Journal of Biomolecular NMR, 1 (1991) 421–438
ESCOM

J-Bio NMR 031

Statistical analysis of double NOE transfer pathways in proteins as measured in 3D NOE-NOE spectroscopy

Geerten W. Vuister[a,*], Rolf Boelens[a], André Padilla[b] and Robert Kaptein[a,**]

[a]*Bijvoet Center for Biomolecular Research, NMR Spectroscopy, University of Utrecht, Padualaan 8, 3584 CH Utrecht, The Netherlands*

[b]*Centre CNRS-INSERM de Pharmacologie-Endocrinologie, Rue de la Cardonille, F-34094 Montpellier Cedex, France*

Dedicated to the memory of Professor V.F. Bystrov

Received 17 May 1991
Accepted 2 August 1991

Keywords: Statistical analysis; 3D NMR spectroscopy; NOE; Assignment

SUMMARY

The recent development of three-dimensional NMR spectroscopy has alleviated the problem of overlap of resonances. However, also for the 3D experiments resonance assignment strategies have usually relied upon knowledge about spin systems, combined with information about short (sequential) distances. For doubly ($^{15}N/^{13}C$)-labelled molecules, a novel assignment strategy has been developed. In this paper we address the possibilities of an assignment strategy for proteins, based solely upon the use of NOE data. For this, the 3D NOE-NOE experiment seems most suitable. Therefore, we have made a theoretical evaluation of double NOE transfer pathways in 28 protein crystal structures. We identify 95 connectivities which are most likely to be observed as cross peaks in a 3D NOE-NOE spectrum of a protein. Given the occurrence of one of these 95 connectivities, we evaluate the chances of occurrence for the others. Analysis of these conditional probabilities allowed the construction of five patterns of related, highly correlated cross peaks which resemble the conventional idea of spin systems to some extent and may provide a basis for assignment and secondary structure analysis from 3D NOE-NOE data alone.

INTRODUCTION

The recent development of three-dimensional (3D) (Griesinger et al., 1987a,b; Vuister and Boelens, 1987; Fesik and Zuiderweg, 1988; Oschkinat et al., 1988; Vuister et al., 1988; Marion et al., 1989a) and four-dimensional (4D) NMR techniques (Kay et al., 1990a; Clore et al., 1991;

*Present address: Laboratory of Chemical Physics, NIDDK, National Institutes of Health, Bethesda, MD 20892, U.S.A.
** To whom correspondence should be addressed.

Zuiderweg et al., 1991) has increased the range of molecules amenable to high-resolution NMR spectroscopy. The increased resolution, obtained by the introduction of extra domains, resolves the overlap of cross peaks, thus allowing assignment and structural analysis.

In the first stage of protein structure determination by NMR, resonance assignment is the principle aim. Traditional methods of assignment and structural analysis usually have relied on the well-known approach of combining information about spin systems, obtained from COSY (Aue et al., 1976) or 2D HOHAHA (Braunschweiler and Ernst, 1983; Bax and Davis, 1985), with distance information obtained from NOE experiments (Jeener et al., 1979). COSY or 2D HOHAHA spectra yield patterns of cross peaks (i.e. the spin systems), which are linked by the evaluation of $d_{\alpha N}$, $d_{\beta N}$ and d_{NN} NOEs in 2D NOE spectra (Wüthrich, 1986).

In the second stage NOE-based proton-distance information is used in distance-geometry calculations (Havel and Wüthrich, 1984; Braun and Gō, 1985) and restrained molecular-dynamics calculations (Clore et al., 1985; Kaptein et al., 1985) to obtain conformations of the protein consistent with the NMR data. This approach has successfully been applied to the resonance assignment and structure determination of approximately 70 proteins in solution.

The 3D NOE-HOHAHA (Oschkinat et al., 1988; Vuister et al., 1988) and 3D HOHAHA-NOE (Oschkinat et al., 1989; Padilla et al., 1990) experiments combine both scalar interaction and dipolar interaction in one experiment. It is for this reason that these were the first versatile homonuclear 3D experiments for proteins. From these experiments all necessary information for assignment can be obtained (Padilla et al., 1990). Furthermore, cross-peak volumes and patterns can be related to secondary structure elements (Oschkinat et al., 1990; Vuister et al., 1990). In the first heteronuclear 3D experiments, the chemical shift dispersion of the heteronucleus (i.e. ^{15}N) was used to edit homonuclear 2D NOE or 2D HOHAHA spectra into less complex subspectra (Fesik and Zuiderweg, 1988; Marion et al., 1989a,b; Zuiderweg and Fesik, 1989).

Although the problem of overlap is alleviated by these 3D experiments, sequential assignment was still based on the combination of spin-system information and short-distance sequential NOE connectivities. Since the intrinsic linewidth of the resonances becomes larger with increasing size of the molecules, experiments (2D or 3D) based upon homonuclear scalar interactions (COSY, HOHAHA) become almost impossible for larger proteins. Clearly, important information needed for the traditional sequential assignment approach is then missing. In case of heteronuclear techniques, this problem has been overcome by a novel assignment strategy, based upon the use of doubly ($^{15}N/^{13}C$)-labelled molecules in $^{13}C/^{1}H$ COSY/TOCSY-type experiments (Bax et al., 1990; Fesik et al., 1990; Kay et al., 1990b) to obtain spin-system information and $^{15}N/^{13}C/^{1}H$ triple-resonance experiments (Bax et al., 1988; Wester et al., 1988; Ikura et al., 1990; Montelione and Wagner, 1990) to obtain sequential assignments.

However, since double isotopic labelling is not always possible, we would like to explore the possibilities of a different approach. As outlined above, the assignment of proton-proton NOEs is the key step in structural analysis, since this is the source of distance constraints used in distance-geometry and restrained molecular-dynamics calculations. We will evaluate the possibilities for obtaining these assignments from 3D spectra in the absence of a *J* coupling analysis. The 3D NOE-NOE experiment (Boelens et al., 1989; Breg et al., 1990) seems most suitable for this purpose, for a number of reasons: first, the analysis of the 3D NOE-NOE spectrum of a protein shows that virtually all information present in 'traditional' *J* and NOE-based spectra, can also be found in the 3D NOE-NOE spectrum (Padilla et al., manuscript in preparation). Secondly, the

double NOE transfer of the 3D NOE-NOE experiment allows novel connectivities, as compared to other (homonuclear) 3D experiments identifying, for example, tripeptide fragments (Breg et al., 1990). Thirdly, the redundancy in information allows multiple checking of resonance assignments (Padilla et al., manuscript in preparation). Finally, the NOE effect is expected to increase with increasing size of the molecules (Neuhaus and Williamson, 1989), making this experiment more suitable for the larger biomolecules.

Biomolecules are made up of regular building blocks and for this reason a number of characteristic short- and medium-range distances occur. For proteins they have been extensively evaluated by Wüthrich and co-workers (for a review see Wüthrich, 1986) and they provide the basis of the previously mentioned sequential assignment strategy. The existence of specific NOE patterns in relation to secondary structure elements like α-helical and β-sheet domains, has been employed in the main-chain directed assignment strategy (Englander and Wand, 1987). Furthermore, the presumption of helical structure allows assignments of DNA fragments to be made on the basis of NOE data alone (Scheek et al., 1984). It has also been realised before that, because of holonomic constraints (i.e. chemical bonding), correlations between distance constraints can be expected. In this respect, the concept of structural templates was introduced (Hempel, 1989). As a result it can be expected that the observation of NOE cross peaks will be correlated too.

Although intuitively the intra-residual and sequential cross peaks are expected to be predominant in the spectrum, only a statistical analysis can correctly identify to which extent this presumption is valid. For this, we have evaluated 28 crystal structures with a resolution better than 0.2 nm, from the Brookhaven Protein Data Bank (Bernstein et al., 1977). We have identified 95 connectivities which are most likely to be observed as cross peaks in a 3D NOE-NOE spectrum of a protein. Furthermore, we have constructed clusters of related highly correlated 3D cross peaks to identify patterns which resemble the conventional idea of spin systems to some extent and may provide a basis for assignment and secondary structure analysis from 3D NOE-NOE data alone.

METHODS

The aim of the present analysis is to classify connectivities resulting from the double NOE transfer, which may be observed as cross peaks in 3D NOE-NOE spectra. For this, we must have information about the spins involved in the connectivity and be able to calculate the expected cross-peak volume.

