Journal of Biomolecular NMR, **18:** 129–137, 2000.
KLUWER/ESCOM

Sequence-specific NMR assignment of proteins by global fragment mapping with the program MAPPER

Peter Güntert*, Michael Salzmann**, Daniel Braun & Kurt Wüthrich
Institut für Molekularbiologie und Biophysik, ETH-Hönggerberg, CH-8093 Zürich, Switzerland

Received 5 May 2000; Accepted 11 July 2000

Key words: automated assignment, NMR assignment of proteins, program MAPPER, sequence-specific assignment, triple resonance experiments

Abstract

A new program, MAPPER, for semiautomatic sequence-specific NMR assignment in proteins is introduced. The program uses an input of short fragments of sequentially neighboring residues, which have been assembled based on sequential NMR connectivities and for which either the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts or data on the amino acid type from other sources are known. MAPPER then performs an exhaustive search for self-consistent simultaneous mappings of all these fragments onto the protein sequence. Compared to using only the individual mappings of the spectroscopically connected fragments, the global mapping adds a powerful new constraint, which results in resolving many otherwise intractable ambiguities. In an initial application, virtually complete sequence-specific assignments were obtained for a 110 kDa homooctameric protein, 7,8-dihydroneopterin aldolase from *Staphylococcus aureus.*

Introduction

Improved efficiency of NMR assignments in proteins by partial or full automation has attracted much attention (for a recent review see Moseley and Montelione, 1999) because sequence-specific assignments are the foundation for three-dimensional structure determination and other detailed investigations on conformation, dynamics and function by NMR (Wüthrich, 1986). Overall, one is usually faced with the situation that connectivities between neighboring amino acid residues can readily be established for short stretches of the polypeptide chain, using either sequential nuclear Overhauser effects (NOE) (Billeter et al., 1982; Wagner and Wüthrich, 1982) or data from triple resonance experiments with ^{13}C,^{15}N-labeled proteins (Bax and Grzesiek, 1993). In typical globular proteins the placement of such spectroscopically assembled peptide segments in the amino acid sequence ('sequence-specific assignment') is in most cases unambiguous if the segment contains at least three residues and the amino acid types are known (Wüthrich, 1986). However, although straightforward in principle, unique NMR identification of the amino acid types is a limiting factor when using either homonuclear ^{1}H-NMR with smaller proteins (Wüthrich, 1983) or the data available from sequential assignments with triple resonance NMR techniques (Bax and Grzesiek, 1993). It is therefore of interest to develop procedures for support of sequence-specific assignments based on spectroscopic identification of short peptide segments with only partially known sequence, for example, –Ala–(Arg or Lys)–(Glu or Gln or Met)– (e.g., Friedrichs et al., 1994; Meadows et al., 1994; Morelle et al., 1995; Zimmerman et al., 1997).

This paper presents the program MAPPER, which is a tool for obtaining sequence-specific resonance assignments on the basis of spectroscopically assembled short segments of sequentially connected residues ('fragments'). These fragments are mapped onto the polypeptide primary structure using partial knowledge of their amino acid types. When using sequential

*To whom correspondence should be addressed. E-mail: guenter@mol.biol.ethz.ch
**Present address: Bruker AG, Industriestrasse 26, CH-8117 Fällanden, Switzerland.

Table 1. $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shift reference values for use with the program MAPPER[a]

Amino acid, *R*	$\tilde{\omega}_R^{\alpha} \pm \Delta\tilde{\omega}_R^{\alpha}$ (ppm)[b]	$\tilde{\omega}_R^{\beta} \pm \Delta\tilde{\omega}_R^{\beta}$ (ppm)[b]
Ala	52.59 ± 2.17	18.72 ± 1.95
Arg	56.52 ± 2.47	30.73 ± 1.92
Asn	53.15 ± 1.95	38.13 ± 1.79
Asp	54.06 ± 2.00	40.24 ± 2.00
Cys	56.23 ± 3.38	37.09 ± 6.33
Gln	55.62 ± 2.29	29.28 ± 2.13
Glu	56.62 ± 2.43	29.81 ± 1.89
Gly	44.81 ± 1.48	–
His	55.54 ± 2.40	29.95 ± 2.91
Ile	60.88 ± 2.74	38.83 ± 2.57
Leu	54.76 ± 2.19	42.25 ± 2.51
Lys	56.41 ± 2.09	32.59 ± 2.05
Met	55.10 ± 1.91	32.23 ± 2.62
Phe	57.16 ± 2.28	39.87 ± 2.15
Pro	62.85 ± 1.37	31.74 ± 1.54
Ser	57.74 ± 1.92	63.93 ± 1.75
Thr	61.76 ± 2.56	69.31 ± 1.98
Trp	57.34 ± 2.52	29.24 ± 2.12
Tyr	57.07 ± 2.37	38.78 ± 2.30
Val	61.37 ± 2.76	32.87 ± 2.12

