Journal of Biomolecular NMR, 4 (1994) 79–96
ESCOM

J-Bio NMR 152

A computer-based protocol for semiautomated assignments and 3D structure determination of proteins

Robert P. Meadows, Edward T. Olejniczak and Stephen W. Fesik*

Pharmaceutical Products Division, Abbott Laboratories, Abbott Park, IL 60064, U.S.A.

Received 25 May 1993
Accepted 12 August 1993

Keywords: Nuclear magnetic resonance; Automated spectral analysis; NMR solution structure

SUMMARY

A strategy is presented for the semiautomated assignment and 3D structure determination of proteins from heteronuclear multidimensional Nuclear Magnetic Resonance (NMR) data. This approach involves the computer-based assignment of the NMR signals, identification of distance restraints from nuclear Overhauser effects, and generation of 3D structures by using the NMR-derived restraints. The protocol is described in detail and illustrated on a resonance assignment and structure determination of the FK506 binding protein (FKBP, 107 amino acids) complexed to the immunosuppressant, ascomycin. The 3D structures produced from this automated protocol attained backbone and heavy atom rmsd of 1.17 and 1.69 Å, respectively. Although more highly resolved structures of the complex have been obtained by standard interpretation of NMR data (Meadows et al. (1993) *Biochemistry*, **32**, 754–765), the structures generated with this automated protocol required minimal manual intervention during the spectral assignment and 3D structure calculations stages. Thus, the protocol may yield an approximate order of magnitude reduction in the time required for the generation of 3D structures of proteins from NMR data.

INTRODUCTION

Originally developed in the early 1980s (Wider et al., 1982; Wüthrich et al., 1982; Wüthrich, 1983, 1986), NMR-based protein structure determination has been used to elucidate the 3D structures of hundreds of small proteins (for reviews see Bax, 1989; Clore and Gronenborn, 1989). In the first step of the three-stage process, the NMR signals are assigned to the individual nuclei of the protein. This is accomplished by correlating the NMR resonances of the amino acid spin systems by using scalar connectivities, identifying the spin systems by amino acid types, and sequentially assigning the spin systems from either ^{1}H-^{1}H NOESY data or from correlations observed in newly developed triple-resonance correlation experiments (Ikura et al., 1990; Kay et al., 1990b). Next, distance and angular restraints are derived from analysis of the NOESY data

*To whom correspondence should be addressed.

and vicinal three-bond coupling constants, respectively. Finally, these restraints are used in a calculational strategy (Havel and Wüthrich, 1985; Nilges et al., 1988; Scheek et al., 1989) to produce the 3D structures.

Although many protein structures have been solved by the NMR method, the process of NMR-based structure determination is still tedious and time consuming. For proteins with molecular weights > 10 kDa, analysis of the NMR data often requires several months of both computational and human resources. One way to reduce the time required for NMR-based structure determinations is to implement computer algorithms that perform the mundane and monotonous chores of spectral analysis. Among the aspects of NMR spectral analysis that have been subjected to computer automation are peak-picking (Pfändler et al., 1985; Glaser and Kalbitzer, 1987; Hoch et al., 1987; Stoven et al., 1989; Garret et al., 1991) and the identification of spin systems by amino acid types (Neidig et al., 1984; Pfändler et al., 1985; Novic et al., 1987; Grahn et al., 1988; Weber et al., 1989; Eccles et al., 1991). Once spins systems have been correctly identified, computer strategies have been developed that make use of extensive bookkeeping and consistency checking (Billeter et al., 1988) to obtain the sequential assignments. A number of these methods have been encoded into computer programs (Cieslar et al., 1988; Eads and Kuntz, 1989; Kleywegt et al., 1989; Kraulis, 1989; Catasti et al., 1990; Van de Ven, 1990; Eccles et al., 1991; Bernstein et al., 1993) that permit interactive or semiautomatic resonance assignment.

For unlabeled proteins, sequential assignments are made from NOE (through space) interactions. However, for isotopically labeled proteins, new heteronuclear triple-resonance experiments allow sequential assignments of protein backbone resonances from scalar (through bond) connectivities. These new data offer a more direct, and hence, less ambiguous assignment method. Recently, a semiautomated procedure for assigning the backbone resonances in proteins from triple resonance data was described (Ikura et al., 1990). This work demonstrated the viability of heteronuclear NMR data for automatic sequential assignments, an important initial step of a more fully automated assignment procedure. However, a complete assignment protocol which includes the side chain resonance assignments and automated identification of structural restraints from NOESY data would be highly desirable.

In this report we describe a protocol for the semiautomated generation of backbone, side chain, and NOE assignments from heteronuclear multidimensional NMR data. The protocol is illustrated by generating the assignments and 3D structure of the FK506 binding protein (FKBP, 107aa) (Harding et al., 1989; Siekierka et al., 1989) when bound to the immunosuppressant, ascomycin (Arai et al., 1962; Hatanka et al., 1988). With this approach, the time required to produce high-resolution 3D structures from NMR data is dramatically reduced.

MATERIALS AND METHODS

Sample preparation

Recombinant human FKBP was cloned from a Jurkat T cell cDNA library and expressed at high levels in *E. coli* using the translational coupling to the 5' end of the *E. coli* kds B gene (Pilot-Matias et al., 1993). [U-^{15}N]FKBP and [U-^{13}C,^{15}N]FKBP were purified from cells grown on a minimal medium containing [^{15}N]ammonium chloride alone or in combination with [U-^{13}C]acetate (Venters et al., 1991), using ion exchange and size exclusion chromatography (Edalji et al., 1992). The protein was concentrated to ~ 3 mM and exchanged into a H_2O or ^2H_2O solution

(pH = 6.5) containing potassium phosphate (50 mM), sodium chloride (100 mM), and dithiothreitol-d_{10} (5 mM). FKBP/ascomycin complexes (1/1) were prepared by incubating labeled FKBP with an excess amount of unlabeled ascomycin for 24–48 h at room temperature. The excess ascomycin was removed by centrifugation.

NMR

All NMR spectra were collected at 30 °C on a Bruker AMX600 (600 MHz) spectrometer. NMR spectra were processed and analyzed using in-house written software on Silicon Graphics computers. The ^{15}N-resolved 3D NOESY-HSQC (Fesik and Zuiderweg, 1988; Marion et al., 1989), 4D [^{13}C,^{1}H,^{15}N,^{1}H] NOESY (Kay et al., 1990a), 4D [^{13}C,^{1}H,^{13}C,^{1}H] NOESY (Clore et al., 1991; Zuiderweg et al., 1991), and 2D [^{1}H,^{15}N] HSQC (Bodenhausen and Ruben, 1980) experiments were acquired and processed as described in Meadows et al. (1993). The ^{13}C spectral widths in the 4D [^{13}C,^{1}H,^{13}C,^{1}H] NOESY experiment were folded in the same manner as the 4D [^{13}C,^{1}H,^{15}N,^{1}H] NOESY. The phase ramps for the two indirectly detected carbon dimensions in the 4D [^{13}C,^{1}H,^{15}N,^{1}H] NOESY experiment were 360° and 180°. The 4D [^{13}C,^{1}H,^{15}N,^{1}H] HNCAHA and HN(CO)CAHA spectra were collected and processed as previously described (Olejniczak et al., 1992). The 4D [^{13}C,^{1}H,^{15}N,^{1}H] HC(CO)NH-TOCSY spectrum was acquired by using the pulse sequence and parameters as described in Logan et al. (1992).

