Journal of Biomolecular NMR, **11:** 31–43, 1998.
KLUWER/ESCOM

Automated backbone assignment of labeled proteins using the threshold accepting algorithm

Michael Leutner, Ruth M. Gschwind, Jens Liermann, Christian Schwarz, Gerd Gemmecker & Horst Kessler*
Institut für Organische Chemie und Biochemie, TU München, D-85747 Garching, Germany

Received 14 May 1997; Accepted 24 July 1997

Key words: automated assignment, combinatorial minimization, proteins, threshold accepting

Abstract

The sequential assignment of backbone resonances is the first step in the structure determination of proteins by heteronuclear NMR. For larger proteins, an assignment strategy based on proton side-chain information is no longer suitable for the use in an automated procedure. Our program PASTA (Protein ASsignment by Threshold Accepting) is therefore designed to partially or fully automate the sequential assignment of proteins, based on the analysis of NMR backbone resonances plus C^{β} information. In order to overcome the problems caused by peak overlap and missing signals in an automated assignment process, PASTA uses *threshold accepting*, a combinatorial optimization strategy, which is superior to simulated annealing due to generally faster convergence and better solutions. The reliability of this algorithm is shown by reproducing the complete sequential backbone assignment of several proteins from published NMR data. The robustness of the algorithm against misassigned signals, noise, spectral overlap and missing peaks is shown by repeating the assignment with reduced sequential information and increased chemical shift tolerances. The performance of the program on real data is finally demonstrated with automatically picked peak lists of human nonpancreatic synovial phospholipase A_2, a protein with 124 residues.

Introduction

The sequential assignment of the protein backbone signals is the basic step in the NMR structure determination process for larger proteins. Since this is a tedious and time-consuming task, automation is highly desirable. For small proteins, numerous programs have been developed to partially or fully automate the assignment using the classical strategy first outlined by Wüthrich (1986); for a detailed review see Zimmerman and Montelione (1995). This approach identifies the proton spin systems in homonuclear COSY/TOCSY experiments and then connects them sequentially via information obtained from NOESY experiments. While suitable for smaller proteins (<10 kDa), this approach usually fails for larger proteins due to increasing signal overlap. The strategy can be expanded to larger proteins by the use of 3D ^{15}N-edited TOCSY and NOESY experiments, as demonstrated by the program ALFA (Bernstein et al., 1993). However, this extension is again limited to ca. 12–15 kDa proteins, due to the relaxation-sensitive ^{1}H,^{1}H-TOCSY transfer. Experiments like HC(C)NH-TOCSY and HC(CO)NH-TOCSY (Logan et al., 1992; Montelione et al., 1992) rely on the faster ^{13}C,^{13}C-TOCSY transfer and provide a unique combination of sequential proton spin system information. But even this approach requires deuterated samples when applied to proteins in the 20 kDa range or larger (for a review see Sattler and Fesik (1996)).

Therefore, a generally applicable program for the automated assignment of larger proteins should not rely on side-chain information in the initial sequential assignment process. More suitable is a suite of heteronuclear 3D experiments tracing the protein backbone (Grzesiek and Bax, 1992), including C^{β} information (Wittekind and Mueller, 1993; Yamazaki et al.,

* To whom correspondence should be addressed.

1994). For example, recently the backbone assignment of the 37 kDa Trp repressor / DNA complex could be manually obtained with this strategy on deuterated samples (Shan et al., 1996).

Common processing software, like Felix (MSI/ Biosym), Triad (Tripos), Aurelia (Neidig et al., 1995), and Xeasy (Bartels et al., 1995), usually provide only a graphical interface for the data. They mainly offer sophisticated book-keeping to aid the human 'step-by-step' way of assignment, while their facilities for automated assignment are limited. To automate certain tasks, some packages (e.g., Triad, Felix) provide built-in macro languages. The program GARANT (Bartels et al., 1996) uses a totally different approach to sequence-specific assignment. It mainly performs an intelligent peak-picking of the NMR spectra, based on a structure homology approach. It is therefore limited to families of homologous proteins, where the error rate for the calculation of chemical shifts from secondary structure predictions is low.

Even when tracing the backbone connectivities, programs have to deal with the ambiguity of experimental data, caused by heavy signal overlap in larger proteins. Here programs following a direct approach like the CONTRAST macro set (Olson and Markley, 1994) or AUTOASSIGN2 (Zimmerman and Montelione, 1995) can fail due to the 'combinatorial explosion' problem, where the decision tree created from all possible neighbors is not solvable anymore. The use of 4D experiments (Friedrichs et al., 1994; Meadows et al., 1994) to solve the overlap problem is limited because of reduced signal intensity resulting in missing peaks and because of restricted digital resolution (Bax and Grzesiek, 1993).