[a]This table lists chemical shift values that have been assembled in 1995 from a database of the following 25 proteins, which have been checked and corrected for consistent referencing as described in footnote b: *Antennapedia* homeodomain (Qian et al., 1993), cyclophilin (Ottiger et al., 1997), DNA-binding domain of Gal4 (Shirakawa et al., 1993), interleukin-1 receptor antagonist (Stockman et al., 1994), staphylococcal nuclease (Wang et al., 1992), Fk506 binding protein (Xu et al., 1993), ribonuclease H (Yamazaki et al., 1993), development-specific Ca^{2+}-binding protein S (Bagby et al., 1994), ferrocytochrome (Caffrey et al., 1994), interleukin-1β (Clore et al., 1990), 26-10 antibody Vl domain (Constantine et al., 1993), glucose permease IIa domain (Fairbrother et al., 1992), pathogenesis-related protein P14a (Fernández et al., 1997), urokinase-type plasminogen activator (Hansen et al., 1994), BPTI (Hansen et al., 1995), second RNA-binding domain of the sex-lethal protein (Lee et al., 1994), Pec-60 (Liepinsh et al., 1994), N-terminal domain of DNA-polymerase β (Liu et al., 1994), barstar (Lubienski et al., 1994), villin (Markus et al., 1994), tendamistat (Matter and Kessler, 1995), ferredoxin (Oh and Markley, 1990), IIIglc (Pelton et al., 1991), interleukin-4 (Powers et al., 1992), ovomucoid third domain (Robertson et al., 1990). We continue to use these values in our group, since they show no significant discrepancy to the continuously updated corresponding table in the BioMagResBank (www.bmrb.wisc.edu), which nowadays relies on a much bigger set of assigned proteins. Continued use of 'our' chemical shift data is favored by us since it enables a comparison of results obtained over the years with the use of different approaches. For applications of MAPPER it can readily be substituted by the current BioMagResBank database, which leads to virtually identical results.

[b]$\tilde{\omega}_R^{\alpha}$ and $\tilde{\omega}_R^{\beta}$ are the average values, and $\Delta\tilde{\omega}_R^{\alpha}$ and $\Delta\tilde{\omega}_R^{\beta}$ the standard deviations of the chemical shifts in the aforementioned 25 proteins for which nearly complete sequence-specific assignments are available. The shifts were referenced consistently following Wishart et al. (1995), and obvious outliers were removed during the preparation of the database.

resonance assignments based on triple resonance experiments such as HNCA (Kay et al., 1990; Montelione and Wagner, 1990), CBCA(CO)NH (Grzesiek and Bax, 1992) and HNCACB (Wittekind and Mueller, 1993), residue-specific information consists primarily of the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shift values (Richarz and Wüthrich, 1978; Oh et al., 1988; Grzesiek and Bax, 1993). Alternatively, this information may consist of a classification into spin system types by homonuclear ^{1}H-NMR (Wüthrich, 1986) or by other techniques (e.g., Dötsch et al., 1996; Schubert et al., 1999). The MAPPER package includes a library of the average values and standard deviations of the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts of the 20 amino acid residues (Table 1). The input for the program consists of the amino acid sequence of the protein, and the list of fragments of sequentially connected residues, with information on the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ shift values and/or identification of the compatible amino acid types from other sources.

The program MAPPER

In the following description of the MAPPER algorithm for use with chemical shift data for characterization of amino acid types, $\Re = \{\text{Ala}, \ldots, \text{Val}\}$ denotes the set of the 20 standard amino acids. $R(k) \in \Re$ is the amino acid type at position $k = 1, \ldots, N$ in the protein sequence of N residues. The reference chemical shift value and its standard deviation for the atom a of the amino acid type $R \in \Re$ are given by $\tilde{\omega}^a_R \pm \Delta\tilde{\omega}^a_R$ (Table 1). Fragments, F_i, are numbered from 1 to N_F. A fragment F_i spans $n(i) + 1$ sequentially adjacent residues, and $\omega^a_j(i)$ denotes the experimental chemical shift for the atom $a \in A_j(i)$ at the residue position $j = 0, \ldots, n(i)$ within the fragment, where $A_j(i)$ denotes the set of atoms at position j in the fragment F_i for which chemical shift values are available.

In the first step of the MAPPER assignment procedure each fragment is treated independently, and all locations in the amino acid sequence of the protein are determined to which one or several of the fragments can be mapped. If sufficient relevant information is available for a given fragment, an unambiguous sequence-specific assignment may be obtained already from this 'individual mapping' step (Wüthrich, 1983, 1986; Grzesiek and Bax, 1993), but conceptually each individual mapping is only part of an intermediate result that represents the input for the subsequent 'global mapping'. To determine the acceptable individual mappings for a given fragment i, the sum of the squared deviations of the chemical shift values in F_i from the corresponding reference chemical shift values at the positions $k, \ldots, k + n(i)$ in the amino acid sequence of the protein is computed:

$$\chi^2(i;k) = \sum_{j=0}^{n(i)} \sum_{a \in A_j(i)} \left[\frac{\omega^a_j(i) - \tilde{\omega}^a_{R(k+j)}}{\Delta\tilde{\omega}^a_{R(k+j)}} \right]^2 . \quad (1)$$

Assuming that the distributions for the chemical shifts are Gaussian and that the attempted mapping is correct, the probability that the magnitude of the sum of the squared relative chemical shift deviations exceeds the value computed in Equation 1 is given by the χ^2 probability function $Q\left(\chi^2(i;k)\,|\nu_i\right)$ (Equation 6.2.18 in Press et al., 1986), where $\nu_i = \sum_{j=0}^{n(i)} |A_j(i)|$ is the number of known chemical shifts in the fragment F_i (corresponding expressions may be derived with the assumption of different, non-Gaussian chemical shift distributions). Acceptable individual mappings have a value of $Q\left(\chi^2(i;k)\,|\nu_i\right)$ above a user-defined threshold Q_0. $Q_0 = 1\%$ was used for the calculations in this paper.