Structure refinement

Structure calculations employed a distance geometry/dynamical simulated annealing (DG/SA) method. Calculations and structural analysis used the program X-PLOR 3.1 (Brünger, 1992) operating on a Silicon Graphics 4D-360. Input to the DG protocol included the covalent geometry, NOE-derived distance restraints, and specific hydrogen bond restraints. The resulting structures were refined by high-temperature (2000 K) restrained molecular dynamics followed by dynamical simulated annealing (Nilges et al., 1988). Distance restraints were applied with force constants of 50 kcal · mol^{-1} · Å^{-2} and were given upper bounds of 3.5 Å for backbone–backbone restraints and 5.0 Å for all other NOE-derived distances. Corrections for center averaging (Wüthrich et al., 1983) were applied to the upper bounds where appropriate. Hydrogen bond restraints were given bounds of 1.8–2.3 Å (H → O), and 2.5–3.3 Å (N → O). Stereospecific assignments of the methyl groups were obtained from manual inspection of the HSQC spectrum of partially ^{13}C-labeled protein, via the method of Wüthrich and coworkers (Neri et al., 1989; Senn et al., 1989). No other stereospecific assignments were used, and no dihedral angle restraints were applied to the protein during any stage of the refinement. In addition, since the NMR data was collected on the FKBP/ascomycin complex, 36 intermolecular and 88 intra-ligand NOE restraints were included during the structure refinement (Petros et al., 1991). These data were obtained and analyzed previously in the usual manner.

Software

Included below are descriptions of the FORTRAN-77 programs and subroutines that were used in this work. The names of the routines are listed, followed by their functional descriptions.

PEAK-PICK. The PEAK-PICK algorithm identifies the positions and calculates the integrated volumes of the cross peaks in an NMR spectrum. In order to minimize the effects of streaks and other experimental artifacts, peak-picking begins by dividing the full NMR spectrum

into small overlapping sections called blocks. For each of these blocks, an average value of the noise level is obtained by randomly selecting points in the block and calculating their average and variance. In successive loops only data points whose squared deviation from the average is less than eight times the variance of the previous loop are retained. The noise level is determined by using three loops, each loop sampling 5% of the points in the block. In addition, the number of erroneous peak-picks was reduced in the heteronuclear triple-resonance experiments by restricting the searches to a defined set of possible [^{15}N,NH] frequencies obtained from a high-resolution [^{15}N,NH] HSQC spectrum. Such 'directed peak-picking' results in the identification of a larger number of valid resonances through the use of low threshold levels.

Locations of the peaks are determined by finding local n-dimensional extrema. For a point to be identified as a peak maximum, a volume and height criteria must be satisfied. Every peak must have an n-dimensional volume for a 3^n cube that is greater than 3n times the standard deviation of the noise in the block. Peak positions are improved in each dimension by parabolic interpolation of the intensity of the surrounding points. The final n-dimensional positions are stored at twice the digital resolution of the spectrum, since for well-resolved peaks, the interpolation of peak position is very accurate.

Minor adjustments are made to peak-pick lists before beginning the computerized protocol. Manual inspection of these data is facilitated by a graphical data analysis package that has been developed in-house. The analysis program displays cross peaks from the peak list as boxes overlaid onto contour plots of the NMR data. Peak-picks can then be entered or deleted by 'clicking' on the box. Positions of the new peak-picks can be optimized using the same procedure as the automated peak picking. Small, systematic differences in the chemical shifts observed in the different experiments were compensated for by the use of a computer program that adjusts the frequencies of cross peaks in a multidimensional peak-pick list.

The experiment spectral widths and phase of the cross peaks are used to calculate the correct, unfolded cross peak chemical shifts. In practice only carbon chemical shifts are folded and all folded resonances are defined by both a carbon and proton resonance frequency. The experimental sweep widths are chosen so that the aliphatic resonances are folded no more than once into the observed sweep width (i.e., experimental sweep width = 1/3 the aliphatic carbon frequency range). Since the phase ramp for the folded dimension of interest is chosen to be 180°, the folded peaks have the opposite sign of nonfolded peaks. In experiments where two carbon chemical shift dimensions are obtained, the diagonal peak dimension is given a 360° phase ramp while the cross peak dimension has a 180° phase ramp. Given these experimental conditions, carbon resonances are unfolded by using the sign of the resonance and its observed proton chemical shift. If a peak is folded and its proton chemical shift is < 3.0 ppm, then the resonance is unfolded by subtracting the sweep width from the observed peak position. For proton chemical shifts > 3.0 ppm the sweep width is added to the observed position. This procedure gives the correct result for nearly all resonances observed in amino acids.

RESTYPE. Possible residue types are assigned to the experimental spin systems by comparing the observed side chain resonances to values that define both the expected and excluded chemical shifts for each of the 20 common amino acids (Grob and Kalbitzer, 1988; Olejniczak et al., 1992). Assignment of residue type is based upon an empirical scoring system as illustrated in Fig. 1. For each resonance in the experimental group that agrees with the mask for a particular amino acid, a value of 1 is added to the score for that amino acid. Resonances that disagree with

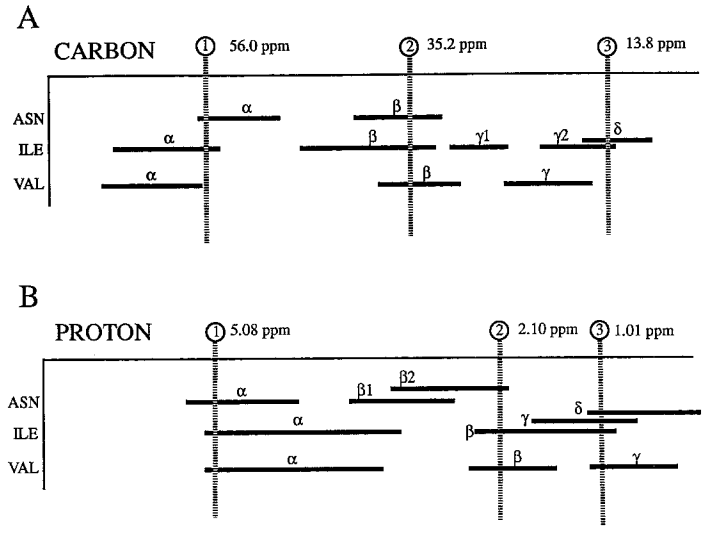

Fig. 1. Schematic representation of the residue typing procedure. For clarity, data is shown for only three amino acids. ^{13}C and ^{1}H chemical shift masks for Asn, Val, and Ile amino acids are depicted as horizontal lines and are compared to a representative set of experimentally observed resonances (circles labeled 1, 2, 3). Chemical shift ranges were obtained from Olejniczak et al., (1992) and Grob and Kalbitzer, (1988). Empirical scores (see text) for the three amino acids are −1 (Asn), +2 (Val), and +8 (Ile).

the mask are penalized by −1. Because methyl groups have very distinct chemical shifts, they are given a higher score (+3) if positively identified, and a lower score (−3) if missing. Several residues have characteristic features and are scored based upon the presence or absence of these features. In addition, the topology of the residue is also considered in the scoring. For example, the maximum score for any resonance type cannot exceed the number of resonances in the amino acid and carbon chemical shift degeneracy is required for the methylene groups.

The number of resonances with high (< 50 ppm), mid-range (37–25 ppm) and degenerate carbon chemical shifts as well as the number of resonances with low (> 2.0 ppm) proton chemical shifts are tabulated and this information is also used for scoring. Specifically, the scores for Thr and Ser are increased by +1 if two or more resonances have high carbon chemical shifts. The scores for Arg, Ile, Leu and Lys are increased if two or more low proton chemical shifts are observed. The scores for Asp, Tyr and Phe are increased by +1 while Glu, Gln, Arg and Pro are penalized by −1 if no mid-range shifts are observed. Likewise, Ile and Val are penalized by −3 if no carbon chemical shift characteristic of methyl groups are observed. At the end of scoring, the nonnegative scores are sorted. If the difference between the top two scores is > 5, then only those amino acids with the top score are included into the list of possible residue types. If the difference is < 5, all residue types with the top two scores are retained. If both proton and carbon chemical shifts are available, then scoring is based on a comparison to both the proton and carbon shift ranges for each resonance observed in an amino acid. In this case, scoring is done pairwise so that a match for a carbon chemical shift must be reflected by a match for it's attached proton. If either carbon or proton fail then the score for that residue is decreased.