3D NOE-NOE cross-peak volumes

The 3D cross-peak volume is dependent on the total transfer efficiency $E_{3D}(abc)$ involving protons *a*, *b*, and *c*, which can be written as the product of the transfer efficiencies of the two NOE mixing periods τ_{m1} and τ_{m2}

$$E_{3D}(abc) = E_{NOE1}(ab)\ E_{NOE2}(bc) \tag{1}$$

For a simple two-spin system the transfer of magnetization during a NOE mixing period can be approximated in the slow-motion limit, assuming a rigid, isotropically tumbling molecule (Neuhaus and Williamson, 1989) for a given τ_m by

$$E_{NOE}(pq) \propto r_{pq}^{-6} \tag{2}$$

where r_{pq} represents the distance between spins p and q. The 3D cross-peak transfer efficiency can now be expressed as

$$E_{3D}(abc) \propto (r_{ab}\, r_{bc})^{-6} \tag{3}$$

if both mixing periods have the same length.

Statistical analysis

Relevant stochastic variables to characterise the 3D NOE-NOE cross peak could be 'intensity', 'type of proton', 'type of NOE' and 'secondary structure element'. Thus, we introduce the probability distribution function P(I, TYPE1, STRUCT1, TYPE2, STRUCT2, TYPE3, STRUCT3, N1, N2). The stochastic variable* I (Intensity) is in essence continuous. However, it is common practice to evaluate NOEs in discrete terms as 'strong', 'medium', 'weak', and unobservable'. The TYPE1, TYPE2, and TYPE3 variables denote the type of the first, second, and third spin involved in the magnetization transfer pathway, respectively. We classify these into the categories 'NH', '$C^{\alpha}H$', '$C^{\beta}H$', and 'others'. The variables STRUCT1, STRUCT2, and STRUCT3 denote the secondary structure element for the first, second, and third spin involved in the magnetization transfer pathway, respectively. We classify these into categories 'helix', 'sheet' or 'extended', 'turn', and 'undefined'. The N1 and N2 variables will be related to the residue numbers i, j, and k of the spins involved in the magnetization transfer pathway in the following way: N1 = i − j, and N2 = k − j. The values of N1 and N2 will classify the NOE as intra-residual (N1, N2 = 0), sequential (N1, N2 = ±1), medium range (N1, N2 = ±2, ±3) or long range (N1, N2 > 3 or N1, N2 < −3).

The probability distribution function P(I, TYPE1, STRUCT1, TYPE2, STRUCT2, TYPE3, STRUCT3, N1, N2) then has

$$4 \times (4 \times 4)^3 \times 9^2 = 1\,327\,104$$

possible values (intensity 4, type 4 and secondary structure element 4, for each of the three protons, N1 and N2 each 9).

A marginal distribution function P'_{abc}(N1,N2) can be constructed from P(I, TYPE1, STRUCT1, TYPE2, STRUCT2, TYPE3, STRUCT3, N1, N2) as follows

$$\begin{aligned} P'_{abc}(N1,N2) = \; & P(I = \{\text{medium, strong}\}, \\ & TYPE1 = a, STRUCT1 \in \{\text{helix,sheet,turn,undefined}\}, \\ & TYPE2 = b, STRUCT2 = \{\text{helix,sheet,turn,undefined}\}, \\ & TYPE3 = c, STRUCT3 = \{\text{helix,sheet,turn,undefined}\}, \\ & N1, N2) \end{aligned} \tag{4}$$

* In accordance with mathematical traditions, lower case identifiers (i.e. n1, n2, etc.) will denote an actual value of a stochastic variable, whereas uppercase identifiers (i.e. N1, N2, etc.) will be used in case of the variable itself (Billingsley, 1986).

Since the subscripts a, b, and c and variables N1 and N2 define a connectivity*, the chance of the event $P'_{abc}(N1 = n1, N2 = n2)$ can also be expressed as $P''(C_{abc}(i+n1,i,i+n2))$, which is the chance of occurrence of the cross peak $C_{abc}(i+n1,i,i+n2)$.

Conditional probability

For the conditional probability, the chance of occurrence $C_{abc}(i+n1,i,i+n2)$ under the condi-

TABLE I
CRYSTAL PROTEIN STRUCTURES FROM THE BROOKHAVEN PROTEIN DATA BANK (Bernstein et al., 1977) USED IN THE STATISTICAL ANALYSIS

Protein	PDB entry	Resolution[a] (nm)	Protons	Evaluated 3D cross peaks
Phospholipase A2	1BP2	0.17	882	166626
Crambin	1CRN	0.15	314	50759
L7/L12 ribosomal protein	1CTF	0.17	520	128110
Erythrocruorin	1ECD	0.14	1017	224646
Glutathione peroxidase	1GP1	0.20	2888	755670
Insulin	1INS	0.15	754	155677
Lysozyme	1LZ1	0.15	977	224148
Plastocyanin	1PCY	0.16	711	170617
Beta-trypsin	1TPP	0.14	1577	390549
Ubiquitin	1UBQ	0.18	627	163252
Actinidin	2ACT	0.17	1550	396873
Acid proteinase	2APP	0.18	2198	548869
Azurin	2AZA	0.18	1894	450174
Carbonic anhydrase form B	2CAB	0.20	1923	477881
Cytochrome P450CAM	2CPP	0.16	3149	816171
Citrate synthase	2CTS	0.20	3410	861096
Hemoglobin	2HHB	0.17	4332	1010354
Lysozyme	2LZM	0.17	1321	318749
Ovomucoid third domain	2OVO	0.15	392	68448
Bence-Jones protein	2RHE	0.16	803	167904
Proteinase A	2SGA	0.15	1189	304273
Dihydrofolate reductase	3DFR	0.17	1252	280995
Native elastase	3EST	0.16	1541	329278
Thermolysin	3TLN	0.16	2269	559921
Cytochrome *c*	451C	0.16	604	120258
Flavodoxin	4FXN	0.18	1060	285310
Carboxypeptidase A	5CPA	0.15	2336	617726
Papain	9PAP	0.16	1575	390739

[a] Upper limit.

* The connectivity notation introduced by Vuister et al. (1990) can easily be used for identifying 3D NOE-NOE connectivities in an unambiguous way. The 3D NOE-NOE cross peak between spin a of residue i, spin b of residue j, and spin c of residue k can be identified as: $C[NOE,NOE]_{abc}(i,j,k)$ or shorthand: $C_{abc}(i,j,k)$. Throughout this paper the shorthand notation will be used to identify connectivities.

tion that the cross peak $C_{xyz}(p,i,q)$ is present, will be expressed as $P(C_{abc}(i+n1,i,i+n2)|C_{xyz}(p,i,q))$.

Twenty-eight protein structures with a resolution better than 0.2 nm were selected from the Brookhaven Protein Data Bank (Bernstein et al., 1977) (cf. Table 1). Definitions for the secondary structure elements were taken from the headers of these PDB files. Hydrogen atoms were added with INSIGHT V2.5 (Biosym Technologies). The resulting structures were directly used without energy minimisation by the program SAND (Statistical Analysis of NOE-NOE Data). SAND first generates the pair list for all proton pairs corresponding to a distance smaller than 0.5 nm. It then creates from this pair list the possible 3D NOE-NOE connectivities and records the occurrences in the relevant categories.

The intensity categories are defined by the following equation:

$$\begin{aligned}\text{unobservable} &< (0.4\text{ nm}, 0.4\text{ nm}) \leq \text{weak} < (0.325\text{ nm}, 0.325\text{ nm}) \leq \text{medium} \\ &< (0.25\text{ nm}, 0.25\text{ nm}) < \text{strong}\end{aligned} \tag{5}$$

where the distances between the brackets define a 3D cross peak with an intensity corresponding to these distances. These values were chosen on the basis of observable 3D cross peaks in the 3D NOE-NOE spectrum of pike parvalbumin in H_2O (Padilla et al., manuscript in preparation).

CSAND (Conditional Statistical Analysis of NOE-NOE Data) is a program that analyses the simultaneous occurrence of a number of predefined 3D NOE-NOE connectivities. The 3D connectivities are generated in the same way as in the program SAND. The 95 most likely types of cross peaks, as obtained by the program SAND, were used to define the connectivities for CSAND.