In order to reach an assignment for heavily overlapped data, combinatorial minimization strategies as used in the program ALFA (Bernstein et al., 1993) seem to be more promising, especially when applied to the backbone data, as shown by the program ALPS (Morelle et al., 1995) or more recently in the program of Lukin et al. (1997). A limitation of the classical simulated annealing procedure (Kirkpatrick et al., 1983) is the slow convergence of the algorithm. In larger proteins, the solution space grows rapidly with the number of residues and cannot be searched extensively in practical time scales, unless additional constraints are used to reduce the search space (such as the protein sequence in combination with the secondary structure (Lukin et al., 1997)).

The combinatorial minimization strategy used in PASTA, called *threshold accepting* (Dueck and Scheuer, 1990; Dueck et al., 1993), has proven to be significantly faster and providing better solutions than simulated annealing in solving the *traveling salesman problem* (TSP (Lawler et al., 1985)) and in structure minimizations (Morales et al., 1992). The complexity of the backbone assignment problem for proteins is comparable in size to the mentioned TSP. Therefore, it is likely that threshold accepting will succeed in protein backbone assignment with a minimum number of restraints. For example, the determination of the amino acid type, based on the C^{α}-C^{β} chemical shift combination, is quite ambiguous and therefore not essential for the primary sequential assignment. In this paper we will show on simulated data that the algorithm finds the global minimum, i.e., the correct assignment for several published proteins. It is further shown that the correct backbone assignment with automatically picked peak lists from triple-resonance spectra of human nonpancreatic synovial phospholipase A_2 (hnps-PLA_2) (124 residues) can be achieved just with sequential backbone (C^{α}/C^{β}/CO) information.

Materials and Methods

Description of the algorithm

PASTA uses ASCII peak lists in the same format as they are exported by common NMR processing software (Felix (MSI/Biosym), Triad (Tripos)). Due to the expected overlap of sequential information, the program is written to use any combination of C^{α}, C^{β}, CO, and H^{α} chemical shifts as source of information for the sequence-specific assignment. The flexible design of the program will accept all possible combinations of common 3D spectra. The processing of data from new experiments providing suitable information can be easily implemented.

First, an initial set of pseudo-residues (i.e., data structures to be filled with the backbone information based on $^{1}H^{N}/^{15}N$ pairs) is created from the peak lists of the ^{15}N-HSQC and/or HNCO spectra. Intraresidual and sequential information (i.e., C^{α}, C^{β}, CO, and H^{α} shifts) is then automatically added by searching the peak lists of the appropriate 3D triple-resonance spectra for the corresponding $^{1}H^{N}/^{15}N$ pairs. In general, spectral information such as HNCACB and HN(CO)CACB data is combined to enable the program to uniquely distinguish between the (i) and (i$-$1) signals of a $^{1}H^{N}/^{15}N$-pair. If this discrimination is not possible due to missing (i$-$1) information, then (i) and (i$-$1) signals are distinguished according to their