Special consideration is given for glycine residues and amino acids with overlapping ^{15}N and

NH chemical shifts. Both of these special cases are identified automatically from the 3D HN(CO)CA data. In the first case, glycines can be uniquely identified by their characteristic $^{13}C^{\alpha}$ chemical shift (between 42 and 48 ppm). If a resonance is observed at this frequency, and the [^{15}N, NH] is not degenerate, then only glycine is included into the list of possible residue types. Residues with overlapping amides are recognized by observing multiple [$^{13}C^{\alpha}$, H^{α}] resonances at a given [^{15}N, NH] frequency. If this occurs, no penalties (i.e., negative scores) are applied for the presence of unexpected peaks and the amino acids with the top four scores are retained in the possible residue type list.

DIPEPTIDE. Residues that are adjacent in the primary sequence are identified by using the DIPEPTIDE algorithm. As shown in Fig. 2A, the algorithm begins by superimposing the [^{15}N,NH] frequencies observed in both the HN(CO)CAHA and the HNCAHA experiments. This allows the $^{13}C^{\alpha}$ and H^{α} frequencies for residue (i) to be differentiated from those of the (i − 1) residue. Next (Fig. 2B), the $^{13}C^{\alpha}$ and H^{α} frequencies are used to search the HN(CO)CAHA data for the (i + 1) [^{15}N,NH] resonances.

LINK. Using this algorithm, dipeptides are joined into longer fragments by identifying peptides with common [^{15}N, NH, $^{13}C^{\alpha}_{(i-1)}$, $H^{\alpha}_{(i-1)}$] resonances. The list of possible adjacent residues is extended by using the n-peptide list to find two n-peptides with (n − 1) overlapping and common [^{15}N, NH, $^{13}C^{\alpha}_{(i-1)}$, $H^{\alpha}_{(i-1)}$] resonances. This gives (n + 1) adjacent [^{15}N, NH, $^{13}C^{\alpha}_{(i-1)}$, $H^{\alpha}_{(i-1)}$] resonances whose possible residue type can be compared to the known protein sequence. All of the (n + 1) peptide fragments which have residue types that are consistent with the protein sequence are retained.

SEQCHECK. The residue types for a n-peptide fragment are checked for consistency with the protein sequence by using the SEQCHECK algorithm. All of the possible combinations of residue types for each residue of the peptide fragment are tested. If none of the possible combinations match a section of the known protein sequence, then the peptide fragment is eliminated. Amino acids such as prolines, which do not have amide protons, do not give side chain informa-

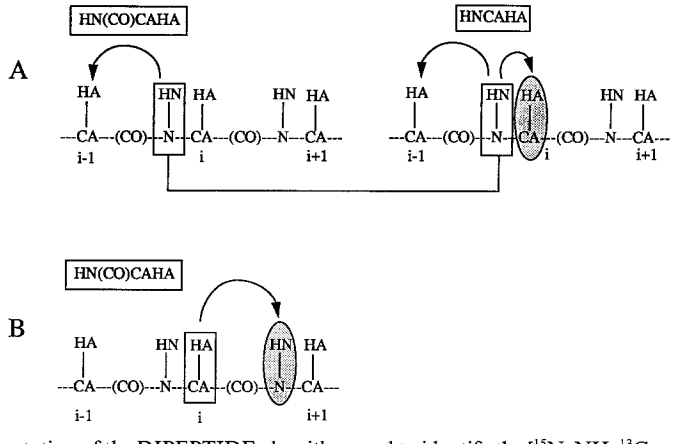

Fig. 2. Schematic representation of the DIPEPTIDE algorithm used to identify the [^{15}N, NH, $^{13}Ca_{(i-1)}$, $Ha_{(i-1)}$] resonances of two adjacent residues. The two-step process begins as shown in (A), where identical [^{15}N,NH] resonances (rectangular boxes) from HN(CO)CAHA and HNCAHA experiments are used to determine the intraresidue [H^{α},C^{α}] resonance (oval). Next, this resonance set is used in step (B) to match identical [H^{α},C^{α}] resonances to find the [^{15}N, NH] resonance of the (i + 1) residue.

tion for the (i − 1) residue in a 4D HC(CO)NH-TOCSY experiments. These gaps are taken into account when comparing the possible matches of the fragment to the protein sequence. Data from selective labels are also checked at this stage. If the frequencies of one of the residues in the peptide fragment uniquely matches those of an exclusive label, then only assignments consistent with this exclusive label are retained.

ATOMASSIGN. The final, atom-specific [^{13}C,^{1}H] assignments are determined after the sequential assignment process is completed. The input to the routine is the amino acid type and the set of observed resonances. The routine uses the chemical shift ranges from the appropriate amino acid mask to assign the individual resonances. In the first step, only resonances that uniquely fit a single chemical shift range are assigned. The remaining resonances are then allowed to match the remaining atom types. If a resonance does not fall within an allowable chemical shift range, it is not assigned. In cases of spectral overlap, only resonances sets that have unique chemical shift ranges for both of the degenerate residues are assigned.

NOEASSIGN. Possible NOE assignments are made by a simple algorithm which matches frequencies from a chemical shift database to the frequencies of peak-picks from a NOESY spectrum. Input includes the chemical shift database, the digital resolution of each dimension, any systematic differences observed between the experimental NOESY and the database, the number of allowed matches, and the error allowed for each frequency match. The output consists of the list of possible assignments for each cross peak in the NOE peak-pick file ranked in order of increasing cumulative error.

XYZFILTER. Once a structure is available, the XYZFILTER algorithm effectively uses the calculated interatomic distances to remove ambiguities in the list of possible NOE assignments. For example, when an NOE has more than one possible assignment, then all of the possibilities are evaluated by checking their distances in the current structure. If only one of the combinations has an interproton distance < 6–8 Å, then it is that single, conformationally correct assignment that is assumed valid and included into the current NOE restraint list.

PRECHECK. A simple check of the accuracy of the NOE assignment list is obtained by comparing the NOE assignments to the predicted secondary structure. The PRECHECK algorithm first predicts the secondary structure from the α-proton chemical shifts by making use of the chemical shift index, or CSI (Wishart et al., 1992). It then checks the NOE assignments against the secondary structure and eliminates unlikely restraints that are observed within a given secondary structural element.

REDUNDANCY. An additional check for accuracy in the restraint list is obtained from a simple redundancy filter. In essence, the redundancy filter retains only those restraints that are corroborated by other restraints. Correctness is assumed if any of the atoms in that same residue or atoms in sequential neighbors have NOE contacts to the residue receiving the NOE or its sequential neighbors. The amount of corroborating data required to verify a particular NOE-derived restraint is adjusted to reflect the confidence in the spectral quality. Obviously, use of this filter in the initial stages of analysis may eliminate correct restraints; however, this is preferable to including interproton distances derived from incorrectly assigned NOEs. If incorrect restraints are included in the initial folding stages, the structures, which act as seeds for the remaining NOE assignments, will ultimately lead to conformations which become further and further removed from the actual conformation.

FINDHBOND. A list of possible hydrogen bond acceptors is generated from the Cartesian

coordinates for those backbone amides that undergo slow deuterium exchange. User-defined criterion control the selection of hydrogen bond acceptors. Typically, to be included as a hydrogen bond, the heavy atom O–N distance must be within 4.00 Å and the NH–O angle must fall within 180° ± 40°. To be included as a restraint, a hydrogen bond must be observed in > 66% (user defined) of the calculated 3D structures, and in those conformations where the hydrogen bond was not observed, the amide must not be hydrogen bonded to a different acceptor. By using these criterion, several unambiguous hydrogen bond pairs may be generated at each iteration.