RESULTS AND DISCUSSION

Statistically relevant connectivities

The number of connectivities evaluated by SAND are listed in Table 1. The total amounts to 10435074. To evaluate these data, we decided to consider some marginal distribution functions, comprising the more reliable cross peaks in the categories medium and strong. Cross peaks in the weak category are expected to be found in the real 3D NOE-NOE spectrum only under favourable conditions. With each type of proton a distinct chemical-shift range can be associated (Gross and Kalbitzer, 1988). Therefore, various connectivities will be located in different regions of the 3D spectrum and can be discriminated. Ninety-five connectivities were selected on the basis of a high probability of occurrence or a high selectivity for either 'all-helix' or 'all-sheet', i.e. those instances in which *all* three resonances involved in the magnetization transfer pathway are located in either helical domains or β-sheet conformation. The results for the αNN, NNN, αβN, and βαN connectivities are listed in Tables 2A-D, respectively*.

As an illustration, a visual representation of the marginal distribution function $P'_{\alpha NN}(N1,N2)$ (cf. Table 2A) is shown in Fig. 1. Naturally, connectivities resulting from the simultaneous occurrence of the $d_{\alpha N}(i,i)$, $d_{\alpha N}(i,i+1)$ and $d_{NN}(i,i\pm1)$ will be predominant in this distribution. Relative-

* A listing of all 95 connectivities is available from the authors upon request.

TABLE 2A
MOST SIGNIFICANT ENTRIES IN THE MARGINAL DISTRIBUTION FUNCTION $P'_{\alpha NN}(N1,N2)$

Connectivity	Distribution (%)[a]	Helical[b] (%)	β-Sheet[c] (%)
$C_{\alpha NN}(<i-3,i,i-1)$	3	46	7
$C_{\alpha NN}(<i-3,i,i+1)$	2	44	7
$C_{\alpha NN}(i-3,i,i-1)$	6	72	1
$C_{\alpha NN}(i-3,i,i+1)$	5	71	1
$C_{\alpha NN}(i-1,i,<i-3)$	4	0	53
$C_{\alpha NN}(i-1,i,i-1)$	16	34	17
$C_{\alpha NN}(i-1,i,i+1)$	15	37	13
$C_{\alpha NN}(i-1,i,>i+3)$	4	0	53
$C_{\alpha NN}(i,i,i-1)$	12	49	4
$C_{\alpha NN}(i,i,i+1)$	12	50	4

[a] The total number of entries of $P'_{\alpha NN}(N1,N2)$ amounts to 30 064.
[b] All three resonances involved in the magnetization-transfer pathway are located in helical domains.
[c] All three resonances involved in the magnetization-transfer pathway are located in β-sheet domains.

ly selective are the $C_{\alpha NN}(i,i,i\pm1)$ connectivities, which combine intra-residual and sequential information in one connectivity, the $C_{\alpha NN}(i-1,i,i-1)$ connectivity identifying one sequential residue through two NOEs and the $C_{\alpha NN}(i-1,i,i+1)$ connectivity identifying a tripeptide fragment.

The $P'_{\alpha NN}(N1,N2)$ marginal distribution function is of interest because of the well-known results of the analysis of short distances in proteins by Billeter et al. (1982), which provides the basis for the sequential assignment strategy. From their results it was concluded that if both the $d_{\alpha N}$ and d_{NN} NOEs can be identified, then in 95% of the cases this would correspond to the sequential

TABLE 2B
MOST SIGNIFICANT ENTRIES IN THE MARGINAL DISTRIBUTION FUNCTION $P'_{NNN}(N1,N2)$

Connectivity	Distribution (%)[a]	Helical[b] (%)	β-Sheet[c] (%)
$C_{NNN}(<i-3,i,<i-3)$	3	1	66
$C_{NNN}(i-2,i,i-1)$	6	50	0
$C_{NNN}(i-2,i,i+1)$	5	54	1
$C_{NNN}(i-1,i,i-1)$	23	54	4
$C_{NNN}(i-1,i,i+1)$	20	55	2
$C_{NNN}(i-1,i,i+2)$	4	59	1
$C_{NNN}(i+1,i,i+1)$	23	54	3
$C_{NNN}(i+1,i,i+2)$	5	57	1
$C_{NNN}(>i+3,i,>i+3)$	3	0	68

[a] Two entries representing the same connectivity, i.e. $C_{NNN}(i,j,k)$ and $C_{NNN}(k,j,i)$, are counted only once. The total number of entries of $P'_{NNN}(N1,N2)$ amounts to 13 591.
[b] All three resonances involved in the magnetization-transfer pathway are located in helical domains.
[c] All three resonances involved in the magnetization-transfer pathway are located in β-sheet domains.

TABLE 2C
MOST SIGNIFICANT ENTRIES IN THE MARGINAL DISTRIBUTION FUNCTION $P'_{\alpha\beta N}(N1,N2)$

Connectivity	Distribution (%)[a]	Helical[b] (%)	β-Sheet[c] (%)
$C_{\alpha\beta N}(i-3,i,i)$	7	79	0
$C_{\alpha\beta N}(i-3,i,i+1)$	4	81	1
$C_{\alpha\beta N}(i,i,i)$	32	38	20
$C_{\alpha\beta N}(i,i,i+1)$	22	47	12
$C_{\alpha\beta N}(i+1,i,i)$	3	72	2
$C_{\alpha\beta N}(i+1,i,i+1)$	2	51	11

[a] The total number of entries of $P'_{\alpha\beta N}(N1,N2)$ amounts to 39 880.
[b] All three resonances involved in the magnetization-transfer pathway are located in helical domains.
[c] All three resonances involved in the magnetization-transfer pathway are located in β-sheet domains.

NOEs, linking residues i and i+1. Their analysis was based on the comparison of the simultaneous occurrence of the $d_{\alpha N}(i,j)$ and $d_{NN}(i,j)$ NOEs for $(i-j)=0,\pm 1$ with the simultaneous occurrence of the $d_{\alpha N}(i,j)$ and $d_{NN}(i,j)$ NOEs for $(i-j)\neq 0,\pm 1$ (cf. Fig. 2, both grey areas). However, this comparison is allowed only because the intra-residual $d_{\alpha N}(i,i)$ is assumed to be known from comparison with COSY or 2D HOHAHA spectra.

In our analysis this is not the case and thus the combinations $d_{\alpha N}(i,j)$ and $d_{NN}(k,l)$ with $(i-j)=0,-1$ $(k-l)\neq \pm 1$, and $d_{\alpha N}(i,j)$ and $d_{NN}(k,l)$ with $(i-j)\neq 0,-1$ $(k-l)=\pm 1$ are also considered in the analysis (cf. Fig. 2, both blank areas). Thus, summing the categories $C_{\alpha NN}(i,i,i-1)$, $C_{\alpha NN}(i,i,i+1)$, $C_{\alpha NN}(i-1,i,i+1)$ and $C_{\alpha NN}(i-1,i,i-1)$ yields a value of 55% (cf. Table 2A), and not the high selectivity of 95%, as one would expect at first glance. In Fig. 2 the percentages corresponding to the four different categories, as obtained from the marginal distribution

TABLE 2D
MOST SIGNIFICANT ENTRIES IN THE MARGINAL DISTRIBUTION FUNCTION $P'_{\beta\alpha N}(N1,N2)$

Connectivity	Distribution (%)[a]	Helical[b] (%)	β-Sheet[c] (%)
$C_{\beta\alpha N}(<i-3,i,i+1)$	4	5	31
$C_{\beta\alpha N}(i,i,<i-3)$	3	2	42
$C_{\beta\alpha N}(i,i,i)$	24	37	20
$C_{\beta\alpha N}(i,i,i+1)$	20	28	19
$C_{\beta\alpha N}(i,i,i+2)$	4	32	3
$C_{\beta\alpha N}(i,i,i+3)$	6	74	1
$C_{\beta\alpha N}(i,i,>i+3)$	6	34	23
$C_{\beta\alpha N}(i+1,i,i+1)$	4	2	42
$C_{\beta\alpha N}(i+3,i,i)$	4	81	1
$C_{\beta\alpha N}(i+3,i,i+1)$	2	82	1
$C_{\beta\alpha N}(i+3,i,i+3)$	2	84	1
$C_{\beta\alpha N}(>i+3,i,i+1)$	4	4	27

[a] The total number of entries of $P'_{\beta\alpha N}(N1,N2)$ amounts to 40 461.
[b] All three resonances involved in the magnetization-transfer pathway are located in helical domains.
[c] All three resonances involved in the magnetization-transfer pathway are located in β-sheet domains.