In the second step of the MAPPER assignment, an exhaustive search for 'global mappings', i.e., simultaneous, self-consistent mappings of all fragments, is performed on the basis of the accepted individual mappings. Similar approaches have been used in other semiautomated assignment programs (Friedrichs et al., 1994; Meadows et al., 1994; Morelle et al., 1995; Zimmerman et al., 1997). A global mapping is consistent if all N_F fragments can be mapped onto the sequence such that only permissible overlap occurs between any two fragments. The C-terminal residue in a fragment may carry an 'overlap' attribute to indicate that it can be placed at a sequence location which is already occupied by the N-terminal residue of another fragment. This situation occurs routinely if fragments are assembled by analyzing 'strips' (Bartels et al., 1995) taken from triple resonance spectra that provide both intraresidual and sequential connectivities to C^{α} and C^{β} atoms for a given backbone amide group. Two fragments may then share a common residue position only if the corresponding chemical shift values for the same atoms match within a user-defined tolerance of, typically, 0.4 ppm for ^{13}C. In nearly complete sets of short fragments there are usually many individual mapping possibilities, which may, however, be combined only in a limited number of ways into global mappings. An efficient exhaustive search for global mappings, which consists, in principle, of N_F nested

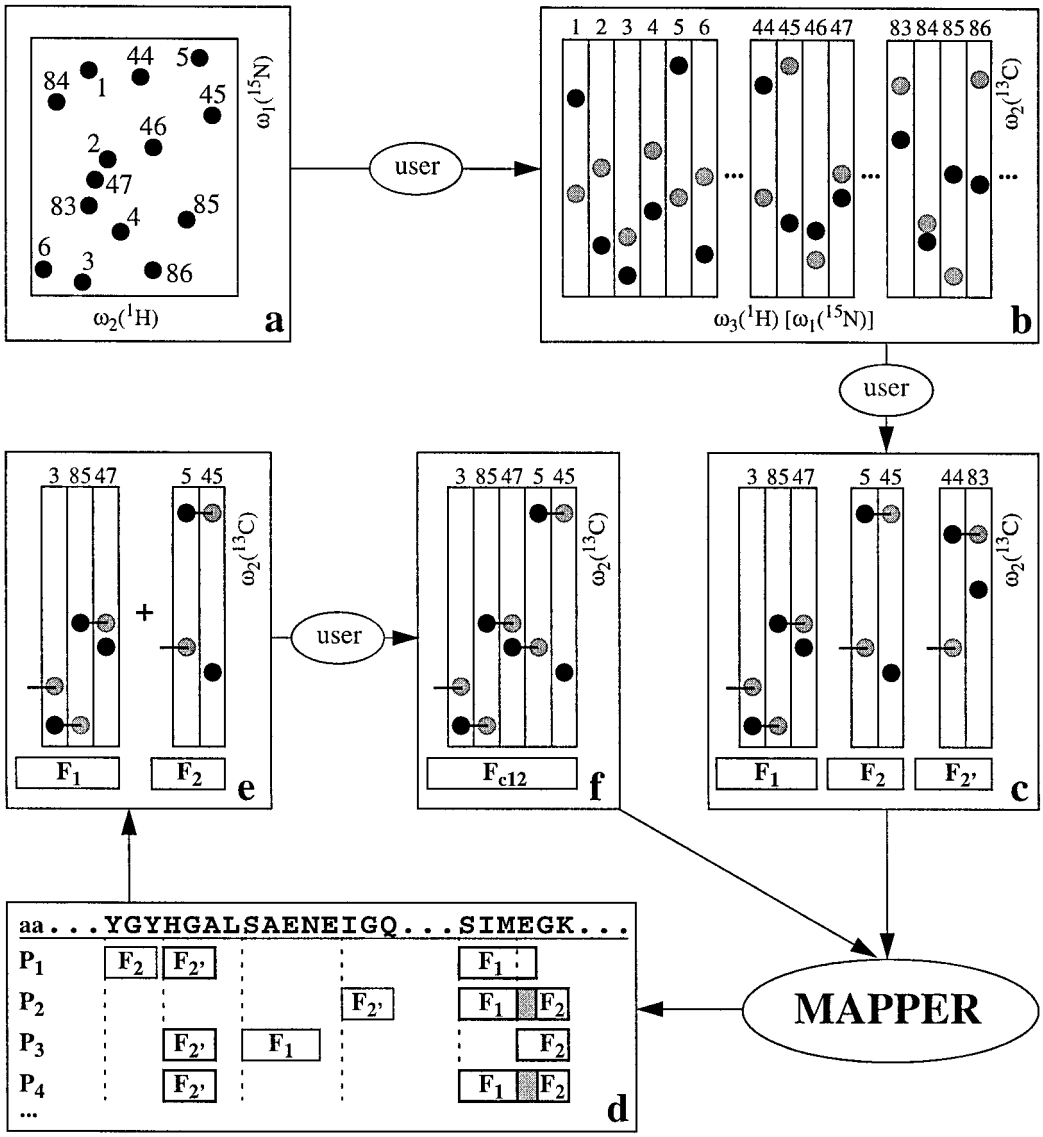

Figure 1. Schematic representation of an assignment process using the programs XEASY and MAPPER. (a) All peaks in a 2D [^{15}N,^{1}H]-HSQC spectrum are picked, leading to a peak list that contains the ^{15}N and ^{1}H frequencies for each amide moiety of the polypeptide backbone. (b) The ^{15}N and ^{1}H chemical shifts define the locations of 2D [ω_2(^{13}C), ω_3(^{1}H)] strips from the 3D triple resonance experiments used for the sequential assignment of ^{13}C,^{15}N- or ^{2}H,^{13}C,^{15}N-labeled proteins. The arbitrarily numbered schematic [ω_2(^{13}C), ω_3(^{1}H)] strips shown here are from a three-dimensional HNCA experiment and were taken at the ω_1(^{15}N) position and centered about the ω_3(^{1}H) chemical shift of each amide ^{1}H^{N}–^{15}N moiety of the polypeptide chain. The black and gray circles indicate intraresidual and sequential peaks in the spectrum, respectively. (c) Sequential connectivities are established interactively by matching the chemical shifts of sequential and intraresidual peaks in pairs of [ω_2(^{13}C), ω_3(^{1}H)] strips, as indicated by the horizontal lines. The steps (a) to (c) are supported by XEASY, and the resulting short peptide fragments with unambiguously assigned sequential connectivities, F_1, F_2 and $F_{2'}$, are used as input for MAPPER. (d) Output from MAPPER: The amino acid sequence of the protein is shown at the top, and each of the rows P_1, P_2, ... represents one possible 'global mapping', i.e., one possibility for simultaneously mapping the three fragments F_1, F_2 and $F_{2'}$ to specified positions in the polypeptide chain. The vertical dotted lines indicate the various possible N-terminal ends for a fragment, and the thickness of the boxes around F_1, F_2 and $F_{2'}$ indicates the quality of the individual fit as expressed by $\chi^2(i;k)$ (Equation 1) and $Q(\chi^2(i;k)|\nu_i)$ (see text), where a thick line indicates a good fit. P_2 and P_4 propose a sequential connection between the fragments F_1 and F_2, as indicated by the gray shading. (e) The sequential connectivity between the two fragments F_1–F_2 proposed by MAPPER is further evaluated interactively by inspection of the original spectra. If the two fragments can be unambiguously connected, the new, longer fragment F_{c21} (f) is used as part of the input for the next round of MAPPER calculations. The alternative of connecting F_1 with $F_{2'}$, which has a sequential peak at the same ω_2(^{13}C) chemical shift as F_2, can be ruled out since none of the global mappings from MAPPER would be consistent with a sequential connectivity F_1–$F_{2'}$.