RESULTS AND DISCUSSION

Sequential and specific resonance assignments

A flowchart (Fig. 3) shows the protocol used for obtaining the chemical shift assignments. Input for the chemical shift assignment procedure includes the peak-picks from the various multidimensional NMR experiments, the protein sequence, tolerances for allowable variances in the chemical shift of peaks, and any additional information supplied by the user such as the [^{15}N,NH] shifts from selective labels or previously determined assignments.

After peak-picking, residue types are assigned for each [^{15}N,NH] pair by using the routine RESTYPE. The residue typing process is shown schematically in Fig. 1. The signals observed for a residue are shown at the top of the figure, and the chemical shift ranges for three amino acids are shown below. If the chemical shift of an observed resonance intersects both the carbon and proton chemical shift range for an amino acid resonance, then the score for that amino acid is increased. If it is outside the observed chemical shift range, then the score is decreased. Using all of the available chemical shift information improves the assignment of resonance type. In this example, the proton chemical shifts alone are not able to differentiate between Ile and Val. Likewise, using only the C^{α} and C^{β} chemical shifts, one cannot differentiate between Ile and Asn. Because methyl scores are weighted more heavily relative to nonmethyls, the scoring protocol treats Ile and Val as special cases. For these two amino acids, an experimental resonance must match at least one of the expected methyl group chemical shifts, otherwise the score for the amino acid is decreased by three. This results in the very different scores of −1 (Asn), +2 (Val), and +8 (Ile) for these three amino acids. Based upon the scoring protocol, the residue type most consistent with these experimental data is Ile. Importantly, this may be accomplished even though only three of a possible six Ile resonances were observed.

The next step of the assignment procedure produces a list of possible adjacent spin systems that are identified from the HN(CO)CAHA and HNCAHA triple resonance experiments using the routine DIPEPTIDE (see Fig. 2). The process involves matching the [^{15}N,NH] resonance frequencies observed in the two experiments to identify the [$^{13}C^{\alpha}$,H^{α}] resonance of the adjacent residue (Fig. 2A). This resonance set is then used to identify the [^{15}N,NH] of the next residue (Fig. 2B). The dipeptide fragments are extended to longer peptide fragments by using the routine LINK which aligns common [^{15}N, NH, $^{13}C^{\alpha}_{(i-1)}$, $H^{\alpha}_{(i-1)}$] resonances of shorter fragments into peptide fragments of one residue more. Each of these longer peptides is compared for consistency to the protein sequence by using the routine SEQCHECK. The peptides are iteratively extended until lists of all of the possible tri- to hexapeptide fragments are obtained.

The final sequential assignments are then made by aligning the various peptide fragments by using the routine SEQCHECK. At first, all of the hexa- and pentapeptide fragments are aligned

into the sequence. These larger fragments are generally placed into the sequence uniquely. Inconsistencies can occur for the assignment of residues at the ends of these peptide fragments. These misassignments occur because the end residues are only constrained by the residue type of one adjacent residue. These are typically resolved by considering the assignments of overlapping peptide sequences where the residues in question are in the middle of the peptide fragments and are thus constrained by the residue typing information of two adjacent residues. Gaps that remain in the sequential assignments are resolved with progressively smaller peptide fragments. For each unassigned residue, peptide fragments are found that align into the sequence and do not disagree with the 'best' amide assignments made with a longer fragment. The best assignments from each iteration are stored before going to the next iteration with a shorter peptide fragment. The smallest fragment used in this process are tripeptides.

After the sequential assignments are made, the side chain resonances are assigned to specific atoms by ATOMASSIGN. This involves comparing each observed resonance to the allowed

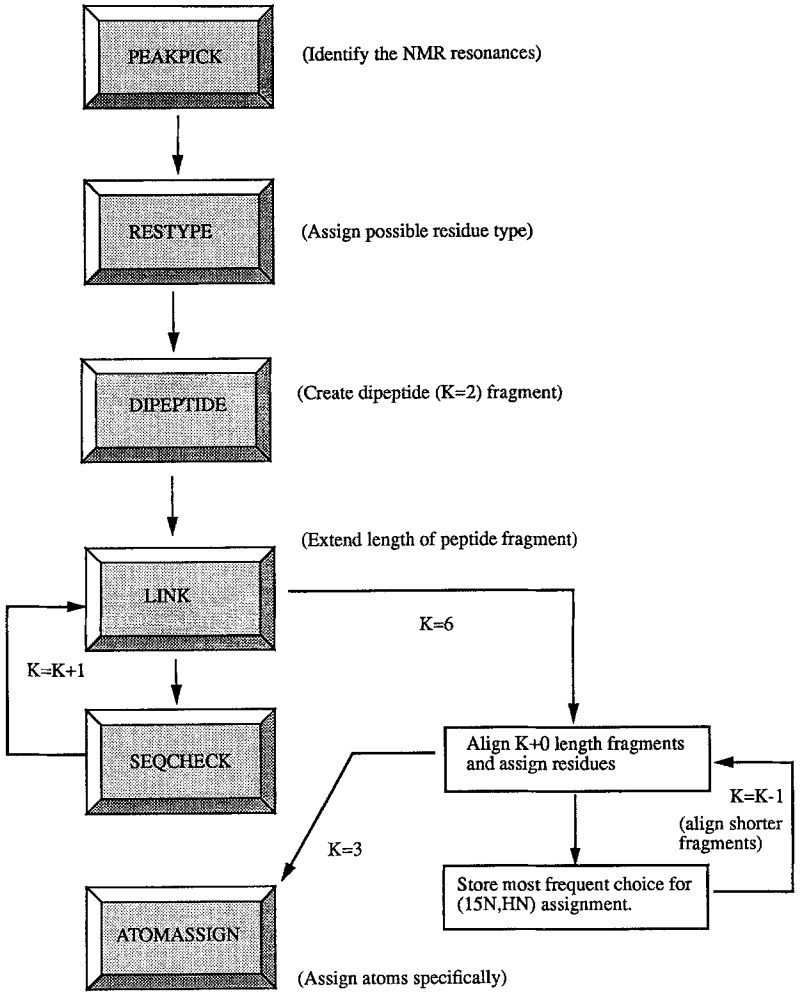

Fig. 3. Flow chart showing the chemical shift assignment strategy.

chemical shift ranges in the shift mask. If a reasonance is within the observed range for its assigned amino acid, it is assigned that atom type. Side chain resonances defined by degenerate [^{15}N,NH] frequencies are only assigned if the chemical shift range for that resonance is distinct. For residues next to prolines, only backbone resonances are defined. The remaining side chain resonances for these residues need to be identified from other experiments (e.g., a 4D HCCH-TOCSY).

NOE assignments

A stepwise procedure for obtaining NOE assignments and distance restraints is illustrated in Fig. 4. The first step in the protocol is to use the completed chemical shift database to identify unambiguously possible NOE assignments by NOEASSIGN. Optimum performance of this routine is only obtained if the chemical shift database from the automated assignment procedure is complete and consistent with the chemical shifts observed in the NOESY experiments. Systematic deviations between the NOE and the database are easily corrected, yet random fluctuations on the order of several data points (due to pH, temperature, isotope shifts, etc.) are more problematic and may be partially overcome with the use of a 'dynamical' database (Eccles et al., 1991).

