Journal of Biomolecular NMR, **26:** 335–344, 2003.
KLUWER/ESCOM

IBIS – A tool for automated sequential assignment of protein spectra from triple resonance experiments

Sven G. Hyberts & Gerhard Wagner
Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts MA 02115, U.S.A.

Received 7 January 2003; Accepted 11 March 2003

Key words: automated assignments, backbone assignments, NMR, protein structure, triple resonance experiments

Abstract

We have developed a tool for computer-assisted assignments of protein NMR spectra from triple resonance data. The program is designed to resemble established manual assignment procedures as closely as possible. IBIS exports its results in XEASY format. Thus, using IBIS the operator has continuous visual and accounting control over the progress of the assignment procedure. IBIS achieves complete assignments for those residues that exhibit sequential triple resonance connectivities within a few hours or days.

Introduction

Sequence-specific resonance assignments are the basis for solution structure determination of proteins by NMR spectroscopy. Assignment strategies relying on homonuclear NMR methods have been established more than two decades ago (Wagner and Wüthrich, 1982; Wüthrich et al., 1982), and triple resonance assignment strategies have become available soon afterwards (Ikura et al., 1990; Kay et al., 1990; Montelione and Wagner, 1989, 1990). Since then, the tools for assignments have been refined continuously. Nowadays, typically a set of key triple resonance experiments is recorded to establish sequential assignments. However, the procedures for analyzing the spectra may differ depending on the style of the individual researcher. Nevertheless, a number of attempts have been made to achieve automated assignments with computer programs based on homonuclear experiments (Nelson et al., 1991; Oschkinat et al., 1991; Xu et al., 1994) and heteronuclear triple resonance experiments (Friedrichs et al., 1994; Hare and Prestegard, 1994; Leutner et al., 1998; Lukin et al., 1997; Meadows et al., 1994; Moseley et al., 2001; Olson and Markley, 1994; Oschkinat and Croft, 1994; Zimmerman et al., 1993, 1994, 1997; Zimmerman and Montelione, 1995). It was also proposed that resonance assignments could be obtained by analysis of ^{13}C and ^{15}N separated NOESY spectra but without using J-correlated spectra (Kraulis, 1994). In parallel to automated assignment procedures, programs have become available for computer-assisted assignments where the operator organizes assignment lists and spectral strips on a screen. Examples are the EASY (Eccles et al., 1991) and XEASY programs (Bartels et al., 1995), or the program NMRview (Johnson and Blevins, 1994). Most proteins are still assigned with tools like these despite the efforts to develop fully automated assignment procedures. A shortcoming of most currently available automated assignment procedures is that the operator needs to validate the assignments afterwards, which may take significantly more time than the computerized assignment process itself. In addition, the operator wants to have quantitative measures for the quality of the assignments on a per-residue basis. Thus, automated assignment procedures should include tools for immediate validation of the assignments and provide criteria to rate the quality of each particular assignment.

Here we present a computer-based assignment tool, IBIS, that is modeled after manual assignment procedures and provides continuous visual operator control over the state of the assignment procedure. The C-program with an X/motif interface creates an input

file for the XEASY program package and presents the assignments in a visual output as sequentially aligned XEASY strips and as a list of chemical shifts. The reliability of the assignments is graded by quality factors. The program typically finds all sequential assignments that are manifested in sequential triple resonance cross peaks.

Results and discussion

Design of IBIS

When designing IBIS we had the goals to (i) obtain sequential assignments rapidly in a way that resembled established manual assignment procedures, to (ii) make sure that validation of the results would not be more time-consuming than the actual assignment procedure, and (iii) we wanted to have a quantitative measure of the quality of the assignments on a residue basis.

IBIS is written in the program language C and contains approximately 12 000 lines of code, whereof 2000 are dedicated to the user interface. Currently, IBIS is compiled on SGI/IRIX and RedHat/Linux. The program is based on comparing the peak lists of triple resonance experiments for sequential matches. The results are rated per residue based on agreement with the distribution of chemical shifts from the current content (November 10, 2001, 2076 entries) of the BioMagResBank (Seavey et al., 1991). IBIS displays the results by sequentially aligning spectral strips of triple resonance experiments in the programs, such as XEASY (Bartels et al., 1995) or NMRview (Johnson and Blevins, 1994). In this way, IBIS provides the user continuous visual control over the progress and quality of the assignment progress. IBIS requires as input a user-defined number of pairs of triple resonance experiments, such as HNCA/HN(CO)CA, HNCO/HN(CA)CO and HN-COCACB/HNCACB. The program accepts $\{^{1}H,^{15}N\}$ chemical shift data either directly from a separately recorded ^{1}H-^{15}N HSQC spectrum or indirectly from a proton-nitrogen projection of triple resonance experiments. IBIS further accepts a (H)CCONH TOCSY experiment to provide side chain assignments. As an option, additional information can be included from residue-type labeling, residue-specific experiments (Dötsch et al., 1996; Ou et al., 2001; Schubert et al., 2001a,b), or other independent and previously obtained knowledge.

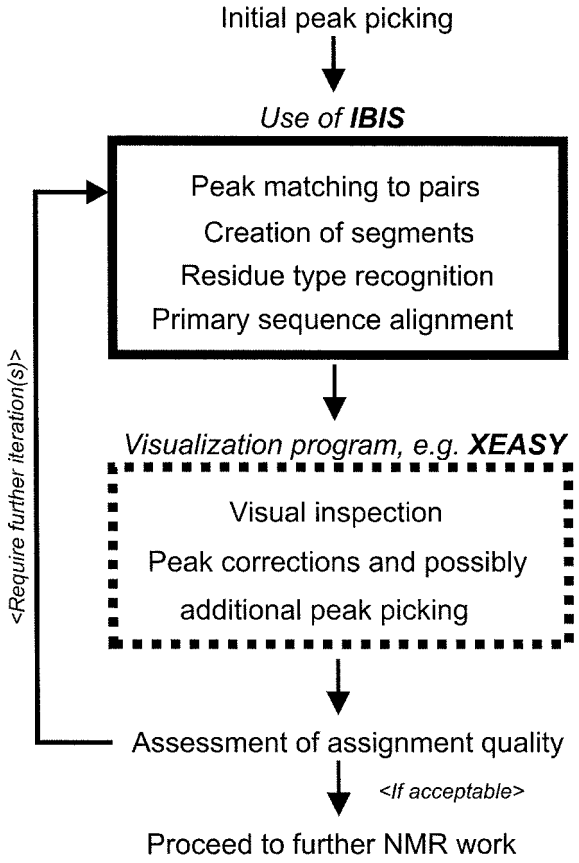

Figure 1. Flow diagram for the assignment program IBIS. The program is initiated by peak picking. The second part achieves the first sequential alignment of spin systems, which is the main task of IBIS. The results are validated by visual inspection if the alignment of triple resonance strips and by analyzing the quality factors obtained from IBIS. This process is iterated until the operator is satisfied with the validation of the results.