loops, is therefore obtained if 'forbidden' branches of the search tree are cut as early as possible. To this end, the loops are nested such that the outermost loops correspond to fragments with few individual mappings, whereas the innermost loops belong to fragments with many individual mapping possibilities. The global mappings found by MAPPER are ranked according to χ^2(global), which is defined as the sum over all fragments of the individual χ^2 values of Equation 1. An overall probability, *Q*(global), for the global mapping is defined in the same way as the above-mentioned probability for individual mappings, $Q\left(\chi^2\left(i;k\right)|\nu_i\right)$. A value of *Q*(global) close to 100% indicates that a global mapping is 'reasonable' in the sense that, overall, the chemical shift deviations are within the range expected statistically on the basis of their standard deviations, $\Delta\tilde{\omega}_R^a$, but it does not exclude the presence of other, similarly reasonable global mappings. The program MAPPER is written in Fortran-77 and is available from P. Güntert.

The assignment procedure using MAPPER for obtaining sequence-specific assignments with amino acid type information from chemical shifts is outlined in Figure 1. At the outset, a list of amide proton and ^{15}N chemical shifts is established by picking the peaks in a 2D [^{15}N,^{1}H]-HSQC spectrum (Figure 1a). This arbitrarily numbered list defines the locations of two-dimensional strips in 3D triple resonance spectra such as HNCA, HNCACB or CBCA(CO)NH, with the ω_2(^{13}C) chemical shift along the third axis (Figure 1b). Sequential connectivities are then identified, for example, with the program XEASY (Bartels et al., 1995), by visually matching the ^{13}C chemical shifts of sequential and intraresidual correlation peaks in different [ω_2(^{13}C), ω_3(^{1}H)] strips (Figure 1c). This yields an initial set of short fragments assembled on the basis of unambiguous connectivities between pairs of strips from the triple resonance spectra. These fragments, for instance F_1, F_2 and $F_{2'}$ in Figure 1c, contain the $^{13}C^{\alpha}$ shifts and, if available, the $^{13}C^{\beta}$ chemical shifts of the sequentially linked residues. Typically, many potential sequential connectivities are ambiguous at this stage because of chemical shift degeneracies, and therefore the unambiguously connected segments remain short. For example, in Figure 1c the fragment F_1 could have been connected equally well with either of the fragments F_2 or $F_{2'}$. MAPPER identifies all possible global mappings, P_1, P_2,..., of this preliminary set of fragments (Figure 1d). Of prime interest are mappings for which MAPPER proposes an overlap between two fragments, as between the fragments F_1 and F_2 for the possibilities P_2 and P_4 (Figure 1d). On the other hand, many potential sequential connectivities between pairs of strips may be ruled out on the basis of the intermediate results from MAPPER, i.e., whenever the two fragments cannot be mapped onto adjacent sequence locations (F_1 and $F_{2'}$ in Figure 1d). Thereby the number of potential sequential connectivities that have to be checked visually by consulting the 3D spectra is obviously reduced when compared to the initial assignments of sequentially neighboring strips (Figure 1e). As a result of this step of the procedure, pairs of fragments can be fused into longer fragments. For example, after visual inspection the two fragments F_1 and F_2 were combined into a new, longer fragment, F_{c12} (Figure 1, e and f), which was then used in the place of F_1 and F_2 as input for a subsequent MAPPER run. The alternative connectivity between the fragments F_1 and $F_{2'}$, which would also be compatible with the coinciding ^{13}C frequencies (Figure 1c), is not supported by MAPPER, since the fragment $F_{2'}$ is mapped with high probability to a different sequence position. This semiautomatic procedure is repeated until all assignment ambiguities are resolved.