The assignments from NOEASSIGN should be checked by PRECHECK, which checks assign-

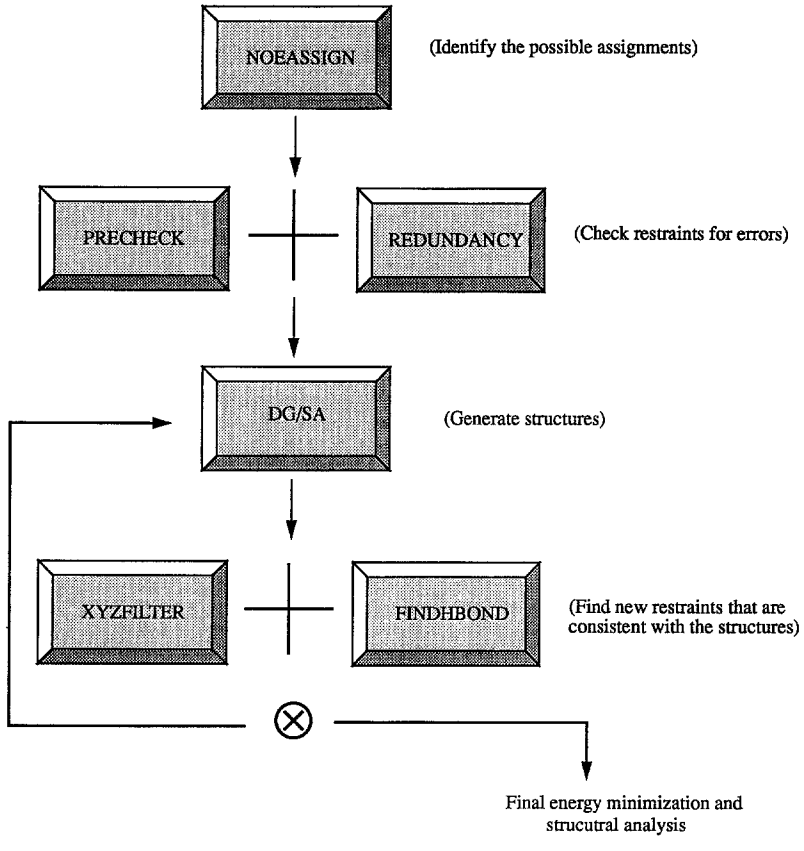

Fig. 4. Flow chart showing the NOE assignment strategy.

ments against the predicted secondary structure, and also by REDUNDANCY, which searches for corroborating NOEs. Checking the consistency of assignments is important to identify incorrect restraints. In practice, NOEASSIGN may produce an incorrect NOE assignment if atoms are missing from the chemical shift database or if significant chemical shift differences exist between the NOE data and the chemical shift database. It is our experience that for such 'noisy' data, 3–5 corroborating neighbors are sufficient to remove incorrect NOE assignments. If a complete and correct assignment set is available, then consistency checking with PRECHECK and REDUNDANCY may not be necessary. After assignment, the unambiguous NOEs are converted into distance restraints which are used in the structure refinement. The structures produced are then used to assign additional NOE restraints (using XYZFILTER) that satisfy both the chemical shift and interatomic distance criteria. In addition, if one has data on the exchange rates of the backbone amides, a list of probable hydrogen bonding partners may be obtained at each iteration by the program FINDHBOND. The cycle is repeated, typically 2–5 times, until no new assignments are added to the NOE list. At the end of the final iteration, one has determined a self-consistent set of NOE assignments that match the chemical shifts and produce low-energy conformations.

EXPERIMENTAL APPLICATION

Chemical shift assignments

The assignment protocol was evaluated by using experimental data obtained on the FKBP/ascomycin complex. The input data included all of the observed correlations from the 4D HNCAHA and HN(CO)CAHA experiment. This included a complete set of correlations to all residues except prolines, Ala^{84}, Met^{49}, and Leu^{97}. In addition, cross peaks for more than 85% of the side chain atoms were obtained from the 4D HC(CO)NH TOCSY experiment. To insure that resonances were not missing due to the mixing time dependence of the TOCSY experiments, two different 4D data sets obtained by using mixing times of 9.8 and 19.6 ms were added and used in the analysis. Many of the [^{15}N,NH] resonances of glycine residues were identified unambiguously from the 3D HNCA and HN(CO)CA data.

In the first step of the procedure, side chain resonance information was used to assign possible residue types for the (i − 1) residue of every [^{15}N,NH] resonance. Nearly all of the residues adjacent (i + 1) to Gly spin systems were identified unambiguously from the 3D HNCA and HN(CO)CA data. For the remaining data, an average of four different amino acids were identified as possible residue types for the residue preceding each [^{15}N,NH] resonance set. Because of the unique nature of their side chain chemical shifts, several Val, Ile and Thr residues were unambiguously identified at this time. In all cases, the correct residue type was among the list of possible amino acids. Next, dipeptides were generated which were then linked together into longer peptide fragments. During the linking stage several incorrect linkages were introduced for residues with nonunique [$^{13}C^{\alpha}$,H^{α}] resonances. However, these misassignments were eliminated by comparing the sequences of the possible peptides to the known protein sequence using SEQCHECK.

Alignment of shorter peptide fragments into the sequence is less accurate than when longer fragments are used. For example, tripeptide fragments were correct only 90% of the time while tetrapeptide fragments were correct 98% of the time. Errors in the assignment of peptide frag-

ments occurred most often next to residues, such as prolines, that have missing correlations. For FKBP, the incorrect choices that occurred were all eliminated by the unambiguous assignment of one of the choices in a longer peptide fragment. FKBP can be assigned correctly by using only tetra- and tripeptide fragments.

In the final step of the protocol, the individual NMR resonances were assigned to specific atom types. This was accomplished using the sequential assignment data and the amino acid masks. Although a small number of side chain assignment errors occurred at this stage for side chain resonances with overlapping chemical shift ranges (e.g., Lys H^{β} and H^{γ}), these errors did not significantly affect the quality of the resulting structures.

In summary, from this first stage of the semiautomated procedure, more than 90% of the resonances of FKBP were assigned (Fig. 5). These assignments were obtained in less than 1 CPU minute on a Silicon Graphics Personal Iris computer. Manual completion of the chemical shift database was necessary before continuing to the NOE assignment stage. Most of the missing assignments were obtained from analysis of HCCH-TOCSY data. Aromatic resonances were identified from TOCSY spectra and were sequentially assigned by observing unambiguous NOEs between aromatic protons and other protons of the same amino acid residue. The few remaining aromatic resonances were assigned after the first iteration of structure calculations.

NOE assignments and structure calculations

NOEASSIGN was used to generate a list of possible NOE assignments from the 3D and 4D NOESY data using the completed chemical shift data base. An error bar of ± 1 experimental NMR data point was used for matching the NOE cross peaks to the chemical shift database. During the first iteration of the NOE assignments, the algorithm identified 557 distance restraints unambiguously from the 3D ^{15}N, 4D $^{15}N/^{13}C$, and 4D $^{13}C/^{13}C$ resolved NOESY data sets. These restraints were then checked against the predicted secondary structures by PRECHECK. No errors in the restraints were noted by this inspection. The procedure thus produced 557 automatically assigned NOE restraints. Of these, 217 were intraresidue, 127 were sequential, 56 were short-range ($|i - j| < 5$), and 157 were long-range NOE restraints. These 557 automatically derived NOE restraints, along with the intermolecular and intraligand restraints, were used in a distance geometry/simulated annealing protocol to generate the first-round structures.