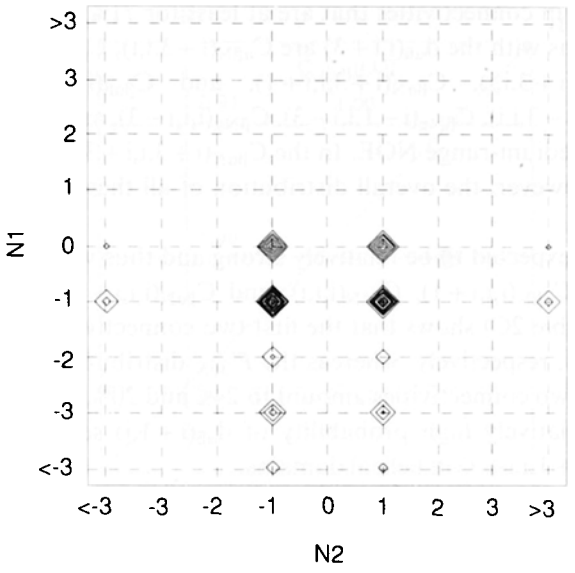

Fig. 1. Two-dimensional contour plot of the marginal distribution function $P'_{\alpha NN}(N1,N2)$ as obtained from the statistical analysis of 3D NOE-NOE connectivities in 28 crystal structures of the Brookhaven Protein Data Bank (Bernstein et al., 1977). Each contour level corresponds to 500 absolute occurrences. The total number of occurrences amounts to 30 064, so each contour level represents 1.7%.

$P'_{\alpha NN}(N1,N2)$ are given. From the fractions in the grey areas the 95% selectivity, similar to the results of Billeter et al. (1982) can be reproduced.

Several interesting connectivities were found. The $C_{NNN}(i-1,i,i+1)$ connectivities (cf. Table 2B) have a high occurrence. A priori, this constitutes information about tripeptide fragments, which are to a large extent located in helical domains.

The combination of intra-residual and sequential NOEs, with the medium-range $d_{\alpha\beta}(i,i+3)$ and

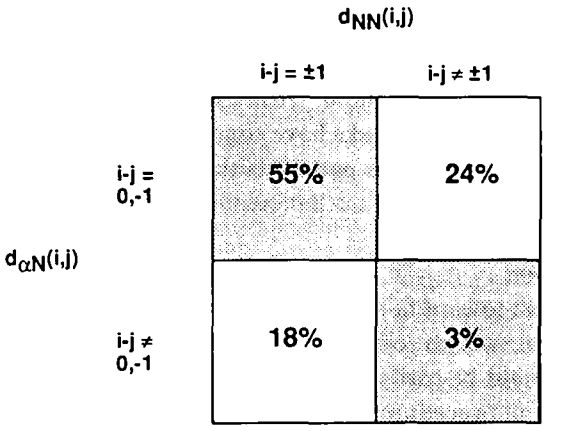

Fig. 2. Schematic representation of four possible categories of $C_{\alpha NN}$ connectivities, resulting from the combination of $d_{\alpha N}(i,j)$ and $d_{NN}(i,j)$ NOEs. The distribution of the four categories, as obtained by SAND, is indicated. Refer to the text for further details.

$d_{\alpha N}(i,i+3)$ NOEs results in connectivities that are at least for 71% exclusive for helical domains. Examples of combinations with the $d_{\alpha\beta}(i,i+3)$ are $C_{\alpha\beta N}(i-3,i,i)$, $C_{\alpha\beta N}(i-3,i,i+1)$, $C_{\beta\alpha N}(i+3,i,i)$, $C_{\beta\alpha N}(i+3,i,i+1)$, $C_{\beta\alpha N}(i+3,i,i)$, $C_{\beta\alpha N}(i+3,i,i+1)$, and $C_{\beta\alpha\beta}(i+3,i,i)$. The $C_{N\alpha N}(i+3,i,i)$, $C_{\alpha NN}(i-3,i,i\pm1)$, $C_{N\alpha N}(i-3,i,i)$, $C_{\beta N\alpha}(i-1,i,i-3)$, $C_{\beta N\alpha}(i,i,i-3)$, and $C_{\beta\alpha N}(i,i,i+3)$ connectivities involve the $d_{\alpha N}(i,i+3)$ medium-range NOE. In the $C_{\beta\alpha N}(i+3,i,i+3)$ both types of medium-range NOEs are combined. However, the overall distribution of all these cross peaks is relatively low (ca. 6%).

The $d_{\alpha\beta}(i,i)$ NOEs are expected to be relatively strong and thus will yield valuable information through the $C_{\alpha\beta N}(i,i,i)$, $C_{\alpha\beta N}(i,i,i+1)$, $C_{\beta\alpha N}(i,i,i)$, and $C_{\beta\alpha N}(i,i,i+1)$ connectivities. Indeed, the $P'_{\alpha\beta N}$ distribution (cf. Table 2C) shows that the first two connectivities have a probability of occurrence of 32% and 22%, respectively, whereas the $P'_{\beta\alpha N}$ distribution (cf. Table 2D) shows that the numbers for the last two connectivities amount to 24% and 20%, respectively. A somewhat unexpected result is the relatively high probability of $d_{\alpha\beta}(i+1,i)$ sequential NOEs, through the $C_{\alpha\beta N}(i+1,i,i)$ and $C_{\alpha\beta N}(i+1,i,i+1)$ in helical domains.

Search for correlated connectivities

The a priori chances of the 95 connectivities are in our view not high enough to reliably base an assignment strategy upon. In practice, no spectroscopist will analyse a spectrum without taking correlations into account. If patterns of highly correlated cross peaks can be constructed, it may be possible to increase specificity. Cross peaks differing only in either ω_1 or ω_3 are found on lines parallel to the ω_1 or ω_3 axis in the 3D frequency space, respectively. With the exception of cases of two-spin overlap, such related cross peaks share a common NOE.

The $C_{\beta\alpha N}$ connectivities may serve as an example. They correlate three important resonances in one cross peak. The $P'_{\beta\alpha N}$ distribution shows that the $C_{\beta\alpha N}(i,i,i)$ and $C_{\beta\alpha N}(i,i,i+1)$ connectivities are 24% and 20% selective, respectively. The $C_{\beta\alpha N}(i,i,i)$, $C_{N\alpha N}(i,i,i)$, $C_{\alpha\beta N}(i,i,i)$, and $C_{N\beta N}(i,i,i)$ together form a pattern in an ω_3 cross section at the NH(i) frequency as depicted in Fig. 3A. The CSAND analysis shows that the last three connectivities have a very high conditional probability with the $C_{\beta\alpha N}(i,i,i)$ connectivity of 1.00, 1.00 and 0.91, respectively. Therefore, if the $C_{\beta\alpha N}(i,i,i)$ cross peak is present, then the $C_{N\alpha N}(i,i,i)$ and $C_{\alpha\beta N}(i,i,i)$ cross peaks are expected too, whereas the $C_{N\beta N}(i,i,i)$ cross peak has a very high chance (0.91) of being observed.

The $C_{\beta\alpha N}(i-1,i-1,i)$, $C_{N\alpha N}(i,i-1,i)$, $C_{\alpha\beta N}(i-1,i-1,i)$, and $C_{N\beta N}(i,i-1,i)$ together form a pattern with an identical appearance in the ω_3 cross section at the NH(i) frequency, as depicted in Fig. 3B. However, the last three cross peaks have conditional probabilities with $C_{\beta\alpha N}(i-1,i-1,i)$ of 0.50, 0.79 and 0.49, respectively. This pattern is therefore expected to occur less often in the spectrum compared to the pattern constructed previously from the $C_{\beta\alpha N}(i,i,i)$ cross peak.