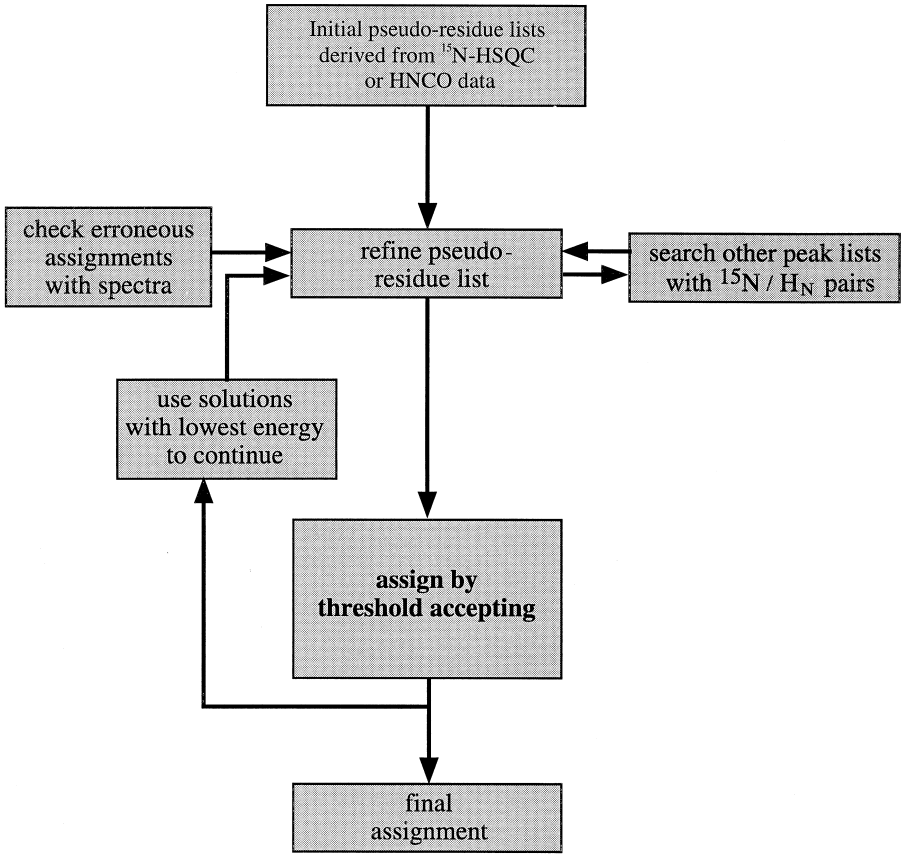

Figure 1. Assignment cycle of PASTA. An initial pseudo-residue list is created from the peak lists of the ^{15}N-HSQC or HNCO spectra. Additional information is added by searching the peak lists of the appropriate 3D experiments. The refinement of the list is done iteratively with the use of the assignment routine.

intensities (implemented for C^{α}, C^{β}, and H^{α}). Noise ridges in the spectra interfering with the identification of real cross peaks or missing signals are pointed out to the user by the program. Chemical shift offsets in any dimension between the different input spectra can be corrected with user given values, relative to the chemical shifts stored in the pseudo-residue list. A flow chart displaying the steps of the assignment process with PASTA is shown in Figure 1.

Our program can also use the shift patterns of C^{α} and C^{β} to perform an amino acid type determination (Grzesiek and Bax, 1993) based on published randomcoil shifts (Wüthrich, 1986; Wishart et al., 1995). Without knowledge of the protein's secondary structure influencing the chemical shift ranges, only glycine, alanine, threonine, serine and, to a lesser extent, valine are uniquely discernible by their chemical shifts (proline is observed in H^N-based spectra only as the (i−1) residue in certain experiments). Thus, a mapping of the result from the assignment process onto the known amino acid sequence may be used as *optional* additional constraint in an extended assignment process. In this extended algorithm, a four-residue wide window is shifted stepwise along the pseudo-residue list and residue types identified from C^{α}/C^{β} shift data are compared with the amino acid patterns of the sequence. This procedure about doubles the computational time for each assignment run, but is able to resolve the fragmentation of the resulting assignment due to overlapping signals. However, it is *not required* for the assignment process, as shown in our test data below.

Usually the pseudo-residue set has to be refined iteratively, in order to remove inconsistencies (e.g., side-chain amide signals and artifacts) from automatically picked lists. Pseudo-residues with too many or too few matching signals in respect to the spectrum are outlined to the user, together with the tolerances

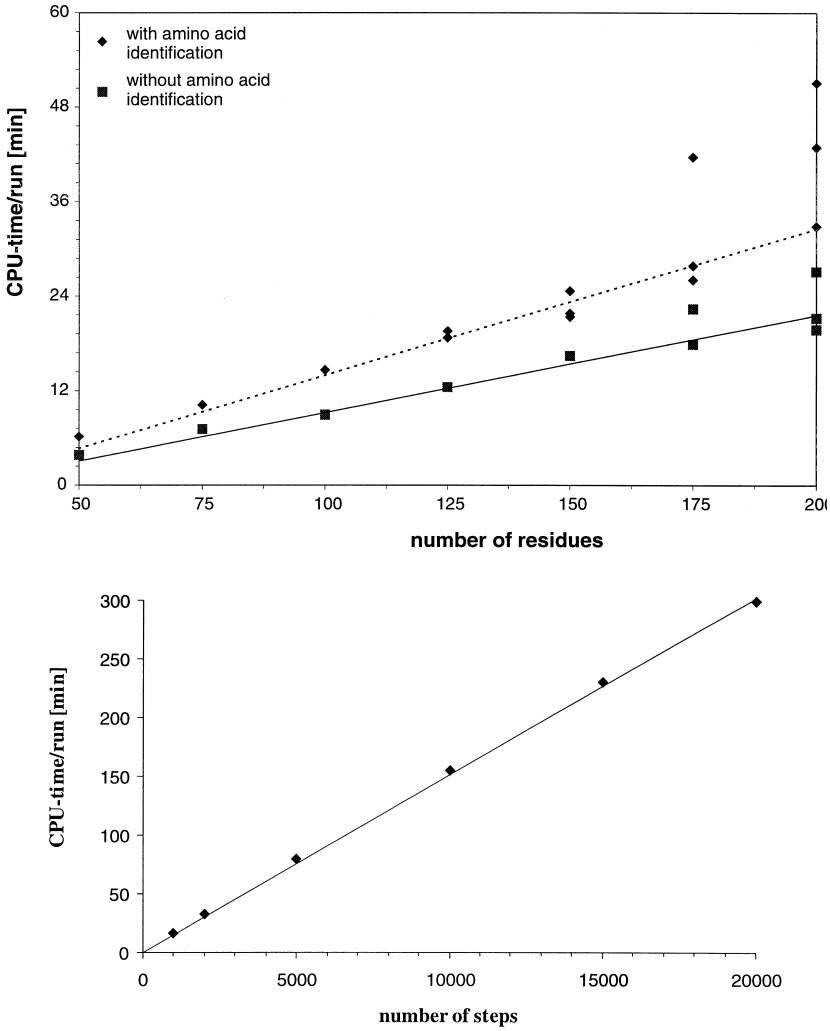

Figure 2. (Top) Time dependence of the program PASTA on the number of residues used, for runs with and without amino acid identification. The trend lines indicate CPU time consumption for an ideal case, while specific sequences can take longer, depending on the amount of signal overlap. (Bottom) CPU time dependence of PASTA on the maximum number of steps for each threshold accepting cycle (for 150 residues with $dE_{start} = 200$).