A flow-chart of the overall assignment process is shown in Figure 1. The assignment process consists of three elements, (i) automated or manual peak picking, (ii) peak matching and sequential alignment, and (iii) visual inspection and validation. If the assignments are not satisfactory, as judged by quality factors and visual inspection, alternate assignments suggested by IBIS are pursued in subsequent assignment cycles until the quality factors are satisfactory. We typically use five to ten cycles until assignments are complete. We typically obtain 80% to 90% of the assignments in this way depending on the signal to noise in the experiment, the number of prolines present, the dispersion of the spectra, and other complicating factors. The procedure obtains all assignments that can typically obtained by visual inspection of spectra, but in much shorter time. Obviously, the program cannot obtain as-

signments that are not manifested in triple resonance cross peaks, which accounts for the fact that we do not usually obtain 100% of the assignments. For a typical protein up to 20 kDa protein, this process takes from a few hour to a few days after the necessary experiments are recorded. The majority of this time is spent validating the assignments; the actual runtime of one IBIS iteration cycle is only a few minutes.

Definitions: Forks, tines, threads and matching

In order to better describe the flow diagram of IBIS (Figure 2), we use the following definitions. The pairs of triple resonance experiments yield connectivities between spins that resemble forks (Figure 3) with two tines where the HNCACB, HNCA, HCCANH, HNCACO establish the intra-residue connectivities, or the I-tines, the associated experiments, HN(CO)CACB, HN(CO)CA, (H)CC-CONH and HNCO establish the sequential connectivities or the S-tines of the forks. Matching of NH chemical shifts between the associated experiments that establish the connectivities between the tines of a fork is called NH matching, matching of the chemical shifts of the S-tine of a fork with the I-tine of the sequentially preceding fork is called fork matching. A series of sequentially connected forks is termed a thread and contains the chemical shifts of the connected residues plus the S-tine of the residue on the N-terminal end, which contains the side chain assignments of the residue preceeding the thread residues. IBIS creates three output files: *ibis.n.*result contains the lists of sequentially aligned spin systems with quality factors describing the degree of confidence in the assignments (Figure 4c). The file *ibis.n.*thread contains a collection of possible thread assignments ordered according to a scoring function (Figure 4b), and *ibis.n.*prot is a file that sequentially aligns triple resonance strips in the XEASY program. In the following we describe how this is used in IBIS to achieve sequential assignments.

Stage 1: Peak picking and creating 3D peak templates

The process by which IBIS operates is depicted in Figure 2. In stage 1, the assignment process is initiated with peak picking a ^{1}H-^{15}N HSQC spectrum and establishing 3D peak templates. The program accepts peak lists generated by either manually picking peaks in a visualization program, such as XEASY, or automatic peak picking with a program, such as CAPP3d (Hyberts and Wagner, unpublished), or Felix Accelrys, Inc. Initially, our intension was to rely primarily on automated peak picking. However, during the initial evaluation of IBIS in our laboratory, peak picking with a cursor on the computer screen was preferred by most users since it is very fast and provides user confidence in the peak list. If a peak list is created automatically, on the other hand, the operator wants to check visually whether the peak picking was reasonable. This is time consuming so that automatic peak picking does not yet provide a gain of assignment speed at this stage of development of our technology. After the peak list from the ^{1}H-^{15}N HSQC has been established, IBIS creates XEASY strips of the pairs of 3D triple resonance experiments at the {^{1}H,^{15}N} frequencies picked in the HSQC. If peaks are manually picked in the ^{1}H-^{15}N HSQC spectrum or the 2D projection of one of the 3D triple resonance spectra, initial 3D peak positions are created with the carbon frequency set at a uniform default value outside the carbon chemical shift range. If peaks are picked automatically in the triple resonance spectra, peak positions in all three dimensions are created.

Stage 2: NH matching of associated triple resonance strips

This stage of IBIS starts with pair matching. IBIS places side by side the associated strips from pairs of triple resonance experiments, such as HNCA and HN(CO)CA. Peak positions in the carbon dimensions of the 3D strips are entered, either by automatic 3D peak picking, or more efficiently by moving in the XEASY strips the peak marks from the initial default positions in the carbon dimensions to the actual peak positions.The number of peaks to be picked in each strip depends on the type of the experiment. For example, one peak is picked in an HN(CO)CA, two in an HNCA, and up to five in an (H)CC(CO)NH. The peaks in the HNCA are classified as sequential and define the S-tine of a fork, or intra-residue and define the I-tine of a fork. Intra-residue and sequential peaks can readily be distinguished from the pair-wise strip alignment with the HN(CO)CA. If only one peak is visible in the HNCA and has the same chemical shift as the peak in the HN(CO)CA it is assumed that sequential and intra-residue peaks are overlapped.