Pattern-generating algorithm

Patterns should be robust toward false information and resemble something of use to the spectroscopist. In order to form such patterns we must define the rules by which they are constructed. Preferably, these rules should be easily implemented in a computer program which assists in a (semi-)automatic assignment procedure. Guided by our experiences with pattern-generating algorithms for 3D HOHAHA-NOE spectra (Kleywegt, 1991; Kleywegt et al., 1991), we have used the following algorithm: First, a 3D cross peak is chosen as the seed of a pattern. New pattern members are added if they have at least one NOE in common with the seed. For example, if the seed

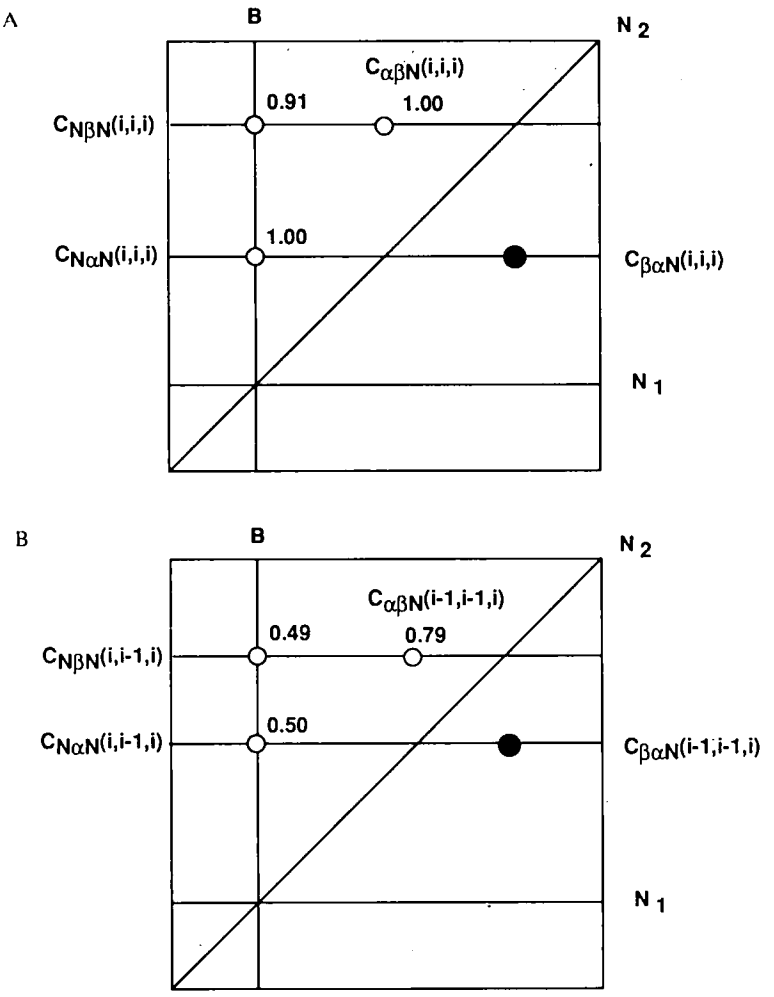

Fig. 3. (A) The cross peaks $C_{\beta\alpha N}(i,i,i)$, $C_{\alpha\beta N}(i,i,i)$, $C_{N\alpha N}(i,i,i)$ and $C_{N\beta N}(i,i,i)$ together form a pattern in the ω_3 cross section at the NH(i) resonance frequency. Given the $C_{\beta\alpha N}(i,i,i)$ cross peak (indicated as the black circle), the remaining three cross peaks are expected with chances 1.00, 1.00, and 0.91, respectively (cf. Table 3). (B) The cross peaks $C_{\beta\alpha N}(i-1,i-1,i)$, $C_{\alpha\beta N}(i-1,i-1,i)$, $C_{N\alpha N}(i,i-1,i)$ and $C_{N\beta N}(i,i-1,i)$ together form a pattern with identical appearance in the ω_3 cross section at the NH(i) resonance frequency. However, given the $C_{\beta\alpha N}(i-1,i-1,i)$ cross peak (indicated as the black circle), the chances for observing the remaining three cross peaks are only 0.79, 0.50, and 0.49, respectively (cf. Table 4). Therefore, this pattern is less likely to occur. Note that $C_{\beta\alpha N}(i-1,i-1,i) \sim C_{\beta\alpha N}(i,i,i+1)$ by a simple shift of the index i.

is $C_{abc}(i,j,k)$ then the connectivities $C_{abP}(i,j,p)$, $C_{Qab}(q,i,j)$, $C_{Rbc}(r,j,k)$, and $C_{bcS}(j,k,s)$, define the new spins of type P, Q, R, and S of residues p, q, r, and s. Sometimes it can be advantageous to consider two cross peaks as seeds of a pattern, thus enlarging the number of possible cross peaks which may be part of the pattern. In contrast, in the case of C_{NNN}-based patterns, we decided to be more stringent (to be discussed shortly).

αβN-based patterns

As discussed before, the αβN and βαN connectivities combine three important spins with each

other. The $C_{\beta\alpha N}(i,i,i)$ and $C_{\alpha\beta N}(i,i,i)$ cross peaks are highly correlated and both can be found in the ω_3 cross section at the NH(i) resonance frequency. For these reasons we considered the construction of a pattern generated from the seeds $C_{\beta\alpha N}(i,i,i)$ and $C_{\alpha\beta N}(i,i,i)$. Table 3 shows the conditional probabilities $P(A \mid C_{\beta\alpha N}(i,i,i))$ and $P(A \mid C_{\alpha\beta N}(i,i,i))$ as obtained by the program CSAND. It may be noted that the intra-residual connectivities are highly correlated, irrespective of the type of secondary structure element. Since cross peaks defining an identical connectivity are considered only once, i.e. $C_{abc}(i,j,k) \sim C_{cba}(k,j,i)$, only nine independent connectivities can be constructed from the NH(i), $C^{\alpha}H(i)$, and $C^{\beta}H(i)$ spins. The observation of each of these nine cross peaks is highly correlated with the observation of both $C_{\beta\alpha N}(i,i,i)$ and $C_{\alpha\beta N}(i,i,i)$. These nine connectivities are schematically shown in Fig. 4A. Not only intra-residual connectivities contribute to this pattern, also some sequential connectivities are highly correlated and will extend the number of resonances comprised by the pattern with the $C^{\alpha}H(i-1)$ and $NH(i+1)$ resonance frequencies. Schematically this is shown in Fig. 4B. In this figure, the darker shaded circles indicate the spins of the seeds of the pattern, whereas the lighter shaded circles indicate spins which would be added by the algorithm. From Table 3, it can also be concluded that in helical domains the pattern may be extended and also comprises the $C^{\beta}H(i-1)$ and $NH(i-1)$ resonance frequencies. They are denoted by blank circles in Fig. 4.

As explained above, from the $C_{\beta\alpha N}(i,i,i+1)$ and $C_{\alpha\beta N}(i,i,i+1)$ connectivities a pattern can be

TABLE 3
CONDITIONAL PROBABILITIES $P(A \mid C_{\beta\alpha N}(i,i,i))$ AND $P(A \mid C_{\alpha\beta N}(i,i,i))$ OBTAINED FROM THE ANALYSIS OF 28 CRYSTAL PROTEIN STRUCTURES BY THE PROGRAM CSAND

A		$P(A \mid C_{\beta\alpha N}(i,i,i))$			$P(A \mid C_{\alpha\beta N}(i,i,i))$		
		All	Helical	β-Sheet	All	Helical	β-Sheet
$C_{\beta\alpha N}$	(i,i,i)	1.00	1.00	1.00	1.00	1.00	1.00
$C_{\alpha\beta N}$	(i,i,i)	1.00	1.00	1.00	1.00	1.00	1.00
$C_{N\alpha N}$	(i,i,i)	1.00	1.00	1.00	1.00	1.00	1.00
$C_{N\beta N}$	(i,i,i)	0.91	0.97	0.86	0.92	0.98	0.87
$C_{\alpha N\alpha}$	(i,i,i)	1.00	1.00	1.00	1.00	1.00	1.00
$C_{\beta N\alpha}$	(i,i,i)	0.97	0.99	0.95	0.97	0.99	0.95
$C_{\alpha\beta\alpha}$	(i,i,i)	1.00	1.00	1.00	1.00	1.00	1.00
$C_{\beta N\beta}$	(i,i,i)	0.91	0.98	0.86	0.92	0.98	0.87
$C_{\beta\alpha\beta}$	(i,i,i)	1.00	1.00	1.00	1.00	1.00	1.00
$C_{\alpha N\alpha}$	(i−1,i,i)	0.90	0.84	0.82	0.90	0.84	0.81
$C_{\beta N\alpha}$	(i,i,i−1)	0.95	0.91	0.81	0.95	0.91	0.81
$C_{N\alpha N}$	(i+1,i,i)	0.91	0.86	0.81	0.92	0.86	0.81
$C_{\alpha\beta N}$	(i,i,i+1)	0.77	0.86	0.62	0.78	0.86	0.63
$C_{\beta\alpha N}$	(i,i,i+1)	0.92	0.84	0.82	0.92	0.84	0.82
$C_{\beta N\beta}$	(i−1,i,i)	0.70	0.82	0.54	0.71	0.82	0.54
$C_{\beta N\alpha}$	(i−1,i,i)	0.71	0.83	0.50	0.72	0.83	0.50
$C_{\beta NN}$	(i,i,i+1)	0.63	0.87	0.19	0.63	0.87	0.18
$C_{\beta NN}$	(i,i,i−1)	0.64	0.87	0.21	0.64	0.87	0.22
$C_{\alpha NN}$	(i,i,i+1)	0.59	0.87	0.11	0.59	0.87	0.10
$C_{\alpha NN}$	(i,i,i−1)	0.58	0.87	0.09	0.58	0.87	0.09