used. If too many signals are encountered, no automatic decision is made, but a list is given containing all possible sequential matches. Thus, the automatic generation of additional false entries in the pseudo-residue list is avoided. Furthermore, pseudo-residues which cannot be sequentially connected (e.g., side-chain amide signals) are identified by the program and can be cross-checked manually with the spectra.

Ambiguity of data as well as missing signals are a challenge for automated sequential assignment. Therefore, a combinatorial minimization strategy was chosen for the assignment algorithm instead of a deterministic approach. The algorithm of threshold accepting is simple and easy to implement. It contains four basic steps:

(1) Start at a random solution x_i.

(2) Generate a new solution x_{i+1} via a random local change of the solution x_i.

(3) Judge the quality of both solutions with a penalty function f. If the value of the penalty function for solution x_{i+1} is not larger than the penalty function for the solution x_i plus a user-defined threshold T, proceed to solution x_{i+1}; else discard solution x_{i+1}.

$$\text{if } [f(x_{i+1}) \leq f(x_i) + T] \text{ then } x_{i+1} \rightarrow x_i \qquad (1)$$

(4) Repeat steps (2) and (3). If for a user-given number of steps no improvement of the current minimum is achieved, reduce the threshold T stepwise to zero. The result corresponds to the best solution encountered during the whole run.

The penalty function or pseudo-energy to be minimized in the sequential assignment is defined as

$$E_{TOT} = E_{MATCH}\{+E_{SEQ}\} \quad (2)$$

E_{MATCH} describes the fit between two adjacent residues and is added up from the pseudo-energy contributions for the individual matching sequential chemical shifts $E_{C\alpha}$, $E_{C\beta}$, $E_{H\alpha}$, E_{CO}, and E_N. For the backbone assignment the following weights have been found to produce the best results: -12 ($^{13}C^{\alpha}$), -12 ($^{13}C^{\beta}$), -10 ($^{1}H^{\alpha}$), -15 (^{13}CO), and -20 (^{15}N). In case of a mismatch for at least one of the available sequential shifts, the pseudo-energy for the whole residue is set to $+130$ instead; any missing sequential information is ignored. Within the tolerances set for the individual nuclei, all matching shifts are equally accepted; there is no preference for the 'best match' to avoid a bias from insignificant chemical shift differences. The optional term E_{SEQ} is an additional feature resulting from mapping the initially obtained assignment onto the amino acid sequence as described above for the extended procedure.

Two different strategies are implemented to obtain a new solution x_{i+1} from the existing solution x_i:

(1) An interchange of two randomly chosen residues (the random number generator used is described as *ran3* in Press et al. (1992)).

(2) A 'cut and paste' of a larger fragment. The starting point, length, and new position of the fragment are again determined by the random number generator.

To allow a global search in the beginning of the assignment process, the starting value of the threshold (dE) has to be large enough, so that nearly every solution is accepted by the algorithm. Thus, dE_{start} must be larger than the contributions of a mismatch (130) minus the possible number of matching sequential informations (n) weighted with their average contribution (-15):

$$dE_{start} = 130 - n \cdot (-15) \quad (3)$$

This will allow the program in the beginning even to disrupt a perfectly matching pair of residues (=loss of $n \cdot (-15)$ *bonus* units) and substitute it with a mismatch (causing an additional 130 units penalty). No explicit bias is used to preserve already correctly aligned fragments. During the course of the PASTA run the pseudo-energies decrease (when already longer fragments of sequentially matching residues have been found); at the same time the threshold T for accepting worse solutions is also gradually lowered. Thus, the algorithm increasingly tends to preserve correctly aligned fragments: if a residue *within* a correct fragment is interchanged with a nonmatching residue, the increasing pseudo-energy would now probably lead to a rejection by the threshold criterion.

The time dependence of the algorithm on the number of residues is shown in Figure 2a. Without mapping the result of the assignment process onto the amino acid sequence, the correlation is nearly linear. The *extended* version of the algorithm including the amino acid identification might show, in the worst case, a square dependence on the length of the list. Nevertheless, the average time dependence is linear even in this case (data not shown).