Side chain carbon assignments are obtained from the (H)CC(CO)NH and (H)CCNH TOCSY experiments, which provide information about the S- and the I-tines, respectively. However, the (H)CCNH TOCSY experiment is often not recorded since it has low sensitivity. In this case, we duplicate the (H)CCONH

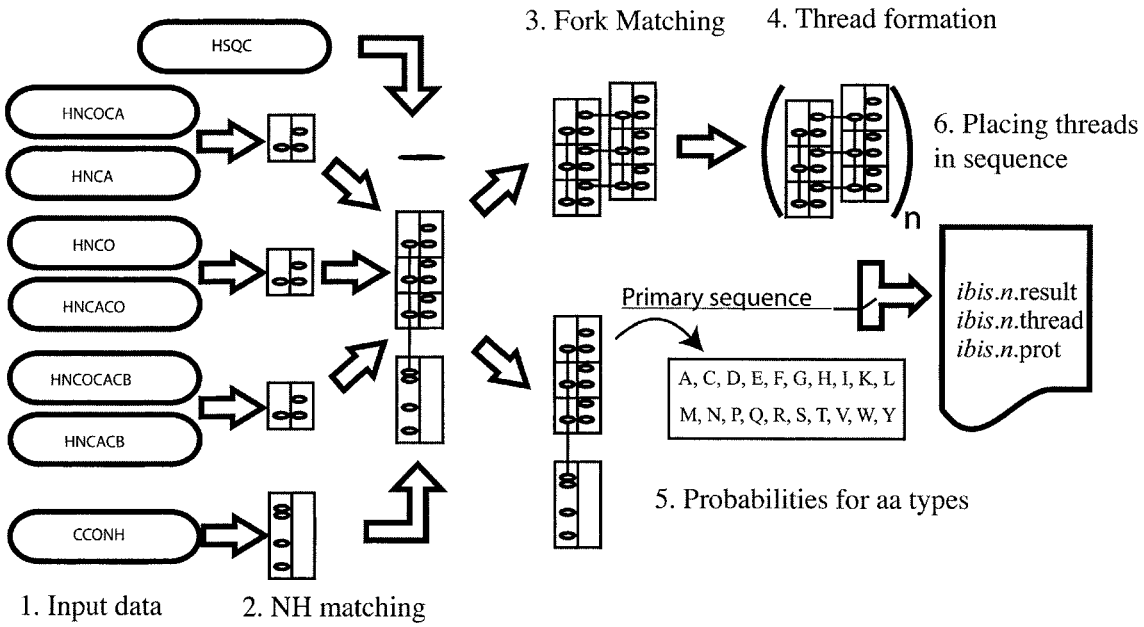

Figure 2. IBIS flow chart: Schematic drawing illustrating the way IBIS places triple resonance strips in sequential order. Ovals indicate the experimental spectra that can be accepted as input data for the assignment task. NH matching creates pairs of triple resonance strips and establishes forks. Fork matching leads to formation of threads of sequentially connected forks. The likelihoods for each tine to belong to a certain amino acid type is calculated by comparison with the BioMagRes Databank and used to place the threads at the correct positions of the primary sequence.

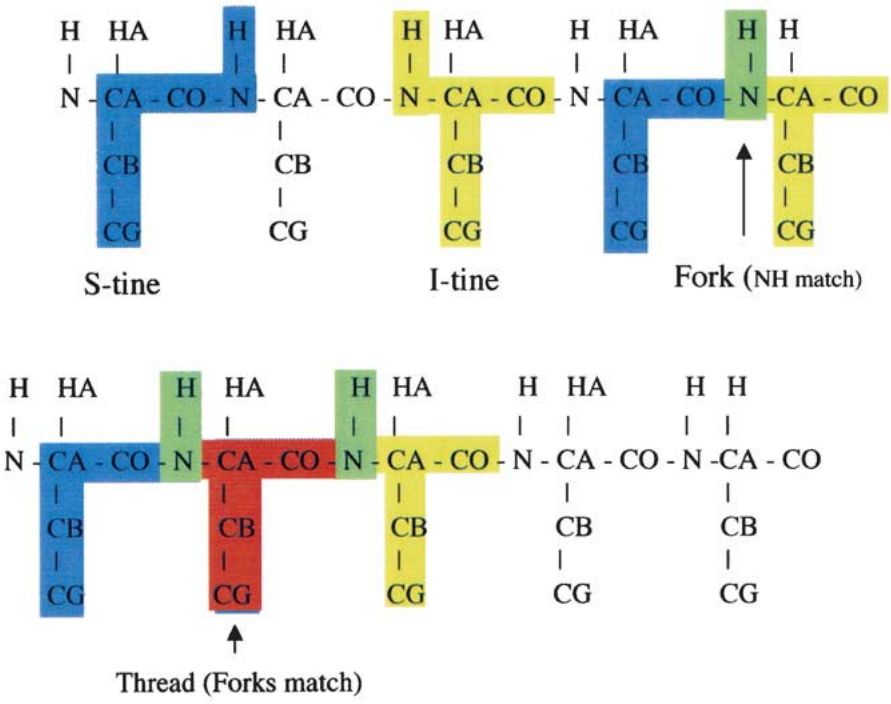

Figure 3. Definition of forks, tines, NH matching and threads. The spin systems that are connected by pairs of triple resonance experiments resemble forks. The S-tine connects the NH of a peptide group with resonances of the sequentially preceding residue, the I-tine contains spins within a residue. I-tines and S-tines are connected by NH matching. Forks are combined to threads by matching the S-tine of a residue to the I-tine of the preceding residue.

a)

```
Thread # 3 chemical shifts

Fork HN       N         CO       CA       CB       Others
-                       174.6330 56.2970  41.2090
01   9.2480   127.4620  175.7060 56.9330  28.5350
-                       175.7060 56.9330  28.5350  44.1820  27.4100
02   7.9290   105.7250  173.6870 45.2260
-                       173.6870 45.2260
03   7.9720   121.7410  174.9210 53.3820  42.8300
-                       174.9210 53.3820  42.8300
04   8.1160   120.4030  176.1680 60.1980  40.1770
-                       176.1680 60.1980  40.1770  17.9490  27.5880  14.0930
05   9.2090   128.4390  173.9810 54.4260  35.5350
-                       173.9810 54.4260  35.5350  32.3740
06   8.3500   121.9980  176.4130 59.2380  39.2930
-                       176.4130 59.2380  39.2930  13.1290  27.8850  17.5780
07   8.7380   122.2110  170.7730 59.3510  69.7250
-                       170.7730 59.3510  69.7250  19.7810
08   7.7700   118.4300  174.2670 54.8990  32.9560
-                       174.2670 54.8990  32.9560
09   8.7190   117.8930  175.8610 54.9390  43.1250
-                       175.8610 54.9390  43.1250
10   9.0320   121.8340  176.2020 54.5450  29.7140
```