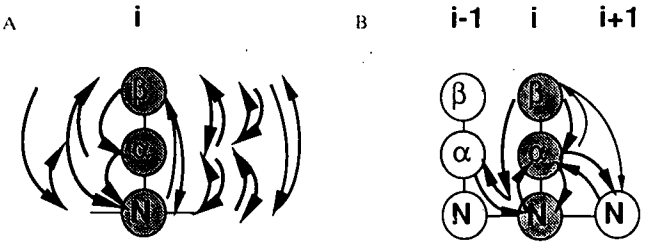

Fig. 4. Schematic representation of the pattern generated from the seeds $C_{\beta\alpha N}(i,i,i)$ and $C_{\alpha\beta N}(i,i,i)$. The darker shaded circles denote the original spins of the seeds. Lighter shaded circles denote the spins which are added because of correlated connectivities. Open circles denote the spins which may potentially be observed in helical domains. The arrows represent connectivities. Thick arrows correspond to high conditional probabilities (cf. Table 3). (A) Intra-residual connectivities. (B) Sequential connectivities.

constructed with identical appearance in the spectrum as that of the completely intra-residual one (compare Figs. 3A and B). For this reason some analogous connectivities are listed in Table 4. Comparison with Table 3 shows that they have lower conditional probabilities. This is illustrated once more by the smaller number and smaller sizes of the arrows in Fig. 5, which schematically represents this pattern.

$C_{N\alpha N}(i+1,i,i)$ and $C_{N\beta N}(i+1,i,i)$ -based patterns

We also evaluated the $C_{N\alpha N}(i+1,i,i)$ and $C_{N\beta N}(i+1,i,i)$ connectivities as potential seeds for a pattern. The intra-residual connectivities are again highly correlated (cf. Tables 5 and 6) and the patterns are expected to encompass the $C^{\alpha}H(i)$, $C^{\beta}H(i)$, NH(i), and NH(i+1) resonance frequencies. In addition, it can be concluded that the $C_{N\alpha N}(i+1,i,i)$-based pattern possibly can be extended with the $C^{\alpha}H(i-1)$ resonance frequency, the $C^{\beta}H(i-1)$ resonance frequency (especially in helical domains) and also, in case of helical domains, the NH(i−1) resonance frequency. Two

TABLE 4
CONDITIONAL PROBABILITIES $P(A|C_{\beta\alpha N}(i,i,i+1))$ AND $P(A|C_{\alpha\beta N}(i,i,i+1))$ OBTAINED FROM THE ANALYSIS OF 28 CRYSTAL PROTEIN STRUCTURES BY THE PROGRAM CSAND

A		$P(A\|C_{\beta\alpha N}(i,i,i+1))$			$P(A\|C_{\alpha\beta N}(i,i,i+1))$		
		All	Helical	β-Sheet	All	Helical	β-Sheet
$C_{\beta\alpha N}$	(i,i,i+1)	1.00	1.00	1.00	0.96	0.94	0.99
$C_{\alpha\beta N}$	(i,i,i+1)	0.79	0.95	0.77	1.00	1.00	1.00
$C_{N\alpha N}$	(i+1,i,i+1)	0.50	0.08	0.73	0.35	0.08	0.70
$C_{N\beta N}$	(i+1,i,i+1)	0.49	0.73	0.35	0.60	0.78	0.36
$C_{\alpha N\alpha}$	(i,i+1,i)	0.50	0.08	0.73	0.35	0.08	0.70
$C_{\alpha\beta\alpha}$	(i,i,i)	1.00	1.00	1.00	1.00	1.00	1.00
$C_{\beta N\beta}$	(i,i+1,i)	0.49	0.73	0.35	0.60	0.78	0.36
$C_{\beta\alpha\beta}$	(i,i,i)	1.00	1.00	1.00	1.00	1.00	1.00
$C_{N\alpha N}$	(i+1,i,i)	0.87	0.81	0.77	0.86	0.84	0.75
$C_{\beta\alpha N}$	(i,i,i)	0.94	0.96	0.97	0.94	0.96	0.96
$C_{\alpha\beta N}$	(i,i,i)	0.94	0.96	0.97	0.94	0.96	0.96

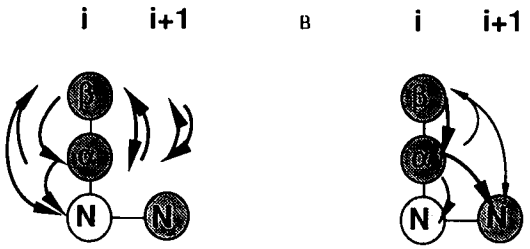

Fig. 5. Schematic representation of the pattern generated from the seeds $C_{\beta\alpha N}(i,i,i+1)$ and $C_{\alpha\beta N}(i,i,i+1)$. The darker shaded circles denote the original spins of the seeds. Lighter shaded circles denote the spins which are added because of correlated connectivities. The arrows represent connectivities. Thick arrows correspond to high conditional probabilities (cf. Table 4). (A) Intra-residual connectivities. (B) Sequential connectivities.

connectivities characteristic for β-sheet conformation can be expected, i.e. the back-transfer $C_{N\alpha N}(i+1,i,i+1)$ and the interesting connectivity $C_{\alpha\alpha N}(i+1,i,i+1)$. The latter combines two sequential NOEs, one of which is the uncommon $d_{\alpha\alpha}(i,i+1)$.

The $C_{N\beta N}(i+1,i,i)$-based pattern possibly can be extended with the $C_{\beta N\alpha}(i,i,i-1)$ connectivity which defines the $C^{\alpha}H(i-1)$ resonance frequency. Several interesting cross peaks may be expected for helical domains, defining the $C^{\beta}H(i-1)$, $NH(i-1)$ and $C^{\alpha}H(i-3)$ resonance frequencies.

NNN-based patterns

Five connectivities in the NNN-domain of the spectrum were also evaluated as possible seeds for a pattern. Patterns from these five seeds were generated under the requirement that a seed $C_{NNN}(i,j,k)$ only allows $C_{NNP}(i,j,p)$, and $C_{QNN}(q,j,k)$ as new pattern members. This more stringent rule is desirable in order to limit the number of potential connectivities. Without it, the pattern can be extended up to five residues along the polypeptide backbone. For example, the less

TABLE 5
CONDITIONAL PROBABILITIES $P(A|C_{N\alpha N}(i+1,i,i))$ OBTAINED FROM THE ANALYSIS OF 28 CRYSTAL PROTEIN STRUCTURES BY THE PROGRAM CSAND

A		$P(A \mid C_{N\alpha N}(i+1,i,i))$		
		All	Helical	β-Sheet
$C_{N\alpha N}$	(i,i,i)	1.00	1.00	1.00
$C_{\beta\alpha N}$	(i,i,i)	0.91	0.95	0.91
$C_{\beta N\alpha}$	(i,i,i)	0.86	0.94	0.84
$C_{\alpha N\alpha}$	(i,i,i)	1.00	1.00	1.00
$C_{\beta\alpha N}$	(i,i,i+1)	0.82	0.76	0.72
$C_{\alpha N\alpha}$	(i−1,i,i)	0.91	0.85	0.79
$C_{\beta N\alpha}$	(i−1,i,i)	0.61	0.80	0.40
$C_{\alpha NN}$	(i,i,i−1)	0.61	0.96	0.14
$C_{N\alpha N}$	(i+1,i,i+1)	0.56	0.14	0.93
$C_{\alpha\alpha N}$	(i+1,i,i+1)	0.33	0.03	0.72

TABLE 6
CONDITIONAL PROBABILITIES $P(A|C_{N\beta N}(i+1,i,i))$ OBTAINED FROM THE ANALYSIS OF 28 CRYSTAL PROTEIN STRUCTURES BY THE PROGRAM CSAND