However, with real data significant deviations from the theoretical time dependence show up, especially for runs with amino acid identification. These can be observed particularly with large or mainly α-helical proteins and are due to heavy overlap of the chemical shift information, causing slower convergence of the algorithm. This phenomenon is clearly visible in Figure 2a for list lengths larger than 150 aa.

The expected linear time dependence of PASTA on the maximum number of steps for each threshold accepting cycle is shown in Figure 2b. If the number of steps is set too small for an effective global search for a given protein, the resulting assignments will show more than two *real errors* (the definition of real errors is given below).

The program PASTA was written in ANSI C and compiled on a Silicon Graphics Indy R4000 (using the *cc -o2* command without further optimization). For basic screen handling the *curses(3X)* library routines are used which are part of the UNIX SVR4. Therefore, PASTA can be easily ported to all systems supporting this package. All input and output is performed via ASCII files for easy manipulation with common editors. A regularly updated version of the program PASTA will be available via the Internet under http://ociialf.org.chemie.tu-muenchen.de/people/jl.

Tests on published data

To test the reliability of the algorithm, pseudo-residue lists were created from the published NMR data of the following four proteins: interleukin 4 (Powers et al., 1992), interferon γ (Grzesiek et al., 1992), calmod-

Table 1. Tolerances for the chemical shifts used in the assignment runs for interferon γ.

H^{α} protons (ppm)	0.010	0.025	0.035	0.050
Heteronuclei $^{13}C^{\alpha}$, C^{β}, CO (ppm)	0.10	0.25	0.35	0.50

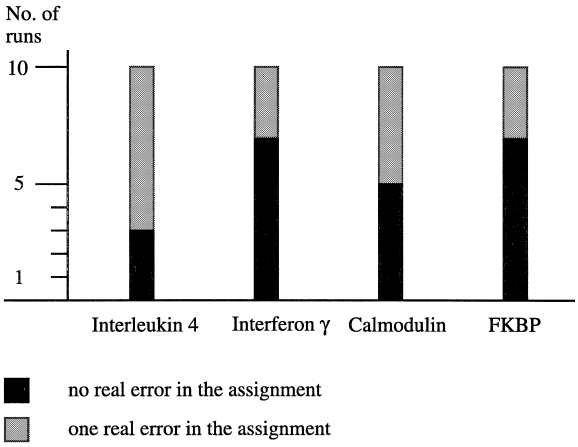

Figure 3. Results of 10 assignment runs using the complete data set with tolerances of 0.010 ppm for the protons and 0.10 ppm for the heteronuclei in all four proteins. No real error corresponds to 100%, one real error to about 99% correct assignment, depending on the length of the different proteins.

ulin (Ikura et al., 1990), and FKBP (Xu et al., 1993). The sequential information comprised C^{α}, CO, and H^{α} shifts for interferon γ and FKBP, whereas the data sets for calmodulin and interleukin 4 also included the C^{β} assignments. PASTA assignment runs with the complete data sets were performed for all four proteins; the amino acid sequence was *not* used as additional information in any test run. Since 'perfect' data sets were provided, relatively low tolerances of 0.01 ppm for the H^{α} protons and 0.1 ppm for the heteronuclei were used.

In the case of interferon γ the assignment was repeated with four different sets of chemical shift tolerances (Table 1).

To test the robustness of PASTA against missing data, additional data sets lacking various amounts of sequential information were created from the pseudo-residue list of interleukin 4:

(1) one specific sequential information removed: C^{α}, C^{β}, CO, or H^{α};

(2) two specific types of sequential information removed: (C^{α} and C^{β}), (CO and H^{α}), or (C^{β} and H^{α}); and

(3) randomly removed sequential information: 10%, 20%, or 50%.

All assignment runs were done with a starting threshold of $dE_{start} = 200$ and 15 000 steps.

For the evaluation of the time dependence of the program on the length of the residue list, artificial residue lists with the following numbers of amino acids were created from the calmodulin data set (148 aa) (for lists exceeding this length, different stretches of residues were added from interleukin 4 data): 50, 75, 100, 125, 150, 175, 200. All the above runs were done with $dE_{start} = 150$ and 1000 steps.

For measuring the time dependence on increasing the search space, the following numbers of steps were used: 1000, 2000, 5000, 10 000, 15 000, 20 000. These runs were done with $dE_{start} = 150$ and a residue-list out of calmodulin containing 100 amino acids.

Application to hnps-PLA_2

For the assignment of hnps-PLA_2 (Kramer et al., 1989; Seilhamer et al., 1989) a 2D $^{1}H^{N}/^{15}N$ projection was calculated from the 3D HNCA spectrum and picked with the Triad/Sybyl peak-picking routine. The obtained $^{1}H^{N}/^{15}N$ pairs served as input to create the initial pseudo-residue list. This initial list was then refined with the optimization algorithm of the program, using the combination of the peak lists from the HNCA/HN(CO)CA spectra. The sequential assignment was carried out first with the C^{α} information only. In a second step, the peak lists of the HNCACB/HN(CO)CACB or, alternatively, the combination HNCO/HNHA/HCACO were searched on the basis of the refined pseudo-residue list. The sequential assignment was then carried out with the combinations (C^{α}, C^{β}), (C^{α}, CO), (C^{α}, C^{β}, protein sequence), and (C^{α}, C^{β}, CO). Tolerances of 0.15 ppm were used for C^{α}, C^{β} and CO in all assignment runs. All assignment runs were done with $dE_{start} = 200$ and 15 000 steps. The CPU time per run was about 160 min on an SGI R4600SC/133 MHz.