b)

```
Score      Fork:        01           02           03           04           05           06           07           08
                        09           10
(166.47%) [ TYR_028] ,7> ARG_029 <5,7> GLY_030 <8,9> ASP_031 <7,6> ILE_032 <8,8> MET_033 <4,5> ILE_034 <8,8> THR_035 <5,4> HIS_036 <6
                     ,6> PHE_037 <5,5> GLU_038 <7
(149.01%) [ PHE_043] ,8> GLU_044 <6,7> GLY_045 <8,9> LEU_046 <7,5> CYS_047 <6,5> ASN_048 <4,5> GLU_049 <6,4> VAL_050 <3,1> ARG_051 <5
                     ,4> ASP_052 <7,8> MET_053 <6
(128.22%) [ PHE_037] ,8> GLU_038 <6,7> PRO_039 <0,0> SER_040 <3,3> ILE_041 <8,8> SER_042 <2,3> PHE_043 <9,8> GLU_044 <4,1> GLY_045 <3
                     ,3> LEU_046 <7,8> CYS_047 <6
(126.22%) [ ASN_048] ,5> GLU_049 <6,7> VAL_050 <2,2> ARG_051 <4,2> ASP_052 <7,6> MET_053 <4,5> CYS_054 <7,5> SER_055 <3,2> PHE_056 <4
                     ,3> ASP_057 <7,8> ASN_058 <5
(124.66%) [ ASP_052] ,6> MET_053 <4,6> CYS_054 <3,4> SER_055 <3,3> PHE_056 <8,8> ASP_057 <3,3> ASN_058 <6,5> GLU_059 <4,1> GLN_060 <5
                     ,3> LEU_061 <7,8> PHE_062 <4
(121.72%) [ PHE_056] ,8> ASP_057 <2,3> ASN_058 <2,3> GLU_059 <4,1> GLN_060 <5,4> LEU_061 <2,2> PHE_062 <9,8> THR_063 <5,4> MET_064 <7
                     ,7> LYS_065 <4,4> TRP_066 <7
(121.72%) [ LEU_082] ,6> GLU_083 <6,7> GLU_084 <2,0> ALA_085 <6,4> PHE_086 <8,8> ARG_087 <2,2> LEU_088 <5,4> TYR_089 <4,2> GLU_090 <4
                     ,2> LEU_091 <7,8> ASN_092 <5
(119.16%) [ GLU_069] ,3> GLU_070 <6,7> GLY_071 <8,9> ASP_072 <7,6> PRO_073 <0,3> CYS_074 <5,7> THR_075 <3,2> VAL_076 <3,1> SER_077 <4
                     ,4> SER_078 <4,2> GLN_079 <7
(118.64%) [ LEU_088] ,6> TYR_089 <4,5> GLU_090 <2,0> LEU_091 <7,5> ASN_092 <6,5> LYS_093 <2,2> ASP_094 <6,6> SER_095 <3,2> GLU_096 <4
                     ,2> LEU_097 <7,8> LEU_098 <5
(118.44%) [ ASN_058] ,5> GLU_059 <6,7> GLN_060 <2,1> LEU_061 <7,5> PHE_062 <8,8> THR_063 <2,3> MET_064 <5,4> LYS_065 <3,0> TRP_066 <5
                     ,4> ILE_067 <4,4> ASP_068 <6
```

c)

```
Residue   Top-scor. Thread:Fork q-score   HN      N       CO       CA      CB      Others
            Thread               S : I

ARG_029   [ARG_029]   03:01  < 7 : 5 >   9.248   127.46   175.71   56.93  28.53  CG: 27.41   CD: 44.18
GLY_030   [GLY_030]   03:02  < 7 : 8 >   7.929   105.72   173.69   45.23
ASP_031   [ASP_031]   03:03  < 9 : 7 >   7.972   121.74   174.92   53.38  42.83
ILE_032   [ILE_032]   03:04  < 6 : 8 >   8.116   120.40   176.17   60.20  40.18  CG1: 27.59  CG2: 17.95  CD1: 14.09
MET_033   [MET_033]   03:05  < 8 : 4 >   9.209   128.44   173.98   54.43  35.53  CG: 32.37
ILE_034   [ILE_034]   03:06  < 5 : 8 >   8.350   122.00   176.41   59.24  39.29  CG1: 27.89  CG2: 17.58  CD1: 13.13
THR_035   [THR_035]   03:07  < 8 : 5 >   8.738   122.21   170.77   59.35  69.72  CG2: 19.78
HIS_036   [HIS_036]   03:08  < 4 : 6 >   7.770   118.43   174.27   54.90  32.96
PHE_037   [PHE_037]   03:09  < 6 : 5 >   8.719   117.89   175.86   54.94  43.12
GLU_038   [GLU_038]   03:10  < 5 : 7 >   9.032   121.83   176.20   54.54  29.71
PRO_039                      < - : - >                    175.70   63.98  31.33  CG: 27.85   CD: 50.88
SER_040   [SER_040]   02:01  < 8 : 5 >   7.190   108.89   174.74   56.90  63.39
ILE_041   [ILE_041]   02:02  < 8 : 7 >   7.549   124.50   173.05   61.74  39.00  CG1: 28.59  CG2: 16.96  CD1: 14.54
SER_042   [SER_042]   02:03  < 6 : 5 >   7.976   119.39   175.53   56.36  66.26
PHE_043   [PHE_043]   02:04  < 5 : 5 >   9.273   121.97   177.09   62.07  39.21
GLU_044   [GLU_044]   02:05  < 6 : 6 >   8.919   119.63   178.87   60.07  29.27  CG: 36.75
GLY_045   [GLY_045]   02:06  < 5 : 7 >   8.272   109.22   176.77   46.69
LEU_046   [LEU_046]   02:07  < 5 : 5 >   8.560   125.39   178.46   58.61  41.21  CG: 27.42  CD1: 26.75  CD2: 25.06
CYS_047   [CYS_047]   02:08  < 5 : 4 >   8.111   116.24   176.97   64.77  26.77
ASN_048   [ASN_048]   02:09  < 2 : 6 >   8.058   117.75   177.80   55.98  37.67
GLU_049   [GLU_049]   02:10  < 4 : 7 >   7.913   120.62   179.55   59.02  29.35  CG: 35.59
VAL_050   [VAL_050]   02:11  < 5 : 4 >   8.563   121.15   177.57   66.87  41.43
```