A		$P(A \mid C_{N\beta N}(i+1,i,i))$		
		All	Helical	β-Sheet
$C_{N\beta N}$	(i,i,i)	0.83	0.92	0.75
$C_{\alpha\beta N}$	(i,i,i)	0.92	0.95	0.91
$C_{\beta N\alpha}$	(i,i,i)	0.88	0.94	0.84
$C_{\beta N\beta}$	(i,i,i)	0.83	0.92	0.75
$C_{\alpha\beta N}$	(i,i,i+1)	0.73	0.84	0.53
$C_{N\beta N}$	(i+1,i,i+1)	0.80	0.94	0.71
$C_{\beta N\alpha}$	(i,i,i−1)	0.88	0.92	0.88
$C_{\beta N\beta}$	(i−1,i,i)	0.68	0.80	0.49
$C_{\beta NN}$	(i,i,i−1)	0.65	0.92	0.20
$C_{\beta NN}$	(i,i,i+1)	0.66	0.86	0.14
$C_{\beta N\alpha}$	(i,i,i−3)	0.43	0.74	0.03

stringent rule would allow $C_{NNN}(i-1,i,i+1)$, $C_{NNN}(i-2,i-1,i)$ and $C_{NNN}(i,i+1,i+2)$ to be pattern members.

The C_{NNN}-based patterns are well defined and mainly observed for helical domains (cf. Table 2). Table 7 shows that the five C_{NNN}-based patterns are expected to encompass the NH(i−1), NH(i), NH(i+1), $C^{\alpha}H(i)$, $C^{\beta}H(i)$, $C^{\alpha}H(i-1)$, and $C^{\beta}H(i-1)$ resonance frequencies, although in the case of the $C_{NNN}(i+1,i,i+2)$-based pattern the latter two resonance frequencies are less likely due to the relative lower scores of the $C_{\alpha NN}(i-1,i,i+1)$ and $C_{\beta NN}(i-1,i,i+1)$. Table 7 also illustrates that the $C_{NNN}(i-1,i,i+1)$ connectivity has the highest number (16) of potential correlated connectivities. This pattern is schematically represented in Fig. 6. For all patterns the $C^{\alpha}H(i-3)$ resonance frequency may be defined in helical domains through medium-range connectivities.

The results clearly demonstrate that helical domains are very well defined and are therefore expected to yield reliable patterns. We realise that patterns involving inter-strand connectivities in β-sheet conformation will not show up in our analysis due to the fact that long-range NOEs were not discriminated from intra-strand NOEs. For these types of patterns a specific search would have to be conducted. However, our results show that a β-sheet conformation is less favourable compared to an α-helical conformation in so far as intra-residual, sequential- and medium-range connectivities are concerned.

On the basis of the results presented here, the following assignment strategy could be envisaged. First, patterns are constructed from the seeds in the βαN and αβN domains of the spectrum (αβN set), from seeds in the NαN and NβN domains of the spectrum (NαN/NβN set) and the NNN domain of the spectrum (NNN set). Secondly, the patterns in the three sets can be ordered (correlated) on the basis of similarities in the resonance frequencies encompassed by the patterns. Two patterns from different sets may encompass identical resonance frequencies (compare Figs. 4. and 6). Thus, all the patterns in one set can be compared with the patterns in one of the other two sets, so as to yield the similarly looking patterns of each set. Finally, the task amounts to finding a con-

TABLE 7
CONDITIONAL PROBABILITIES $P(A|B)$ OBTAINED FROM THE ANALYSIS OF 28 CRYSTAL PROTEIN STRUCTURES BY THE PROGRAM CSAND

A	B									
	$C_{NNN}(i-2,i,i-1)$		$C_{NNN}(i-2,i,i+1)$		$C_{NNN}(i-1,i,i+1)$		$C_{NNN}(i-1,i,i+2)$		$C_{NNN}(i+1,i,i+2)$	
	All	Helical	All	Helical	All	Helical	All	Helical	All	Helical
C_{NNN} $(i-2,i,i-1)$	1.00	1.00	0.52	0.51	0.29	0.27	0.47	0.45		
C_{NNN} $(i-2,i,i+1)$	0.41	0.43	1.00	1.00	0.22	0.22			0.20	0.20
C_{NNN} $(i-1,i,i+1)$	0.99	1.00	0.94	1.00	1.00	1.00	0.97	0.99	0.81	0.86
C_{NNN} $(i-1,i,i+2)$	0.31	0.35			0.19	0.20	1.00	1.00	0.24	0.28
C_{NNN} $(i+1,i,i+2)$			0.20	0.22	0.19	0.21	0.29	0.33	1.00	1.00
$C_{\alpha NN}$ $(i-3,i,i-1)$	0.47	0.73			0.59	0.81	0.75	0.94		
$C_{\alpha NN}$ $(i-3,i,i+1)$			0.58	0.72	0.53	0.70			0.48	0.63
$C_{\alpha NN}$ $(i-1,i,i-1)$	0.97	0.98			0.95	0.95	1.00	1.00		
$C_{\alpha NN}$ $(i-1,i,i+1)$			0.98	0.98	0.91	0.98			0.73	0.84
$C_{\alpha NN}$ $(i,i,i-1)$	0.83	0.91			0.80	0.88	0.82	0.86		
$C_{\alpha NN}$ $(i,i,i+1)$			0.95	1.00	0.96	1.00			1.00	1.00
$C_{\beta NN}$ $(i-1,i,i-1)$	0.82	0.94			0.81	0.92	0.90	0.94		
$C_{\beta NN}$ $(i-1,i,i+1)$			0.89	0.91	0.79	0.92			0.66	0.81
$C_{\beta NN}$ $(i,i,i-1)$	0.78	0.88			0.77	0.85	0.80	0.84		
$C_{\beta NN}$ $(i,i,i+1)$			0.89	0.95	0.90	0.95			0.86	0.95
C_{NNN} $(i-1,i,i-1)$	0.97	1.00			0.95	1.00	0.99	1.00		
C_{NNN} $(i+1,i,i+1)$			0.81	0.88	0.77	0.88			1.00	1.00

sistent ordering of all the patterns of the different sets, in conjunction with certain assumptions about the secondary structure of the molecule; e.g. helical domains in case of patterns from the NNN set. Investigations along this line are currently under way and will be reported elsewhere. Preliminary results, however, show already promising results for this approach.

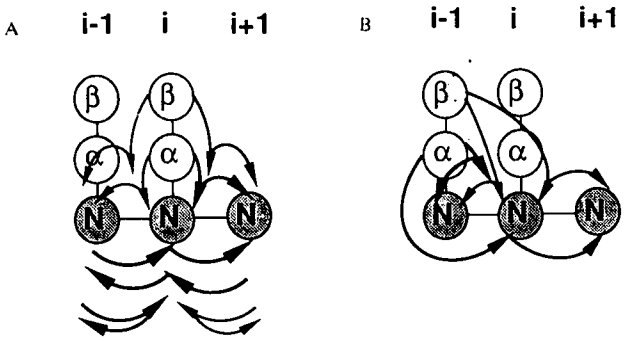

Fig. 6. Schematic representation of the pattern generated from the seed $C_{NNN}(i-1,i,i+1)$. The darker shaded circles denote the original spins of the seed. Lighter shaded circles denote the spins which are added because of correlated connectivities. The arrows represent connectivities. Thick arrows correspond to high conditional probabilities (cf. Table 7). (A) Connectivities involving $C^{\alpha}H(i)$, $C^{\beta}H(i)$, NH(i), NH(i−1), and NH(i+1). (B) Connectivities involving $C^{\alpha}H(i-1)$, $C^{\beta}H(i-1)$, NH(i), NH(i−1) and NH(i+1).

CONCLUSIONS

Our analysis represents a start for a novel assignment strategy in which no use is made of the traditional homonuclear J interaction, but which instead relies solely on the NOE interaction. Since the NOE effect is expected to become larger with increasing size of the molecules, this strategy is expected to be applicable to larger molecules. The 3D NOE-NOE experiment seems most suitable because of the increased resolution this experiment affords, combined with the novel connectivities which provide new and valuable information.