Results and Discussion

Application to artificial data sets of four published proteins

The use of artificial data sets from published assignments allows one to judge the quality of the solutions obtained by the algorithm. Therefore, series of data sets with decreasing quality of the experimental sequential information were used to simulate some as-

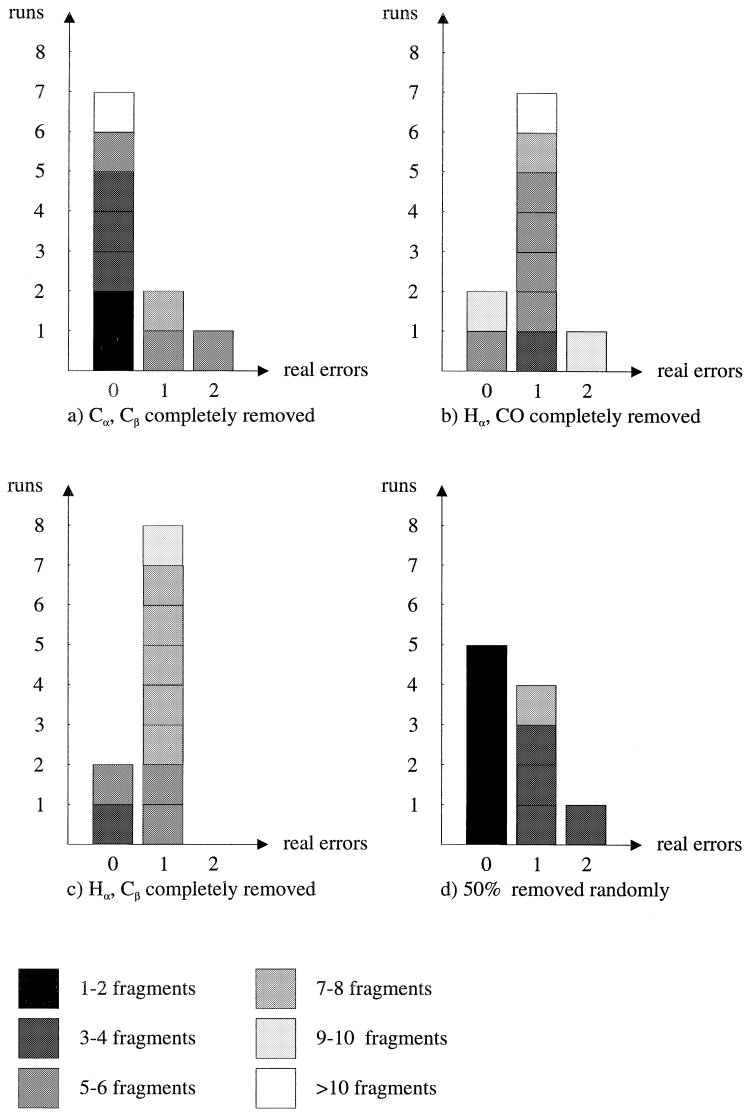

Figure 4. Number of fragments obtained in the calculation of interleukin 4 with artificially reduced data sets. In each case 50% of the sequential information (C^α, C^β, CO, H^α) was omitted. This was achieved either by a complete removal of two nuclei (a, b, c) or by random removal of 50% of all four nuclei (d). The bar graphs show the fragmentation distribution for 10 assignment runs for each of the four data sets divided by the number of real errors.

pects of real data, i.e., missing signals, low resolution of the spectra and noise. However, in real proteins the problems are often clustered in certain areas of the sequence, e.g., signal overlap in unstructured loop regions or missing signals in regions with internal motion. In order to demonstrate the program's ability to cope with this, the application of PASTA is shown also for the real data set of hnps-PLA_2.