Figure 4. Format of the output data of IBIS. (a) Chemical shifts obtained for a thread (# 3 of the N-terminal domain of PKC iota) prior to placement in the sequence (Roehrl et al., 2003). (b) The *ibis.n.*thread file lists all possible placements of the thread at different positions of the protein sequence. The first column lists the value of the overall scoring function as obtained from Equation 3. The quality factors for the S-tine of each residue are placed in front of the amino acid name, followed by a > symbol, those for the I-tine are listed after the amino acid name and a < symbol. The individual quality factors range from 9 (best) to 0 (worst). The assignment picked is based on the best overall score. In addition, the next-scoring choices can readily be ruled out based on the low S-tine score for ARG_051 (second line) or placement of PRO_039 in the middle of a thread. While the example shown here allows an unambiguous choice, other cases may not be so clear and alternate choices can be tried out in subsequent iterations of IBIS. (c) Format of the *ibis.n.*results file. This file contains the finally selected assignment of threads to a stretch of the amino acid sequence. The third column lists the number of the thread followed by the placement of the residue within the thread. The fourth column lists the quality factor-based on thelikelihood that the assigned chemical shifts of the S-tine and the I-tine are consistent with the amino acid type in the $i - 1$ and the i position.

strips to provide a similar format as for the other triple resonance pairs. So far, we are using only carbon side-chain resonances, and up to five side-chain carbon resonances are accepted in each strip.

It should be reiterated that the number of triple resonance experiments used is optional. As a minimum, one pair of triple resonance experiments is required for the program to function but the performance of IBIS increases with more experiments available. In the extreme case, when only a ^{1}H-^{15}N HSQC is available prior to recording triple resonance experiments, IBIS may be used to provide template peak lists for all of the triple resonance spectra as a means for initializing the assignment process and to pursue assignments on the fly while the different triple resonance experiments are acquired.

Although desirable, IBIS doesn't necessarily require a ^{1}H-^{15}N HSQC spectrum to start with. It can directly start from peak picking the ^{1}H-^{15}N projections of the triple resonance experiments and proceed to establishing 3D peak templates and pair matching. If this route is chosen, the signals of asparagine, glutamine and tryptophane side chains are absent, and the user doesn't have to make the decision whether peaks are from backbone or side-chain groups at this point. Whereas identification of side-chain NH_2 signals is often trivial, distinction of tryptophane side-chain NH signals from backbone NH crosspeaks is sometimes not straightforward, in particular in large proteins.

IBIS pursues NH matching by analyzing triple resonance pairs in the order of HNCA/HN(CO)CA, HN(CA)CO/HNCO and HNCACB/HN(CO)CACB, respectively. The user can decide how many triple-resonance pairs to employ. In the optimal case if spectra have a high signal-to-noise ratio and don't suffer from resonance overlap complete assignments can be obtained using a single HNCA/HN(CO)CA pair (Ivanov et al., to be published).

Stage 3: Sequential links and threads

Here, IBIS pursues the concatenation of forks into pairs of forks, and subsequently into longer threads. In this process, the HN, N, CO, CA, CB and other carbon chemical shifts of the S-tine of each fork are compared with those of the I-tines of all other forks. Two forks f_m and f_n are accepted as sequentially linked in the direction $f_m \rightarrow f_n$ if all available carbon shifts of the I-tine of f_m match those of the S-tine of f_n within a specified tolerance, s_o. IBIS always selects the sequential connectivities that have the most matches. If there are two or more possible matches IBIS will use the 'iterative-best-fit' approach where the different possibilities are tested for a best overall fit. As an example, thread #3 obtained for the N-terminal domain of protein kinase C iota (PKC iota) is shown in Figure 4a (Roehrl et al., 2003).

Stage 4: likelihoods for amino acid type assignments based on chemical shifts

In order to place threads on the correct segments of the polypeptide chains, IBIS uses chemical shifts and the number of assigned resonances. The program calculates the likelihood that a tine of a fork belongs to a certain amino acid type. This has two aims, to (i) establish some assignments to residue types, and (ii) validate the sequential connectivities. For this purpose we utilize the content of the BioMagResBank (Seavey et al., 1991). Mean values δ_c^{aa} and standard deviations σ_c^{aa} were calculated for all chemical shifts c of every amino acid *aa* from the content of the data base. We approximated the distribution of the chemical shifts with Gaussians and used these functions to calculate likelihoods that a tine (as defined by the list of chemical shifts for CA, CB, CG and CO) belongs to a particular amino acid type. This is calculated both for the I-tine and the S-tine. The likelihood $\rho(aa|l)$ that a particular tine l belongs to the amino acid type *aa* is defined as:

$$\rho(aa|l) = \frac{0.15*\left(1.0-\frac{\left|n_{\mathrm{exp}}^{aa}-n_{\mathrm{obs}}^{aa}\right|}{n_{\mathrm{exp}}^{aa}}\right)+\sum_c w_{\mathrm{atom}}*G\left(\delta_{c\ \mathrm{exp}}^{aa}-\delta_{c\ \mathrm{obs}}^{aa},\sigma_c^{aa}\right)}{0.15+\sum_c w_{\mathrm{atom}}}, \quad (1)$$

$$G\left(\delta_{c\ \mathrm{exp}}^{aa}-\delta_{c\ \mathrm{obs}}^{aa},\sigma_c^{aa}\right) = e^{-\left(\frac{\delta_{c\ \mathrm{exp}}^{aa}-\delta_{c\ \mathrm{obs}}^{aa}}{\sigma_c^{aa}}\right)^2}. \quad (2)$$

Here the sum runs over all chemical shifts c in the tine. G is the Gaussian function with the standard deviation found for chemical shift c in residue type *aa*. The function in the nominator of Equation 1 prior to the sum sign takes into account whether the observed number of side-chain chemical shifts, n_{obs}, in one tine matches the number, n_{exp}^{aa}, expected for a particular amino acid type *aa*. It is weighted with an empirical factor of 0.15 relative to the chemical shift-matching terms. The numbers w_{atom} give different weights to different atoms and are set to 1.0 for the backbone atoms N, H^N and CO, 1.2 for side-chain carbons, 1.5 for CA and CB, and 0 if the resonance is not assigned.