A total of 30 structures of the protein–ligand complex were produced during the first round of calculations. Of these, 15 were deemed acceptable based upon their low total energies. Statistics on these and subsequent structures are given in Table 1, and a superposition of these conformations is given in Fig. 6B. As shown in Table 1, the energies of these first-round conformations are quite low (E_{tot} = 323 ± 36 kcal). The rmsd between the average conformation and the 15 struc-

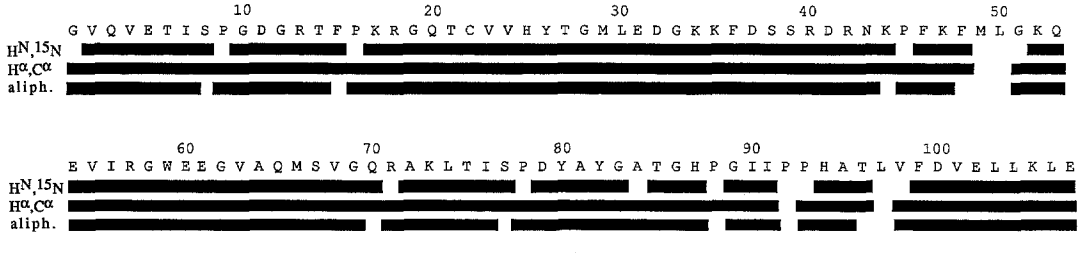

Fig. 5. Schematic depiction of the results of the automatic assignment strategy.

tures was 1.50 ± 0.17 Å and 2.02 ± 0.15 Å for the backbone and all heavy atoms, respectively. Importantly, the deviation from the previously determined high-resolution structure (Meadows et al., 1993) was 2.5 Å for the backbone atoms. This relatively low value indicates that the global fold of the conformations determined with the computer-generated NOE assignments was correct.

The energy-minimized average conformation of the 15 acceptable structures was obtained (see Table 1 for structure statistics) and used by the XYZFILTER program to identify additional NOE assignments and distance restraints. New NOE restraints were accepted if only one of the

TABLE 1
STRUCTURE STATISTICS FOR THE FKBP/ASCOMYCIN COMPLEXES DETERMINED USING THE AUTOMATIC CHEMICAL SHIFT AND NOE ASSIGNMENT METHODOLOGY[a]

	Iteration (1)		Iteration (2)		Final	
Cartesian coordinate RMSD (Å)	Backbone	Heavy atoms	Backbone	Heavy atoms	Backbone	Heavy atoms
<SA> vs <$\overline{SA}$>	1.50 ± 0.17	2.02 ± 0.15	1.20 ± 0.15	1.72 ± 0.20	1.17 ± 0.15	1.69 ± 0.20
<SA> vs NMR	2.45 ± 0.20	3.28 ± 0.20	1.79 ± 0.16	2.63 ± 0.25	1.76 ± 0.16	2.58 ± 0.25
<$\overline{SA}$>r vs NMR	2.15	2.80	1.44	2.20	1.47	2.19
Selected X-PLOR energies (kcal · mol^{-1})	<SA>	<$\overline{SA}$>r	<SA>	<$\overline{SA}$>r	<SA>	<$\overline{SA}$>r
E_{noe}	31.5 ± 9.8	12.3	37.5 ± 14.8	35.8	35.8 ± 14.8	40.6
E_{vdw}	39.8 ± 8.0	43.6	26.9 ± 7.8	33.6	26.2 ± 8.2	42.6
E_{tot}	322.6 ± 35.9	250.1	265.6 ± 41.0	258.3	262.1 ± 40.7	273.8
RMSD from experimental distance restraints (Å)	<SA>	<$\overline{SA}$>r	<SA>	<$\overline{SA}$>r	<SA>	<$\overline{SA}$>r
all	0.030 ± 0.005	0.019	0.025 ± 0.005	0.025	0.024 ± 0.005	0.026
protein	0.030 ± 0.006	0.016	0.025 ± 0.006	0.024	0.023 ± 0.006	0.026
intermolecular	0.019 ± 0.008	0.010	0.010 ± 0.005	0.020	0.011 ± 0.006	0.017
RMSD from idealized X-PLOR geometries	<SA>	<$\overline{SA}$>r	<SA>	<$\overline{SA}$>r	<SA>	<$\overline{SA}$>r
bonds (°)	0.003 ± 0.000	0.003	0.003 ± 0.000	0.003	0.003 ± 0.000	0.003
angles (°)	0.538 ± 0.028	0.509	0.481 ± 0.031	0.494	0.479 ± 0.034	0.498
impropers (°)	0.714 ± 0.170	0.629	0.656 ± 0.058	0.632	0.658 ± 0.054	0.611

[a] <SA> is the ensemble of structures, <$\overline{SA}$> is the mean atomic structure obtained by averaging the superimposed coordinates of the accepted structures, and <$\overline{SA}$>$_r$ is the energy-minimized mean atomic structure. None of the final structures had distance violations > 0.4 Å. Energies were calculated by X-PLOR 3.1 (Brünger, 1992), using a square-well potential for the NOE term (50 kcal · mol^{-1} · Å^{-2}) and a quartic potential (Nilges et al., 1988) for E_{vdw} with the van der Waals radii set to 0.8 times the CHARMM values (Brooks et al., 1983). In addition to the automatically obtained NOE and hydrogen bond restraints for the protein, 36 intermolecular and 88 intraligand NOE restraints (Petros et al., 1991) were used.

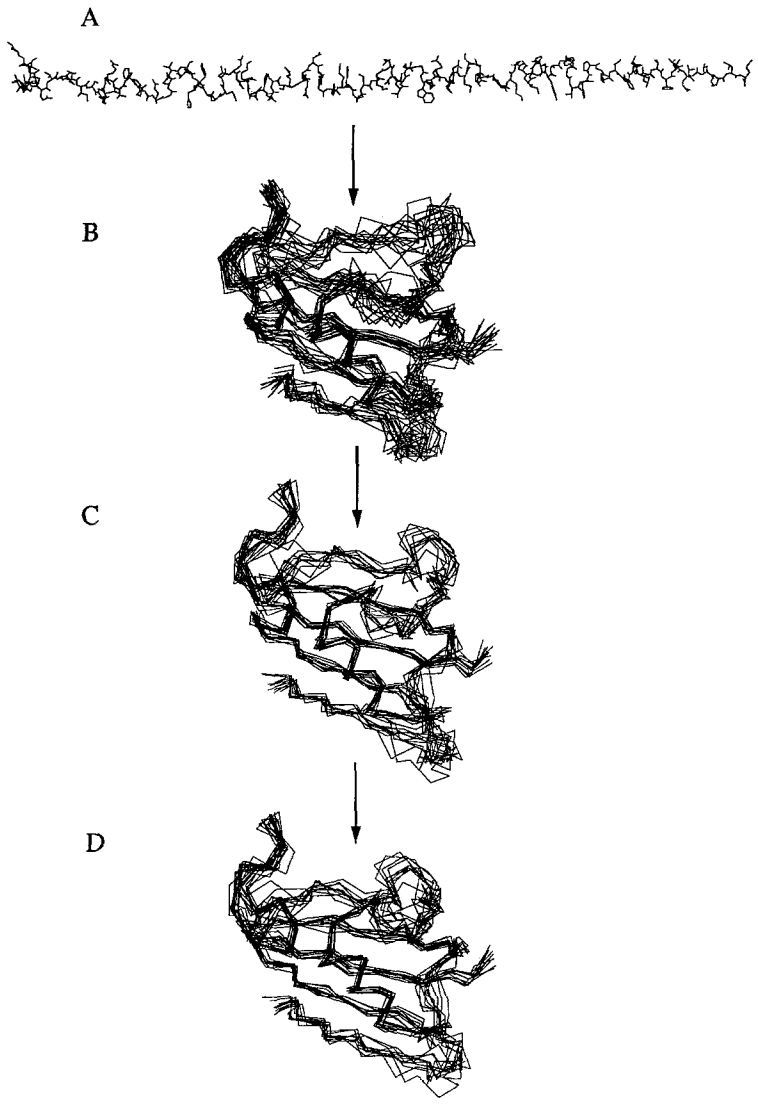

Fig. 6. Cartesian coordinate time line of the FKBP/ascomycin complex (ascomycin not shown) starting from a linear peptide (A). (B) DG/SA conformations generated with 557 unambiguous, automatically assigned restraints. The minimized average conformation of (B) was used to identify new NOE restraints from the list of possible NOE assignments. This resulted in an additional 447 restraints. Computerized Cartesian coordinate analysis also identified four unambiguous hydrogen bonds. A second iteration of DG/SA including 1023 automatically assigned protein NOEs and four hydrogen bonds produced structures (C). Repeating the distance filtering technique produced an additional 60 restraints, and Cartesian coordinate analysis resulted in an additional 15 hydrogen bonds. These were used in a final energy minimization of the (C) conformations to produce the final structures as shown in D.

possible assignment combinations had an interproton distance of less than 8.0 Å in the average minimized geometry. With the initial structure as a guide, 457 new NOE restraints were produced. Of these, 205 were intraresidue, 114 were sequential, 40 were short-range, and 98 were long-range NOE restraints. Since these new restraints were validated by the conformation they

were not checked further. In addition to the NOEs, hydrogen bonding acceptors for four slowly exchanging amides were obtained by FINDHBOND. Thus, the 1014 automatically assigned NOE and eight hydrogen bond restraints, along with the intermolecular and intraligand restraints, were used in the second iteration of structure calculations.