On the basis of the statistical evaluation of 28 crystal structures of the Brookhaven Protein Data Bank (Bernstein et al., 1977), we have identified 95 types of connectivities which are most likely to be observed as cross peaks in the 3D NOE-NOE spectrum of a protein. The results indicate that the a priori selectivity of these 95 connectivities usually is too low to safely base an assignment strategy upon. The large number of connectivities and the relatively low exclusiveness is due to the lack of J coupling information. We showed that our results were in agreement with the analysis of short distances in proteins of Billeter et al. (1982), if we included information about J interactions, which was implicitly assumed in their analysis.

The 95 connectivities served as a basis for the construction of patterns of related and highly correlated connectivities. A simple pattern-generating algorithm allows construction of a pattern starting from one or more seeds. New pattern members can be added under the requirement that they have at least one NOE in common with the seed(s). Given the occurrence of one of the connectivities (B), we evaluated the changes of occurrence of the others (A). Evaluation of these conditional probabilities $P(A \mid B)$ showed that a highly correlated pattern can be constructed from the seeds $C_{\beta\alpha N}(i,i,i)$ and $C_{\alpha\beta N}(i,i,i)$, encompassing the $C^{\alpha}H(i)$, $C^{\beta}H(i)$, NH(i), $C^{\alpha}H(i-1)$ and $NH(i+1)$ resonance frequencies and, in case of helical domains, also the $C^{\beta}H(i-1)$ and $NH(i-1)$ resonance frequencies.

Also the $C_{N\alpha N}(i+1,i,i)$ and $C_{N\beta N}(i+1,i,i)$ connectivities can be used as seeds for a potential pattern, identifying the $C^{\alpha}H(i)$, $C^{\beta}H(i)$, NH(i) and $NH(i+1)$ resonance frequencies. For helical domains, in addition the $C^{\alpha}H(i-1)$, $C^{\beta}H(i-1)$ and $NH(i-1)$ resonance frequencies may be identified, as well as the $C^{\alpha}H(i-3)$ resonance frequency in the case of the $C_{N\beta N}(i+1,i,i)$-based pattern. Beta-sheet domains may yield the $C^{\alpha}H(i+1)$ resonance frequency via the unusual $d_{\alpha\alpha}(i,i+1)$ NOE in case of a $C_{N\alpha N}(i+1,i,i)$-based pattern. Of great value, especially for the helical domains, seem to be the NNN-based patterns, identifying the $C^{\alpha}H(i)$, $C^{\beta}H(i)$, NH(i), $C^{\alpha}H(i-1)$, $C^{\beta}H(i-1)$, $NH(i-1)$ and $NH(i+1)$ resonance frequencies, and possibly also the $C^{\alpha}H(i-3)$ resonance frequency.

We have outlined the first steps toward a novel assignment strategy which could be complementary to recently developed methods based upon $^{15}N/^{13}C$ labelling and heteronuclear 3D and 4D NMR techniques, especially if structural information of homologous proteins is available. Clearly, in case no $^{15}N/^{13}C$-enriched proteins can be obtained, this method would represent a useful alternative.

ACKNOWLEDGEMENTS

This work was supported by The Netherlands Foundation for Chemical Research (SON) with financial aid from the Netherlands Organization for the Advancement of Research (NWO).

REFERENCES

Aue, W.P., Bartholdi, E. and Ernst, R.R. (1976) *J. Chem. Phys.*, **64**, 2229-2246.

Bax, A. and Davis, D.G. (1985) *J. Magn. Reson.*, **65**, 355–360.

Bax, A., Sparks, S.W. and Torchia, D.A. (1988) *J. Am. Chem. Soc.*, **110**, 7926–7927.

Bax, A., Clore, G.M. and Gronenborn, A.M. (1990) *J. Magn. Reson.*, **88**, 425–431.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) *J. Mol. Biol.*, **112**, 535–542.

Billeter, M., Braun, W. and Wüthrich, K. (1982) *J. Mol. Biol.*, **155**, 321–346.

Billingsley, P. (1986) *Probability and Measure*, John Wiley, New York.

Boelens, R., Vuister, G.W., Koning, T.M.G. and Kaptein, R. (1989) *J. Am. Chem. Soc.*, **111**, 8525–8526.

Braun, W. and Gō, N. (1985) *J. Mol. Biol.*, **186**, 611–626.

Breg, J., Boelens, R., Vuister, G.W. and Kaptein, R. (1990) *J. Magn. Reson.*, **87**, 646–651.

Braunschweiler, L. and Ernst, R.R. (1983) *J. Magn. Reson.*, **53**, 521–528.

Clore, G.M., Gronenborn, A.M., Brünger, A.T. and Karplus, M. (1985) *J. Mol. Biol.*, **186**, 435–455.

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

Englander, S.W. and Wand, A.J. (1987) *Biochemistry*, **26**, 5953–5958.

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

Fesik, S.W., Eaton, H.L., Olejniczak, E.T. and Zuiderweg, E.R.P. (1990) *J. Am. Chem. Soc.*, **112**, 886–888.

Griesinger, C., Sørensen, O.W. and Ernst, R.R. (1987a) *J. Magn. Reson.*, **73**, 574–579.

Griesinger, C., Sørensen, O.W. and Ernst, R.R. (1987b) *J. Am. Chem. Soc.*, **109**, 7227–7228.

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

Havel, T.F. and Wüthrich, K. (1984) *Bull. Math. Biol.*, **46**, 673–698.

Hempel, J.C. (1989) *J. Am. Chem. Soc.*, **111**, 491–495.

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

Jeener, J., Meier, B.H., Bachmann, P. and Ernst, R.R. (1979) *J. Chem. Phys.*, **71**, 4546–4553.

Kaptein, R., Zuiderweg, E.R.P., Scheek, R.M., Boelens, R. and van Gunsteren, W.F. (1985) *J. Mol. Biol.*, **182**, 179–182.

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

Kay, L.E., Ikura, M. and Bax, A. (1990b) *J. Am. Chem. Soc.*, **112**, 888–889.

Kleywegt, G.J. (1991) *Computer-assisted Assignment of 2D and 3D NMR Spectra of Proteins*, Ph. D. Thesis, Utrecht.

Kleywegt, G.J., Boelens, R., Cox, M., Llinás, M. and Kaptein, R. (1991) *J. Biomol. NMR*, **1**, 23–47.

Marion, D., Kay, L.E., Sparks, S.W., Torchia, D.A. and Bax, A. (1989a) *J. Am. Chem. Soc.*, **111**, 1515–1517.

Marion, D., Driscoll, P.C., Kay, L.E., Wingfield, P.T., Bax, A., Gronenborn, A.M. and Clore, G.M. (1989b) *Biochemistry*, **28**, 6150–6156.

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

Neuhaus, D. and Williamson, M. (1989) *The Nuclear Overhauser Effect in Structural and Conformational Analysis*, VCH Publishers Inc., New York.

Oschkinat, H., Griesinger, C., Kraulis, P.J., Sørensen, O.W., Ernst, R.R., Gronenborn, A.M. and Clore, G.M. (1988) *Nature*, **332**, 374–376.

Oschkinat, H., Cieslar, C., Holak, T.A., Clore, G.M. and Gronenborn, A.M. (1989) *J. Magn. Reson.*, **83**, 450–472.

Oschkinat, H., Cieslar, C. and Griesinger, C. (1990) *J. Magn. Reson.*, **86**, 453–469.

Padilla, A., Vuister, G.W., Boelens, R., Kleywegt, G.J., Cavé, A., Parello, J. and Kaptein, R. (1990) *J. Am. Chem. Soc.*, **112**, 5024–5030.

Scheek, R.M., Russo, N., Boelens, R., van Boom, J.H. and Kaptein, R. (1984) *Biochemistry*, **23**, 1371–1376.

Vuister, G.W. and Boelens, R. (1987) *J. Magn. Reson.*, **73**, 328–333.

Vuister, G.W., Boelens, R. and Kaptein, R. (1988) *J. Magn. Reson.*, **80**, 176–185.

Vuister, G.W., Boelens, R., Padilla, A., Kleywegt, G.J. and Kaptein, R. (1990) *Biochemistry*, **29**, 1829–1839.

Wester, W.M., Stockman, B.J. and Markley, J.L. (1988) *J. Am. Chem. Soc.*, **110**, 6256–6258.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York.

Zuiderweg, E.R.P. and Fesik, S.W. (1989) *Biochemistry*, **28**, 2387–2391.

Zuiderweg, E.R.P., Petros, A.M., Fesik, S.W. and Olejniczak, E.T. (1991) *J. Am. Chem. Soc.*, **113**, 370–372.