A combinatorial minimization strategy will not necessarily reach the global minimum, i.e., the correct assignment, in all cases. Nevertheless, the result should be a very good solution (i.e., a local minimum) close to the correct assignment. It has to be distinguished between *real errors* introduced by the algorithm (when a sequential residue does not fit with its precursor, i.e., the algorithm fails to find the existing global minimum) and *fragments* produced due to spectral overlap which cannot be resolved with the available data. The correct assignment (no real errors) can easily be recognized among several runs, since real errors *always* lead to an increase in the pseudo-energy. In the case of fragmentation it is not possible

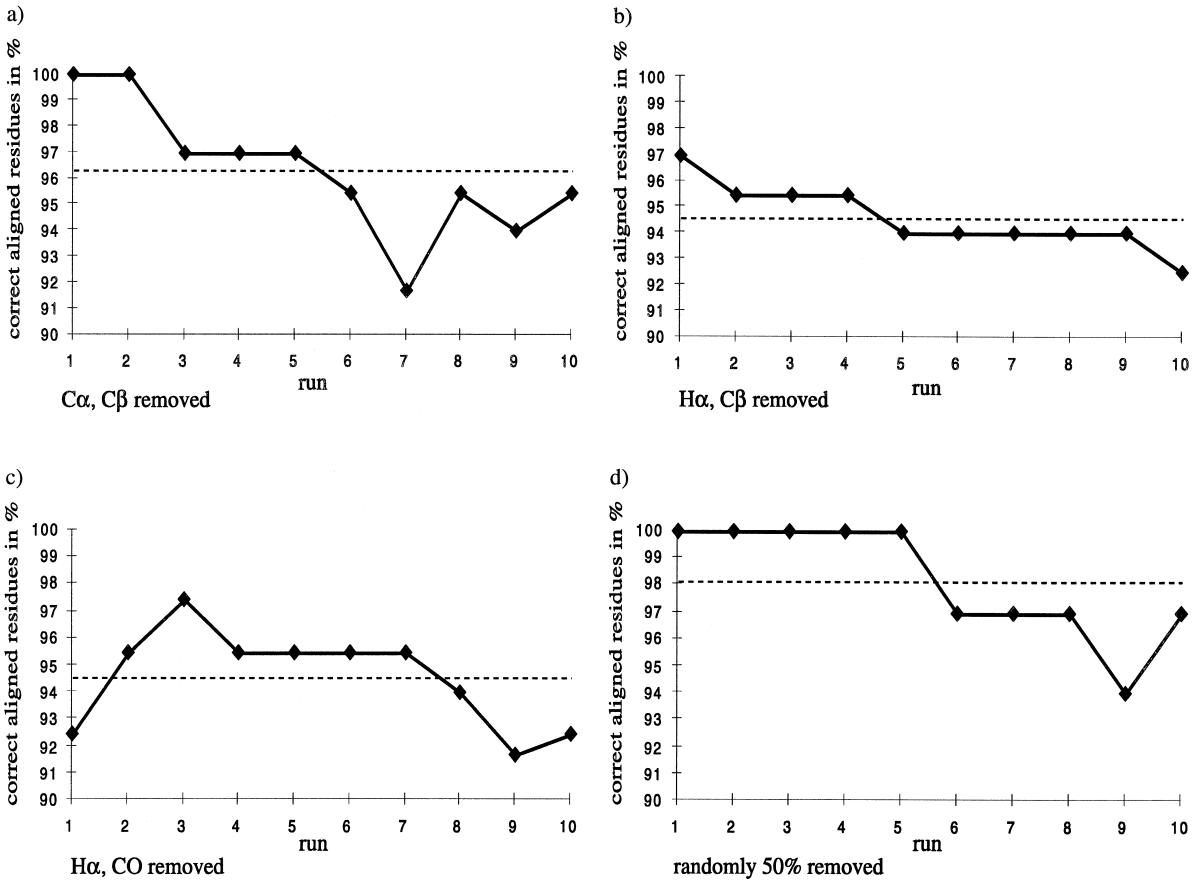

Figure 5. Percentage of correct assignment obtained in the different runs described in Figure 4. Even with significant signal overlap, the number of correctly aligned residues is on average more than 95% (broken lines).

to determine the global minimum without further information, since several equally valid solutions exist with respect to the input data. In all of the following assignment runs using artificial data, the proline residues were included. Thus, only the first element in the list has no sequential information, resulting in one or two fragments in all runs without any real error. This is defined as a 100% correct assignment.

The real errors are randomly distributed over the sequence and their number is not dependent on the amount of information. With regard to the complete data sets of all four test proteins, 55% of the runs do not contain a single real error and 45% contain one real error. In the case of interleukin 4, several sets of reduced information were tested additionally. These were created by omitting one type of sequential information, i.e., by removing all chemical shifts of either C^{α}, C^{β}, CO, or H^{α}, or by removing a certain percentage of data selected randomly from the chemical shifts of C^{α}, C^{β}, CO, and H^{α}. No major changes in the performance of the algorithm with regard to the number of real errors were observed (data not shown).

Using the complete artificial data sets and small tolerances, a 100% correct assignment is achieved for all four test proteins in at least three out of 10 runs (Figure 3). In the other runs, with one real error, three large fragments were produced (ca. 99% correct assignment, depending on the length of the different proteins).

After the complete removal of one of the four types of sequential information (C^{α}, C^{β}, CO, or H^{α}) from the interleukin 4 data set, no significant differences in the results were observed (data not shown).