As in the first iteration, 30 structures were calculated from the DG/SA protocol. Of these, 10 were accepted for further analysis based upon their energies, using the same criterion as the previous iteration. As seen in Table 1, the structures produced by the second iteration were of good quality energetically. At this stage, the rmsd of the mean atomic coordinates was 1.2 and 1.7 Å for the backbone and heavy atoms, respectively (Fig. 6C). Importantly, although the rmsd was smaller than in the first iteration, both with respect to the mean and target conformations, the energies of the conformations did not increase. From these 10 conformations, an average set of Cartesian coordinates was obtained and energy minimized. As before, this conformation served to delineate new NOE restraints from the list of possible NOE assignments. This time a distance cutoff of 6 Å was used to validate the assignments, resulting in only 60 new restraints. Also 17 new hydrogen bond acceptors were located by the FINDHBOND program. These data were included into the final step of the protocol which consisted of energy minimization of the 10 conformations from iteration two. The final restraint list contained 1074 automatically assigned NOE and 38 hydrogen bond restraints, 88 intraligand and 36 intermolecular NOE restraints.

The statistics of the final energy-minimized conformations are shown in Table 1. The 3D structures show good convergence, with a final rmsd from the average conformation of 1.17 ± 0.15 Å and 1.69 ± 0.20 Å for the backbone and heavy atoms, respectively (Fig. 6D). In addition, the energies are low, and there are no NOE violations > 0.4 Å. Also, there is no systematic deviation when compared to the manually derived conformation of the FKBP/ascomycin complex (Figs. 7A, B). Although the structures generated by this semiautomated procedure are not as highly resolved as those that were manually derived, most of the variability is attributed to the lack of both stereospecific assignments and dihedral angle restraints, and also to the cautious categorization of the NOE restraints (3.5 and 5.0 Å).

CAVEATS

In order for this protocol to become routine, several experimental factors must be considered. First, the automated assignment and structure analysis depends on the availability of a near complete set of consistent chemical shifts for the protein. This necessitates special care in the accumulation of data. It is important to diminish errors in peak matching due to chemical shifts observed for different samples. Improvements in peak matching can be obtained by using a single sample in H_2O with NMR experiments conducted using gradients for water suppression (Bax and Pochapsky, 1992; Davis et al., 1992; John et al., 1992; Kay et al., 1992). Shift variation due to Bloch Seigert effects can also be compensated for by the careful design of experiments. In addition, nonlinear errors from temperature differences between experiments may be reduced by precisely controlling the temperature of the sample during data acquisition. With careful temperature control and by using the same sample for all experiments, a much narrower range of tolerances can be used in matching the chemical shifts (unpublished observations).

As is well known, greater spectral resolution significantly improves the automation process. Reducing phase cycling in experiments to minimize the number of scans per increment can greatly

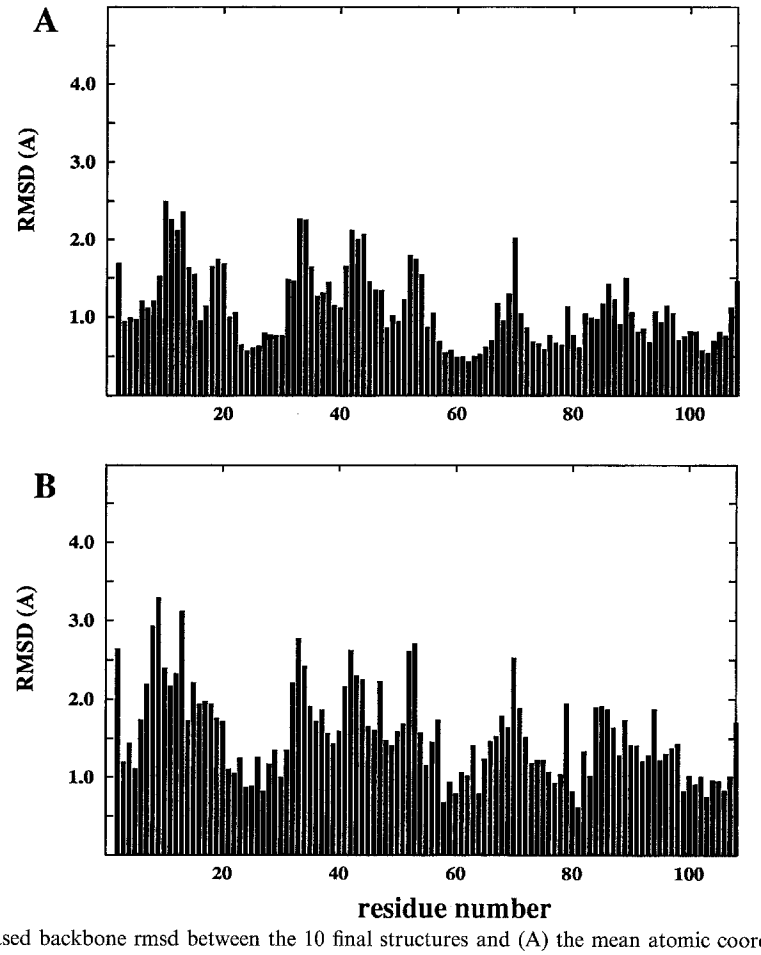

Fig. 7. Residue-based backbone rmsd between the 10 final structures and (A) the mean atomic coordinates for the 10 conformations, and (B), the previously determined high-resolution conformation of the FKBP/ascomycin complex in solution (Meadows et al., 1993).

increase the resolution with almost no loss in total sensitivity. In the triple resonance experiments for example, additional points can be accumulated in the constant time ^{15}N evolution period with little penalty. Reduced phase cycling versions of many of the necessary experiments can be constructed by using gradient pulses rather than phase cycling to suppress artifacts. In many cases, artifacts can be suppressed much more efficiently by gradient pulses than by phase cycling, thus leading to artifact-reduced spectra with fewer scans per time increment.

The automated sequential resonance assignment depends heavily on side-chain resonance information. For FKBP this data can be obtained from the HC(CO)NH-TOCSY experiment (Logan et al., 1992). For larger proteins this data can be obtained from a combination of experiments such as the CBCA(CO)NH (Grzesiek and Bax, 1992) and 4D HCCH-TOCSY (Olejniczak et al., 1992).

Finally, the NOE assignments need a near-complete chemical shift database. This is critical because in the initial stages of the refinement process unambiguously unique NOEs must be identified. It is our experience that completion of the database must be done manually. However,

the manual assignment process is often limited to a small number of protein-dependent special cases, and sequential assignments of aromatic atoms.

CONCLUSIONS

We have described a computer-based protocol that enables the rapid assignment of backbone and side chain resonances from heteronuclear multidimensional NMR experiments. From these data, the NOEs are automatically assigned and used to determine the 3D structure of the protein. The protocol is generalizable and may be used on a variety of proteins. Although not completely automated, the procedure as described above may substantially reduce the time required to generate the assignments and structures of proteins from NMR data.

