha$, $^{13}$C$_\beta$ and $^1$H$_\beta$ resonances*

Mark S. Friedrichs, Luciano Mueller and Michael Wittekind**

Macromolecular NMR Department, Bristol-Myers Squibb Pharmaceutical Research Institute, P.O. Box 4000, Princeton, NJ 08543-4000, U.S.A.

Received 28 December 1993
Accepted 28 March 1994

Keywords: Protein; Automation; Resonance assignment

SUMMARY

A computer algorithm that determines the $^1$HN, $^{15}$N, $^{13}$C$_\alpha$, $^1$H$_\alpha$, $^{13}$C$_\beta$ and $^1$H$_\beta$ chemical-shift assignments of protein residues with minimal human intervention is described. The algorithm is implemented as a suite of macros that run under a modified version of the FELIX 1.0 program (Hare Research, Bothell, WA). The input to the algorithm is obtained from six multidimensional, triple-resonance experiments: 3D HNCACB, 3D CBCA(CO)HN, 4D HNCAHA, 4D HN(CO)CAHA, 3D HBHA(CO)NH and 3D HNHA(Gly). For small proteins, the two 4D spectra can be replaced by either the 3D HN(CA)HA, 3D H(CA)NNH, or the $^{15}$N-edited TOCSY-HSQC experiments. The algorithm begins by identifying and collecting the intraresidue and sequential resonances of the backbone and $^{13}$C$_\beta$ atoms into groups. These groups are sequentially linked and then assigned to residues by matching the $^{13}$C$_\alpha$ and $^{13}$C$_\beta$ chemical-shift profiles of the linked groups to that of the protein's primary structure. A major strength of the algorithm is its ability to overcome imperfect data, e.g., missing or overlapping peaks. The viability of the procedure is demonstrated with two test cases. In the first, NMR data from the six experiments listed above were used to reassign the backbone resonances of the 93-residue human hnRNP C RNA-binding domain. In the second, a simulated cross-peak list, generated from the published NMR assignments of calmodulin, was used to test the ability of the algorithm to assign the backbone resonances of proteins containing internally homologous segments. Finally, the automated method was used to assign the backbone resonances of apokedarcidin, a previously unassigned, 114-residue protein.

INTRODUCTION

A major step in the process of determining protein structures by NMR is the assignment of the many nuclear resonances. This time-consuming task has been traditionally performed manually

* A preliminary account of the research presented in this manuscript was given on a poster at the Frontiers of NMR, Keystone Symposia on Molecular and Cellular Biology, Taos, NM, 1993. The macros and the subroutine code described in this paper are available to anyone who has written permission from Biosym to obtain our modified FELIX version.

** To whom correspondence should be addressed.
and remains a major bottleneck in protein structure determination projects. In efforts to streamline this procedure, a number of groups have proposed automated or semiautomated algorithms to assign homonuclear $^1$H 2D spectra (examples include Cieslar et al., 1988; Weber et al., 1988; Eads and Kunz, 1989; Catasti et al., 1990; Van de Ven, 1990; Eccles et al., 1991; Kleywegt et al., 1991; Nelson et al., 1991; Wehrens et al., 1993; Xu et al., 1993). However, the main obstacle to the successful implementation of automated assignment strategies has been the poor resolution of these spectra; severe overlap can be observed in 2D $^1$H spectra of even small proteins.

The collection of 3D and 4D NMR spectra leads to a large increase in spectral resolution. Although higher dimensionality improves the resolution of 3D homonuclear $^1$H spectra of proteins relative to their 2D counterparts (Griesinger et al., 1987; Vuister and Boelens, 1987), the most dramatic increases in spectral resolution are realized when the new dimensions are the covalent $^{15}$N and/or $^{13}$C resonances (Fesik and Zuiderweg, 1988; Marion et al., 1989). In particular, triple-resonance experiments allow correlation of the $^1$H, $^{13}$C and $^{15}$N resonances of isotopically enriched proteins via large one-bond couplings (Ikura et al., 1990; Kay et al., 1990a; reviewed in Grzesiek and Bax, 1993b). Because the triple-resonance, multidimensional experiments are relatively sensitive and have high resolution, the data are well suited as input for automated assignment algorithms.

A number of protein assignment strategies exploiting multidimensional data have been proposed that could form the basis for automation algorithms (examples include Fesik and Zuiderweg, 1990; Ikura et al., 1990, 1991a; Gao and Burkhart, 1991; Grzesiek et al., 1992; Bernstein et al., 1993; Kleywegt et al., 1993; Logan et al., 1993; Lyons et al., 1993). One approach (Domaille, 1991; Campbell-Burk et al., 1992) uses 4D HCA(NH)H and HCA(CO)NNH (Boucher et al., 1992) or 4D HNCA(HA) and HN(CO)CA(HA) (Kay et al., 1992; Olejniczak et al., 1992a) spectra to correlate the intra- and interresidue backbone $^1$HN, $^{15}$N, $^{13}$Ca and $^1$Ha resonances. This method is particularly powerful when extended to use spectra from recently introduced pulse sequences that correlate complete side-chain $^{13}$C and/or $^1$H resonances with those of the backbone amide proton and nitrogen nuclei in a single 4D (Logan et al., 1992; Clowes et al., 1993) or 3D (Montelione et al., 1992; Grzesiek et al., 1993) experiment. These pulse sequences utilize $^{13}$C isotropic mixing schemes to establish correlations among the side-chain nuclei. However, these side-chain correlation experiments have relatively low sensitivity, and as a result the use of these experiments with this approach is viable only for relatively small proteins.

A second proposed strategy utilizes the recently developed 3D experiments CBCA(CO)NNH (Grzesiek and Bax, 1992a), CBCANNH (Grzesiek and Bax, 1992b), and HNCACB (Wittekind and Mueller, 1993) that correlate the chemical shifts of the backbone amide groups with the $^{13}$Ca and $^{13}$CPh resonances. These two carbon resonances are used to establish sequential connectivities between residues. In addition, the carbon chemical-shift information can be used both to assign residues by amino acid type (Richarz and Wüthrich, 1978; Oh et al., 1988) and to help align sequentially linked residues with the primary protein sequence (Grzesiek and Bax, 1993a). While these pulse sequences yield spectra with good signal-to-noise ratios for relatively large proteins, sequential linkages based on these spectra alone may be ambiguous due to the overlap of the $^{13}$Ca and $^{13}$CPh resonances among the different residues.

Whereas a number of assignment strategies utilizing multidimensional protein spectra have been suggested, fewer accounts of actual automated implementations have been published. Descriptions of algorithms using 3D homonuclear proton data as input have been made (Cieslar