If independent information about the amino acid type is available from labeling by residue type, the likelihood for the intra-residue tine of the fork is set close to one for this amino acid type (precisely this was set to $(4 + \rho(aa|l))/5$). The values for all other residue types can be set to zero. In our current applications, however, we have kept the value at $\rho(aa|l)$ to allow for the possibility that there is spectral overlap or a slight chemical shift change between the selectively and uniformly labeled samples. In all proteins we have assigned with IBIS so far, labeling by amino acid type was not crucial and was only used to confirm the assignments.

Stage 5: Combination of information from threads and aa *type*

At this stage of the assignments, each thread created in Stage 3 is probed against the amino acid sequence to identify its position. For each fork of a thread the likelihood is calculated that it matches the amino acid type of a test sequence, using Equation 1. A composite likelihood is then determined that the thread matches a test sequence by combining the likelihoods for all forks of the thread:

$$\Pi\left(\text{thread}_i, \text{pos}_j\right) = \frac{\ln(n+1)}{n} \sum_{k=0}^{n-1} \frac{\rho(aa[\text{pos}_{j+k-1}]|S)+\rho(aa[\text{pos}_{j+k}]|I)}{2}. \quad (3)$$

The composite likelihood is essentially an average of the individual likelihoods with a logarithmic bonus weight for the length n of a thread. Each member of the sum contains the likelihoods for the S-tine and the I-tine. Subsequently, the resulting likelihoods of assigning a thread to the sequence, starting at position j, $\Pi(\text{thread}_i, \text{pos}_j)$, are sorted according to the scoring function as shown in Figure 4b. The first column gives the overall score as calculated by Equation 3. The numbers placed before and after the amino-acid names and sequence numbers indicate the consistency scores q of tines as assigned to amino acid types, with 9 being the highest and 0 being the lowest. Thus, the sequence, 7> ARG_29 <5,7> GLY_30 in the first line of Figure 4b means that the S-tine of the ARG_29 fork has a relatively high score of belonging to a tyrosine, which precedes ARG_29 in the primary sequence. The score for the I-tine of the ARG_29 fork is 5, and the score for the S-tine of GLY_30 is 7, etc.

At this point, IBIS has identified a set of threads, and for each threads there are different assignment options as shown in Figure 4b. IBIS now assigns first that thread to the primary sequence, that has the largest difference between the top-scoring assignment (at position p) and the next lower-scoring assignment, $\Delta_i(\text{p} - \text{s}) = \Pi(\text{thread}_i, \text{pos}_p) - \Pi(\text{thread}_i, \text{pos}_s)$, at starting position s. Starting position p is the prime top-scoring possible alignment; starting position s is the second highest scoring possible alignment for thread i. A 'possible alignment', is a placement of a thread on the primary sequence that does not overlap with a previously assigned segment. In addition, placements are considered 'not possible' if they contain an internal proline. IBIS hence starts the assignment with the thread that is most uniquely assigned. This is similar to manual assignment preferences. The assignment chosen in this way is saved in the format of Figure 4c. IBIS first assigns the thread with the largest $\Delta_i(\text{p} - \text{s})$. After the assignment of the first segment in the primary sequence is accepted this segment is marked as assigned and is no longer tested against unassigned threads. The procedure is repeated, and threads are assigned in the order of decreasing $\Delta_i(\text{p} - \text{s})$ values until all experimentally connected threads are used up.

The availability of the different possible assignments of threads to sequence positions as shown in Figure 4b provides the operator the chance to validate the assignments by searching for poor scores. As an example, the second-scoring assignment in Figure 4b has a relatively high overall score but the S-tine of ARG_51 scores poorly as connecting to a valine in position 50 and makes it an unlikely assignment. The operator has the option to eliminate such assignment options from the *ibis.n.*thread file.

Figure 4c contains another valuable diagnostic tool. While the first column lists the amino acids the thread was finally assigned to, the second column lists (in brackets) the amino acid sequence of the top-scoring assignment for a thread. Thus, it happens occasionally that the two columns do not agree. This usually raises a red flag indicating that there may be a problem with the assignment of this particular thread. Comparing the first two columns of *ibis.n.*results file has become the first step of results validation.

Stage 6: Display of ordered triple resonance strips, validation and iteration

At this stage, IBIS provides a file with sequentially ordered threads, and the file *ibis.n.*prot is created, which orders triple resonance strips in the XEASY program. An example of sequentially aligned triple resonance strips for the V1 domain of the protein PKC