Application to the protein DHNA

The use of the program MAPPER was a key step in obtaining nearly complete sequential assignments for 7,8-dihydroneopterin aldolase (DHNA) from *Staphylococcus aureus*, a symmetric homooctamer protein of molecular weight 110 kDa with 121 residues per monomer (Salzmann et al., 2000). Fragments were assembled using strips from several TROSY-type triple resonance experiments (Salzmann et al., 1998, 1999). Unambiguous resonance assignments could be established for all C^{α} atoms except Met 1, Gln 2, Pro 103, Ile 114 and Glu 115, and for 77 out of the 111 C^{β} atoms (Salzmann et al., 2000). Examples of the MAPPER input and output files for DHNA are shown in Figure 2, a and c. The deviations of the experimental $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts from the corresponding reference values (Table 1) for the best global mapping, which, in the case of DHNA, corresponds to the correct sequence-specific assignment, are plotted in Figure 2d. Note that, given the standard deviations of the chemical shifts in Table 1, a spread of up to about 4 ppm (corresponding to about two standard deviations) is expected. Although large differences within this spread could clearly jeopardize

a

```
...
fragment 8
   53.70
   59.78              P-
   51.85    43.31     P-
   60.00    34.48     P-
   61.48    69.28     P-
   51.88    41.88     P-
   55.22              P-
   59.39    36.00     P-
   54.09    39.40     P- O

fragment 9
   54.07
   54.08    39.10     P-
   61.29              P- O
...
```

b

```
...
fragment 8
   EKMPQR
   V
   CDFHNSWY
   V
   T
   L
   EKMPQR
   V
   CDFHNSWY  O

fragment 9
   CDFHNSWY
   L
   CDFHNSWY  O
...
```

c

Fragment		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Length		2	5	4	2	2	13	11	9	3	13	9	15	5	12	5	4	8	6
Ind		34	13	11	60	78	1	1	1	19	1	1	1	1	1	3	3	1	1
Ind/Glob		1	1	1	2	3	1	1	1	1	1	1	1	1	1	1	1	1	1
Rank	Score	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
P1	76.82	104	75	2	1	6	7	19	30	38	40	52	61	80	84	95	99	106	116
P2	77.36	.	.	.	6	1	.	.	.	.	.	.	.	.	.	.	.	.	.
P3	78.67	.	.	.	.	114	.	.	.	.	.	.	.	.	.	.	.	.	.
P4	80.13	.	.	.	6	114	.	.	.	.	.	.	.	.	.	.	.	.	.
Best mapping:																			
Rank		10	1	1	22	37	1	1	1	4	1	1	1	1	1	2	1	1	1
Score contrib.		1	2	1	2	1	8	8	7	3	11	9	14	4	10	3	1	9	6
Probability %		52	88	87	16	35	96	77	73	28	80	51	64	70	87	61	99	44	72