REFERENCES

Arai, T., Koyama, Y., Suenaga, T. and Honda, H. (1962) *Antibiotics*, **15**, 231–232.

Bax, A. (1989) *Annu. Rev. Biochem.*, **58**, 223–256.

Bax, A. and Pochapsky, S.S. (1992) *J. Magn. Reson.*, **99**, 638–643.

Bernstein, R., Cieslar, C., Ross, A., Oschkinat, H., Freund, J. and Holak, T.A. (1993) *J. Biomol. NMR*, **3**, 245–251.

Billeter, M., Basus, V.J. and Kuntz, I.D. (1988) *J. Magn. Reson.*, **76**, 400–415.

Bodenhausen, G. and Ruben, D.J. (1980) *J. Chem. Phys. Lett.*, **69**, 185–189.

Brünger, A.T. (1992) *XPLOR manual*, Yale University Press, New Haven, CT.

Brooks, B.R., Brucoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M. (1983) *J. Comput. Chem.*, **4**, 187–217.

Catasti, P., Carrara, E. and Nicolini, C. (1989) *J. Comput. Chem.*, **11**, 805–818.

Clore, G.M. and Gronenborn, A.M. (1990) *CRC Crit. Rev. Biochem. Mol. Biol.*, **24**, 479–564.

Clore, G.M., Kay, L.E., Bax, A. and Gronenborn, A.M. (1991) *Biochemistry*, **30**, 12–18.

Cieslar, C., Clore, G.M. and Gronenborn, A.M. (1988) *J. Magn. Reson.*, **80**, 119–127.

Davis, A.L., Keeler, J., Laue, E.D. and Moskau, D. (1992) *J. Magn. Reson.*, **98**, 207–216.

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.

Edalji, R., Pilot-Matias, T.J., Pratt, S.D., Egan, D.A., Severin, J.M., Gubbins, E.G., Petros, A.M., Fesik, S., Burres, N.S. and Holzman, T.F. (1992) *J. Protein Chem.*, **11**, 213–223.

Fesik, S.W. and Zuiderweg, E.R.P. (1988) *J. Magn. Reson.*, **78**, 588–593.

Garret, D.S., Powers, R., Gronenborn, A.M. and Clore, G.M. (1991) *J. Magn. Reson.*, **95**, 214–220.

Glaser, S. and Kalbitzer, H.R. (1987) *J. Magn. Reson.*, **74**, 450–463.

Grahn, H., Delaglio, F., Delsuc, M.A. and Levy, G.C. (1988) *J. Magn. Reson.*, **77**, 294–307.

Grzesiek, S. and Bax, A. (1992) *J. Am. Chem. Soc.*, **114**, 6291–6293.

Grob, K.H. and Kalbitzer, H.R. (1988) *J. Magn. Reson.*, **76**, 87–99.

Hatanka, H., Kino, T., Miyata, S., Inamura, N., Kuroka, A., Goto, T., Tanaka, H. and Okuhara, M. (1988) *Antibiotics*, **41**, 1592–1601.

Harding, M.W., Galat, A., Uehling, D.E. and Schreiber, S.L. (1989) *Nature*, **341**, 758–760.

Havel, T.F. and Wüthrich, K. (1985) *J. Mol. Biol.*, **182**, 281–294.

Hoch, J.C., Hengyi, S., Kjær, M., Ludvigsen, S. and Poulsen, F.M. (1987), *Carlsberg Res. Commun.*, **52**, 111–122.

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

John, B.K., Plant, D., Webb, P. and Hurd, R. (1992) *J. Magn. Reson.*, **98**, 200–206.

Kay, L.E., Clore, G.M., Bax, A. and Gronenborn, A.M. (1990a) *Science*, **249**, 411–414.

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

Kay, L.E., Keifer, P. and Saarinen, T. (1992) *J. Am. Chem. Soc.*, **114**, 10663–10665.

Kleywegt, G.J., Lamerichs, R.M.J.N., Boelens, R. and Kaptein, R. (1989) *J. Magn. Reson.*, **85**, 186–197.

Kraulis, P.J. (1989) *J. Magn. Reson.*, **84**, 627–633.
Logan, T.L., Olejniczak, E.T., Xu, R.X. and Fesik, S.W. (1992) *FEBS Lett.*, **314**, 413–418.
Marion, D., Driscoll, P.C., Kay, L.E., Wingfield, P.T., Bax, A., Gronenborn, A.M. and Clore, G.M. (1989) *Biochemistry*, **28**, 6150–6156.
Meadows, R.P., Nettesheim, D.G., Xu, R.X., Olejniczak, E.T., Petros, A.M., Holzman, T.F., Severin, J., Gubbins, E., Smith, H. and Fesik, S.W. (1993) *Biochemistry*, **32**, 754–765.
Neidig, K.P., Bodenmueller, H. and Kalbitzer, H.R. (1984) *Biochem. Biophys. Res. Commun.*, **125**, 1143–1150.
Neri, D., Szyperski, T., Otting, G., Senn, H. and Wüthrich, K. (1989) *Biochemistry*, **28**, 7510–7516.
Nilges, M., Clore, G.M. and Gronenborn, A.M. (1988) *FEBS Lett.*, **229**, 317–324.
Novic, M., Oschkinat, H., Pfändler, P. and Bodenhausen, G. (1987) *J. Magn. Reson.*, **73**, 493–511.
Olejniczak, E.T., Xu, R.X. and Fesik, S.W. (1992) *J. Biomol. NMR*, **2**, 655–659.
Petros, A.M., Gampe Jr., R.T., Gemmeker, G., Neri, P., Holzman, T.F., Edalji, R., Hochlowski, J., Jackson, M., McAlpine, J., Luly, J.R., Pilot-Matias, S. and Fesik, S.W. (1991) *J. Med. Chem.*, **34**, 2925–2928.
Pfändler, P., Bodenhausen, G., Meier, B.U. and Ernst, R.R. (1985) *Anal. Chem.*, **57**, 2510–2516.
Pilot-Matias, T.J., Pratt, S.D. and Lane, B.C. (1993) *Gene*, **128**, 219–225.
Scheek, R.M., Van Gunsteren, W.F. and Kaptein, R. (1989) *Methods Enzymol.*, **177**, 204–218.
Senn, H., Werner, B., Messerle, B., Weber, C., Traber, R. and Wüthrich, K. (1989) *FEBS Lett.*, **249**, 113–118.
Siekierka, J.J., Hung, S.H.Y., Poe, M., Lin, C.S. and Sigal, N.H. (1989) *Nature*, **341**, 755–757.
Stoven, V., Mikou, A., Piveteau, D., Guittet, E. and Lallemand, J.Y. (1989) *J. Magn. Reson.*, **82**, 163–168.
Van de Ven, F.J.M. (1990) *J. Magn. Reson.*, **86**, 633–644.
Venters, R.A., Calderone, T.L., Spicer, L.D. and Fierke, C.A. (1991) *Biochemistry*, **30**, 4491–4494.
Weber, P.L., Malikayil, J.A. and Müeller, L. (1989) *J. Magn. Reson.*, **82**, 419–426.
Wider, G., Lee, K.H. and Wüthrich, K. (1982) *J. Mol. Biol.*, **155**, 367–388.
Wishart, D.S., Sykes, B.D. and Richards, F.M. (1992) *Biochemistry*, **31**, 1647–1651.
Wüthrich, K. (1983) *Biopolymers*, **22**, 131–138.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York, NY.
Wüthrich, K., Billeter, M. and Braun, W. (1983) *J. Mol. Biol.*, **169**, 949–961.
Wüthrich, K., Wider, G., Wagner, G. and Braun, W. (1982) *J. Mol. Biol.*, **155**, 311–319.
Zuiderweg, E.R.P., Petros, A.M., Fesik, S.W. and Olejniczak, E.T. (1991) *J. Am. Chem. Soc.*, **113**, 370–372.