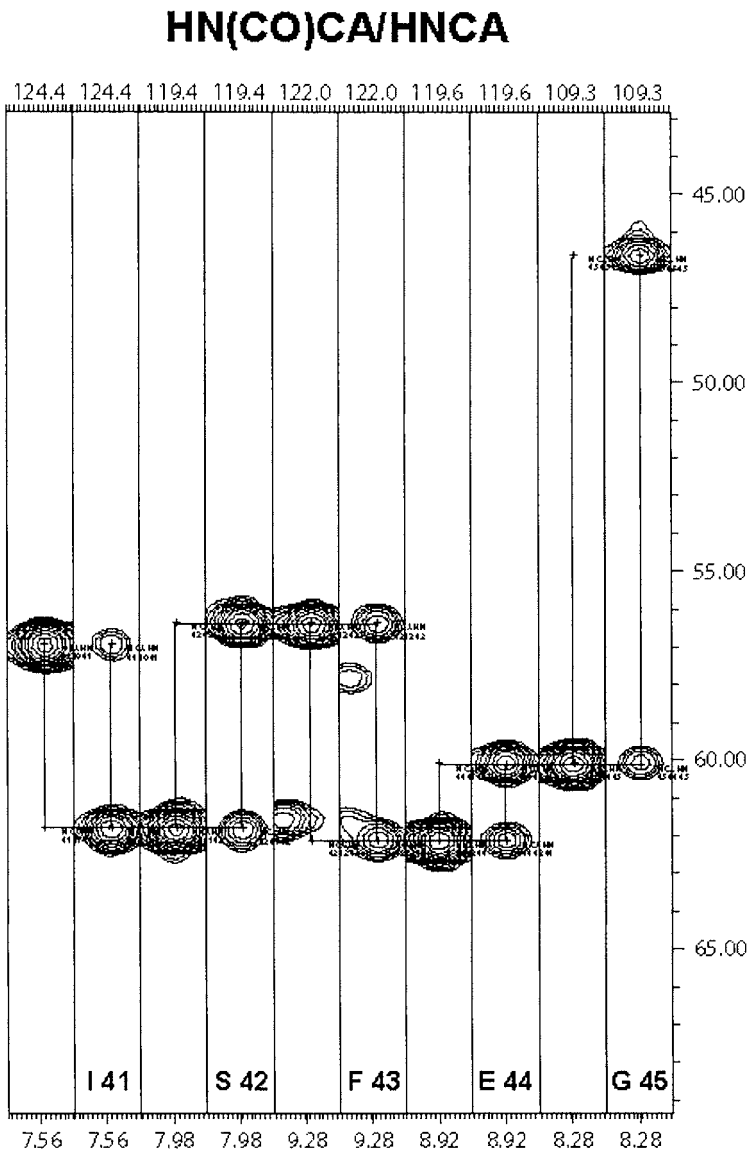

Figure 5. Example of pairs of aligned HN(CO)CA (left) and HNCA (right) triple resonance strips as obtained from the *Ibis.n.*prot file. The data were taken from the assignment of the V1 domain of PKC iota (Roehrl et al., 2003).

iota (Roehrl et al., 2003) is shown in Figure 5. The operator can now validate visually the quality of the assignments. If obvious inconsistencies come up, the peak lists can be corrected, or other assignments of threads to amino-acid sequences can be picked from the *ibis.n.*thread file, and IBIS is run again. For a protein of 10 to 20 kDa we typically run five to ten iterations. For proteins below 100 residues one to three iterations are usually sufficient.

The problem of resolving chemical shift degeneracy

IBIS *per se* does not resolve chemical shift degeneracies. Problems arise essentially only if both H^N and N chemical shifts are degenerate for two residues. In this case, the two S-tines and I-tines can be combined into two forks in two ways. IBIS can readily find the correct pairing in the following way: The program presents the data as interleaved strips and always threads the assignment via the one possibility where chemical shifts of the S-tine are closest to chemical shifts of any I-tine. If pairs of tines are combined to

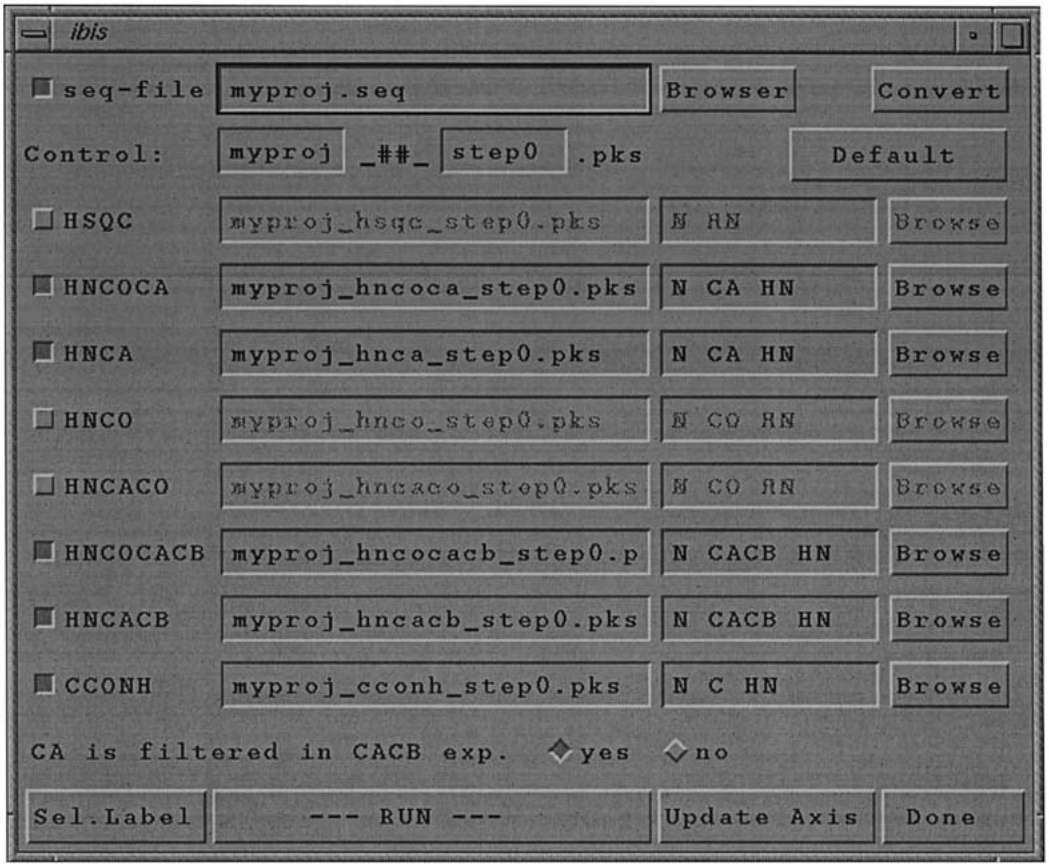

Figure 6. The IBIS user interface allows the operator to load data needed for the assignment process. The top row is used to enter the project name and the amino acid sequence. Peak lists are entered with project name, experiment description and the number of the iteration step. IBIS peak lists have the extensions '.pks' and can be converted into XEASY format by pressing the convert button. The second column of buttons defines the order in which the chemical shifts are organized.

forks in an incorrect way due to resonance degeneracy, IBIS will pick the wrong sequential connection. This will lead to low values of the *q*-sores (Figures 4b and 4c). If this is the case, the spectroscopist may simply first remove suspiciuous degenerate peaks, let IBIS run through another assignment iteration and assign the fragments that do not suffer from chemical shift degeneracies. At this round, only the safe assignments will be taken. Next, the eliminated peaks are reinserted, matching of tines to forks may be switched and another IBIS iteration is initiated. In our experience, the majority of cases of chemical shift degeneracy can be resolved in this manner.