Overall probability: 99.79 %

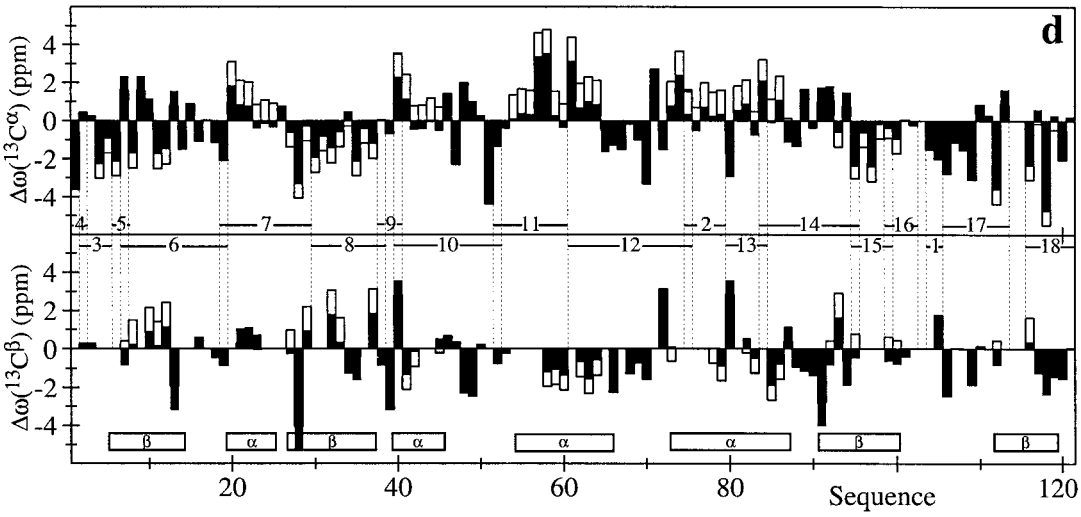

Figure 2. Application of the program MAPPER using amino acid type information from $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts for obtaining sequence-specific resonance assignments in the protein DHNA. (a) Excerpt from the input file for MAPPER. The data for two fragments, F_8 and F_9, which span 9 and 3 residues, respectively, are shown. $^{13}C^{\alpha}$ shifts are listed in the first column. Where available, $^{13}C^{\beta}$ shifts are given in the second column. The string 'P-' means 'this residue cannot be mapped onto a proline in the primary structure' (proline can only occur as the first residue in any given fragment). 'O' indicates that the last residue of the fragment may be mapped onto a position in the amino acid sequence that is already occupied by the first residue of another fragment. (b) Alternative input file using amino acid type classes instead of $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ shifts (see text). The letters indicate allowed amino acid types, which are denoted using the standard one-letter code for amino acids, e.g., 'CDFHNSWY' denotes that the residue in question can be any of the 8 –αCH-βCH$_2$– AMX spin systems, and 'EKMPQR' refers to 'long side chains' as defined by Wüthrich (1986). (c) Output from MAPPER. Each column contains data for one of the 18 fragments that constituted the input for MAPPER, as derived from the NMR experiments (Salzmann et al., 2000), as follows: Fragment: fragment number. Length: number of residues in the fragment. Ind: number of possible individual mappings. Ind/Global: number of individual mappings that form part of a self-consistent global mapping. Next is a table of all global mappings, in this case $P1, \ldots, P4$, containing the total score (Equation 1) and the sequence locations of the first residue in each fragment, where a dot indicates that the fragment is mapped to the same position as in the best global mapping, P_1. The final section of the output affords a statistics of the best global mapping, indicating for each fragment the rank of the individual mapping possibility used in the best global mapping, its score, χ^2(ind), measured as a percentage of the overall score, χ^2(global), and Q(ind), the probability $Q(\chi^2(i;k)\,|\nu_i)$ of the individual mapping used. Note that long fragments, such as number 10, may have a comparatively large contribution to the overall score while still maintaining a high Q(ind) value. Finally, the overall probability of the best global mapping, Q(global), is given (see text). (d) Plots versus the amino acid sequence of the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shift deviations, $\Delta\omega = \omega_j^a(i) - \tilde{\omega}^a_{R(k+j)}$ (Equation 1), between the experimental chemical shifts in the best global mapping and the reference chemical shift values of Table 1. The sequence locations to which the fragments $F_1, \ldots, F_{18}$ are mapped in the best global mapping are indicated by horizontal lines in the center, and the locations of regular secondary structures in DHNA are shown at the bottom of the figure.

the correct sequence-specific assignment of individual residues, they can be accommodated in a meaningful fashion in the present global mapping approach. The well-known variation of the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ chemical shifts in different regular secondary structures (Spera and Bax, 1991) can be accounted for by MAPPER if the secondary structure of the protein under study is known. This is achieved by adding offsets of $+1.25/-0.75$ ppm in α-helical regions or $-0.75/+1.25$ ppm in β-sheet regions to the $^{13}C^{\alpha}/^{13}C^{\beta}$ reference chemical shift values, $\tilde{\omega}^{\alpha}_R$ and $\tilde{\omega}^{\beta}_R$. In the case of DHNA, inclusion of secondary structure information led to the same best global mapping as the one obtained without use of this information, albeit with a significantly reduced global χ^2 value of 57.4 instead of 76.8. In other words, although the secondary structure information is not required for obtaining the correct assignment, use of the known secondary structure chemical shifts (Spera and Bax, 1993) will lead to an improved global χ^2 value if the locations of the regular secondary structures are known.

Discussion and conclusions

The MAPPER algorithm, in which all fragments are simultaneously mapped onto the primary structure, is more powerful in finding the correct sequence-specific assignments than the simpler approach of searching for individual fragment mappings (Wüthrich, 1983; Grzesiek and Bax, 1993), since the global mapping approach excludes mutually contradicting individual fragment mappings. This is apparent from Figure 2c and became particularly obvious in the results of a trial MAPPER run for DHNA with an input of 38 non-overlapping three-residue fragments (fragments containing Pro were not considered) that were attributed the experimental C^{α} and, if available, C^{β} shifts. Using a threshold of $Q_0 = 0.01$ there were between 1 and 69 (on average 10) acceptable individual mappings for these fragments. This would result in 2.1×10^{29} possible combinations of the 38 fragments if contradictions between the placement of individual fragments were not taken into account. For 25 fragments the χ^2 values for the individual fragment mapping (Equation 1) provided a basis for discriminating between the correct assignment and other possibilities, whereas for the remaining 13 fragments the individual mapping with the smallest χ^2 value was not the correct one. The program MAPPER found 2 433 397 possible global mappings, of which the one with the lowest global χ^2 value contained the correct sequential assignments for all fragments.

The global mapping algorithm of MAPPER can readily exploit incomplete amino acid type classifications in the fragments from other sources than ^{13}C chemical shift data. To include this information in the MAPPER algorithm a new notation is defined, as follows: The sequence locations onto which the j-th residue within the fragment F_i can be mapped are described by sets $S_j(i) \subseteq \{1, \ldots, N\}$. In the absence of a priori restrictions on possible mappings, all positions would be allowed, i.e., $S_j(i) = \{1, \ldots, N\}$. Or else, if the j-th residue in the fragment F_i is known

to be glycine, then $S_j\,(i)$ comprises the sequence locations of all glycine residues. To illustrate the use of MAPPER with this type of input data, we consider the grouping of the 20 amino acids into 8 distinct classes that was used in early protein assignments: Gly, Ala, Val, Leu, Ile, Thr, which could usually be identified uniquely, the $-C^{\alpha}H-C^{\beta}H_2-$ AMX spin systems (Ser, Cys, Asp, Asn, Phe, Tyr, His, Trp), and the long side chains (Gln, Glu, Met, Pro, Arg, Lys) (Wüthrich, 1986). Acceptable individual mappings need to be compatible with these restrictions on allowed residue positions: The fragment i can only be mapped to the position starting with residue $k \in \{1, \ldots, N\}$ in the sequence if $(k + j) \in S_j\,(i)$ for all $j = 0, \ldots, n\,(i)$. With this type of input data (Figure 2b) the continuous χ^2 function (Equation 1) is substituted by a binary yes/no choice to indicate whether or not a given mapping is acceptable for MAPPER. Two fragments of the same length that are identical on the level of the amino acid type classification can therefore not be distinguished by MAPPER. The global mapping condition nonetheless constitutes a powerful constraint on the number of possible assignments. For instance, when applying MAPPER in a trial run for DHNA with fragments of length three residues and the aforementioned classification into eight amino acid types, one finds 786 432 possible combinations of individual mappings for the 40 fragments, whereas there are 96 possible global mappings. For completeness' sake it should be added that MAPPER functions also with a mixed input of ^{13}C chemical shift data and amino acid type identifications. This was used, for example, in the experimentally assembled fragments for the MAPPER assignment of DHNA to indicate that proline can be incorporated only at the start of a fragment (Figure 2a).

In conclusion, MAPPER is a new, powerful tool to establish sequence-specific resonance assignments in proteins, which can significantly speed up the overall assignment. The requirement of self-consistency for a global mapping provides a strong additional constraint, which is not taken into account if fragments are treated individually. It can be envisaged that the combination of MAPPER with an automatic method for assembling short fragments of unambiguously connected amino acid residues, for example, from triple resonance experiments with ^{15}N,^{13}C-labeled proteins or from sequential NOE connectivities in unlabeled or ^{15}N-labeled proteins, will yield an efficient and reliable, fully automated method for sequence-specific resonance assignment in proteins.

References

Bagby, S., Harvey, T.S., Kay, L.E., Eagle, S.G., Inouye, S. and Ikura, M. (1994) *Biochemistry*, **33**, 2409–2421.

Bartels, C., Xia, T., Billeter, M., Güntert, P. and Wüthrich, K. (1995) *J. Biomol. NMR*, **6**, 1–10.

Bax, A. and Grzesiek, S. (1993) *Acc. Chem. Res.*, **26**, 131–138.

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

Caffrey, M., Brutscher, B., Simorre, J.P., Fitch, J., Cusanovich, M. and Marion, D. (1994) *Eur. J. Biochem.*, **221**, 63–75.

Clore, G.M., Bax, A., Driscoll, P.C., Wingfield, P.T. and Gronenborn, A.M. (1990) *Biochemistry*, **29**, 8172–8184.

Constantine, K.L., Goldfarb, V., Wittekind, M., Friedrichs, M.S., Anthony, J., Ng, S.C. and Mueller, L. (1993) *J. Biomol. NMR*, **3**, 41–54.

Dötsch, V., Oswald, R.E. and Wagner, G. (1996) *J. Magn. Reson.*, **B110**, 107–111.

Fairbrother, W.J., Palmer, A.G., Rance, M., Reizer, J., Saier, M.H. and Wright, P.E. (1992) *Biochemistry*, **31**, 4413–4425.

Fernández, C., Szyperski, T., Bruyère, T., Ramage, P., Mösinger, E. and Wüthrich, K. (1997) *J. Mol. Biol.*, **266**, 576–593.

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

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

Grzesiek, S. and Bax, A. (1993) *J. Biomol. NMR*, **3**, 185–204.

Güntert, P., Mumenthaler, C. and Wüthrich, K. (1997) *J. Mol. Biol.*, **273**, 283–298.

Hansen, A.P., Petros, A.M., Meadows, R.P., Nettesheim, D.G., Mazar, A.P., Olejniczak, E.T., Xu, R.X., Pederson, T.M., Henkin, J.and Fesik, S.W. (1994) *Biochemistry*, **33**, 4847–4864.