User interface

Figure 6 displays IBIS' current user interface. It allows the input of the amino acid sequence and a variable number of experimental data. The first column represents the names of the peak lists in IBIS format (.pks). The lists can be converted into XEASY format by the convert key. Names of peak lists consist of the project name, experiment type and the iteration step. Experiments that can be accepted at this point include the ^{1}H-^{15}N HSQC and three pairs of triple resonance experiments. Also, the (H)CC(CO)NH TOCSY experiment can be loaded here. Information about labeling by residue type can also be entered through the user interface.

Conclusion

We have developed the IBIS tool that can achieve sequential assignments of protein spectra from triple resonance spectra. The program provides continuous visual access to the progress of the assignments and

allows a validation while the assignments are being achieved. The program is flexible with respect to the number of experiments used. Needless to say, IBIS cannot assign residues that do not exhibit triple resonance cross peaks, and assignments will never reach 100%. However, all assignments that a skilled user can obtain from visual inspections of triple resonance experiments are usually identified by IBIS but much faster. IBIS does not yet use NOE data, which are typically utilized to complete the assignments. In our hands, IBIS can be used to rapidly assign proteins in the range up to 20 kDa and validate assignments in a time frame of a few hours to a few days. The speed of the assignment process depends on the signal-to-noise ratio of the spectra and other complications, such as the presence of exchange broadening. IBIS will be a valuable tool for larger systems but the completeness of signals in triple resonance spectra usually deteriorates with the size of the protein and more operator intervention will be needed. A binary version of IBIS is available from the URL http://gwagner.med.Harvard.edu/ibis/

Acknowledgements

This work was supported by the National Institute of Health (GM47467 and RR-00995). We thank Drs J. Hoch, J. Sun, and M. Roehrl for fruitful discussions and for providing the data to demonstrate the use of IBIS.

References

Bartels, C., Xia, T., Billeter, M., Güntert, P. and Wüthrich, K. (1995) *J. Biomol. NMR*, **5**, 1–10.

Dötsch, V., Matsuo, H. and Wagner, G. (1996) *J. Magn. Reson.*, **B112**, 95–100.

Eccles, C., Guntert, P., Billeter, M. and Wüthrich, K. (1991) *J. Biomol. NMR*, **1**, 111–130.

Friedrichs, M.S., Mueller, L. and Wittekind, M. (1994) *J. Biomol. NMR*, **4**, 703–726.

Hare, B.J. and Prestegard, J.H. (1994) *J. Biomol. NMR*, **4**, 35–46.

Ikura, M., Kay, L.E. and Bax, A. (1990) *Biochemistry*, **29**, 4659–4667.

Johnson, B.A. and Blevins, R.A. (1994) *J. Biomol. NMR*, **4**, 603–614.

Kay, L.E., Ikura, M., Tschudin, R. and Bax, A. (1990) *J. Magn. Reson.*, **89**, 496–514.

Kraulis, P.J. (1994) *J. Mol. Biol.*, **243**, 696–718.

Leutner, M., Gschwind, R.M., Liermann, J., Schwarz, C., Gemmecker, G. and Kessler, H. (1998) *J. Biomol. NMR*, **11**, 31–43.

Lukin, J.A., Gove, A.P., Talukdar, S.N. and Ho, C. (1997) *J. Biomol. NMR*, **9**, 151–166.

Meadows, R.P., Olejniczak, E.T. and Fesik, S.W. (1994) *J. Biomol. NMR*, **4**, 79–96.

Montelione, G.T. and Wagner, G. (1989) *J. Am. Chem. Soc.*, **111**, 5474–5475.

Montelione, G.T. and Wagner, G. (1990) *J. Magn. Reson.*, **87**, 183–188.

Moseley, H.N., Monleon, D. and Montelione, G.T. (2001) *Meth. Enzymol.*, **339**, 91–108.

Nelson, S.J., Schneider, D.M. and Wand, A.J. (1991) *Biophys. J.*, **59**, 1113–1122.

Olson, Jr. J.B. and Markley, J.L. (1994) *J. Biomol. NMR*, **4**, 385–410.

Oschkinat, H. and Croft, D. (1994) *Meth. Enzymol.*, **239**, 308–318.

Oschkinat, H., Holak, T.A. and Cieslar, C. (1991) *Biopolymers*, **31**, 699–712.

Ou, H.D., Lai, H.C., Serber, Z. and Dötsch, V. (2001) *J. Biomol. NMR*, **21**, 269–273.

Roehrl, M.H.A., Hyberts, S.G., Sun, Z.-Y.J., Fields, A.P. and Wagner, G. (2003) *J. Biomol. NMR*, **26**, 373–374.

Schubert, M., Oschkinat, H. and Schmieder, P. (2001a) *J. Biomol. NMR*, **20**, 379–384.

Schubert, M., Oschkinat, H. and Schmieder, P. (2001b) *J. Magn. Reson.*, **148**, 61–72.

Seavey, B.R., Farr, E.A., Westler , W.M. and Markley, J.L. (1991) *J. Biomol. NMR*, **1**, 217–236.

Wagner, G. and Wüthrich, K. (1982) *J. Mol. Biol.*, **155**, 347–366.

Wüthrich, K., Wider, G., Wagner, G. and Braun, W. (1982) *J. Mol. Biol.*, **155**, 311–319.

Xu, J., Straus, S.K., Sanctuary, B.C. and Trimble, L. (1994) *J. Magn. Reson.*, **B103**, 53–58.

Zimmerman, D., Kulikowski, C., Wang, L., Lyons, B. and Montelione, G.T. (1994) *J. Biomol. NMR*, **4**, 241–256.

Zimmerman, D.E. and Montelione, G.T. (1995) *Curr. Opin. Struct. Biol.*, **5**, 664–673.

Zimmerman, D.E., Kulikowski, C.A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C., Powers, R. and Montelione, G.T. (1997) *J. Mol. Biol.*, **269**, 592–610.

Zimmerman, D.E., Kulikowski, C.A. and Montelione, G.T. (1993) *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **1**, 447–455.