Hansen, P.E., Lauritzen, C., Petersen, L.C., Björn, S., Norris, K. and Olsen, O.H. (1995) *J. Cell. Biochem.*, 72–79.

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

Lee, A.L., Kanaar, R., Rio, D.C. and Wemmer, D.E. (1994) *Biochemistry*, **33**, 13775–13786.

Liepinsh, E., Berndt, K.D., Sillard, R., Mutt, V. and Otting, G. (1994) *J. Mol. Biol.*, **239**, 137–153.

Liu, D.J., Derose, E.F., Prasad, R., Wilson, S.H. and Mullen, G.P. (1994) *Biochemistry*, **33**, 9537–9545.

Lubienski, M.J., Bycroft, M., Freund, S.M.V. and Fersht, A.R. (1994) *Biochemistry*, **33**, 8866–8877.

Markus, M.A., Nakayama, T., Matsudaira, P. and Wagner, G. (1994) *J. Biomol. NMR*, **4**, 553–574.

Matter, H. and Kessler, H. (1995) *J. Am. Chem. Soc.*, **117**, 3347–3359.

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

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

Morelle, N., Brutscher, B., Simorre, J.P. and Marion, D. (1995) *J. Biomol. NMR*, **5**, 154–160.

Moseley, H.N.B. and Montelione, G.T. (1999) *Curr. Opin. Struct. Biol.*, **9**, 635–642.

Oh, B.H. and Markley, J.L. (1990) *Biochemistry*, **29**, 3993–4004.

Oh, B.H., Wrestler, W.M., Darba, P. and Markley, J.L. (1988) *Science*, **240**, 908–910.

Ottiger, M., Zerbe, O., Güntert, P. and Wüthrich, K. (1997) *J. Mol. Biol.*, **272**, 64–81.

Pelton, J.G., Torchia, D.A., Meadow, N.D., Wong, C.Y. and Roseman, S. (1991) *Biochemistry*, **30**, 10043–10057.

Powers, R., Garrett, D.S., March, C.J., Frieden, E.A., Gronenborn, A.M. and Clore, G.M. (1992) *Biochemistry*, **31**, 4334–4346.
Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986) *Numerical Recipes. The Art of Scientific Computing*, Cambridge University Press, New York, NY.
Qian, Y.Q., Otting, G., Billeter, M., Müller, M., Gehring, W. and Wüthrich, K. (1993) *J. Mol. Biol.*, **234**, 1070–1083.
Richarz, R. and Wüthrich, K. (1978) *Biopolymers*, **17**, 2263–2269.
Robertson, A.D., Rhyu, G.I., Westler, W.M. and Markley, J.L. (1990) *Biopolymers*, **29**, 461–467.
Salzmann, M., Pervushin, K., Wider, G., Senn, H. and Wüthrich, K. (1998) *Proc. Natl. Acad. Sci. USA*, **95**, 13585–13590.
Salzmann, M., Pervushin, K., Wider, G., Senn, H. and Wüthrich, K. (2000) *J. Am. Chem. Soc.*, **122**, 7543–7548.
Salzmann, M., Wider, G., Pervushin, K., Senn, H. and Wüthrich, K. (1999) *J. Am. Chem. Soc.*, **121**, 844–848.
Schubert, M., Smalla, M., Schmieder, P. and Oschkinat, H. (1999) *J. Magn. Reson.*, **141**, 34–43.
Shirakawa, M., Fairbrother, W.J., Serikawa, Y., Ohkubo, T., Kyogoku, Y. and Wright, P.E. (1993) *Biochemistry*, **32**, 2144–2153.
Spera, S. and Bax, A. (1991) *J. Am. Chem. Soc.*, **113**, 5490–5492.
Stockman, B.J., Scahill, T.A., Strakalaitis, N.A., Brunner, D.P., Yem, A.W. and Deibel Jr., M.R. (1994) *FEBS Lett.*, **349**, 79–83.
Wagner, G. and Wüthrich, K. (1982) *J. Mol. Biol.*, **155**, 347–366.
Wang, J.F., Mooberry, E.S., Walkenhorst, W.F. and Markley, J.L. (1992) *Biochemistry*, **31**, 911–920.
Wishart, D.S., Bigam, C.G., Yao, J., Abildgaard, F., Dyson, H.J., Oldfield, E., Markley, J.L. and Sykes, B.D. (1995) *J. Biomol. NMR*, **6**, 135–140.
Wittekind, M. and Mueller, L. (1993) *J. Magn. Reson.*, **B101**, 201–205.
Wüthrich, K. (1983) *Biopolymers*, **22**, 131–138.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York, NY.
Xu, R.X., Nettesheim, D., Olejniczak, E.T., Meadows, R., Gemmecker, G. and Fesik, S.W. (1993) *Biopolymers*, **33**, 535–550.
Yamazaki, T., Yoshida, M. and Nagayama, K. (1993) *Biochemistry*, **32**, 5656–5669.
Zimmerman, D.E., Kulikowski, C.A., Huang, Y.P., Feng, W.Q., Tashiro, M., Shimotakahara, S., Chien, C.Y., Powers, R. and Montelione, G.T. (1997) *J. Mol. Biol.*, **269**, 592–610.