The situation changes slightly if two types of sequential information are completely removed. While the number of real errors stays constant (see above), an increase in the number of fragments is observed, due to the higher ambiguity of the input data (Figure 4).

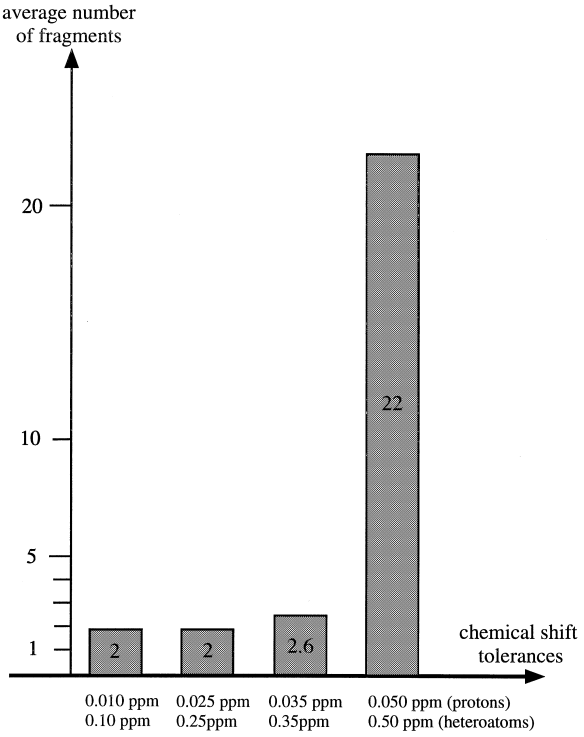

Figure 6. Average number of fragments in the assignment of interferon γ using different tolerance limits. It is obvious that only unrealistically large tolerance limits of 0.05 ppm for the H^{α} protons and 0.50 ppm for the carbon chemical shifts yield an unacceptable number of fragments.

Here, the increased overlap between the remaining sequential information creates a plateau in the solution space, i.e., up to more than 10 fragments without real error, which would cause a 'combinatorial explosion' in the straightforward approach. The average number of fragments produced depends on the type of the remaining sequential information. Under these circumstances, the best performance is given with 50% of the data randomly removed. Fragmentation due to informational degeneracy is an important fact when handling real data. The maximum number of fragments depends only on the specific chemical shifts and the tolerances used. However, this number is often not reached, because the correct successor is chosen accidentally for a residue at a fragmentation site. All possible fragmentation sites are outlined by PASTA and all alternative matching sequential residues are given, so that the correct sequential ordering can easily be found manually.

The differences in overall performance (Figure 5a–c) for the partial sets (C^{α}/C^{β}, C^{β}/H^{α} or CO/H^{α}, removed, respectively) can be correlated to the various degree of overlap (i.e., chemical shift dispersion) for the different types of remaining nuclei, especially in an α-helical protein like interleukin 4. The overall performance is on average 95% of the residues obtained in correct sequential order, in the worst case still 91.7%. When 50% of the sequential information is removed, the best results are achieved with the deletions occurring randomly. Here the program still finds the correct assignment in five out of 10 runs (Figure 5d).

The influence of changing the tolerances for the match between the shifts on the assignment is shown in Figure 6. With tolerances up to 0.035 ppm for the H^{α} protons and 0.35 ppm for the heteronuclei, the correct assignment is achieved nearly without fragmenting the resulting residue string. Only unreasonably high tolerances like 0.050 ppm for the H^{α} protons and 0.50 ppm for the heteronuclei lead to an increased number of fragments (22 fragments on average) in the output. Here again the ambiguity of the input data is too high to receive less fragmentation without additional information.

On the other hand, the algorithm is sensitive to wrong chemical shifts within a pseudo-residue, especially when small tolerances are used. These wrong shift values occur only when noise peaks are mistaken for actually missing signals in the peak-picking process. If the expected number of signals is found, then the wrongly picked peaks are not recognized by the program and give rise to real errors in the assignment, always occurring at the same pseudo-residue and resulting in increased fragmentation. Hence, such misassignments have to be carefully eliminated by manually counterchecking the spectra.

Application to hnps-PLA_2

The semiautomatic backbone assignment of hnps-PLA_2 (124 residues) was done independently from manual assignment. An initial set of 155 pseudo-residues, including noise and the side-chain amides, was obtained from the ^{15}N-HSQC. Since the sequence of hnps-PLA_2 contains two prolines (Pro^{36}, Pro^{122}) and the N-terminal residue is not easily observable, 121 pseudo-residues were expected. The pseudo-residue list was refined manually using the HNCA and HN(CO)CA spectra, i.e., only residues with one cross peak from the HN(CO)CA and at least one from the HNCA were kept. In addition, pseudo-residues containing side-chain amide signals were eliminated. This procedure led to a final list containing 114 residues. The further processing of the data is shown in Figure 7. Furthermore, no spec-

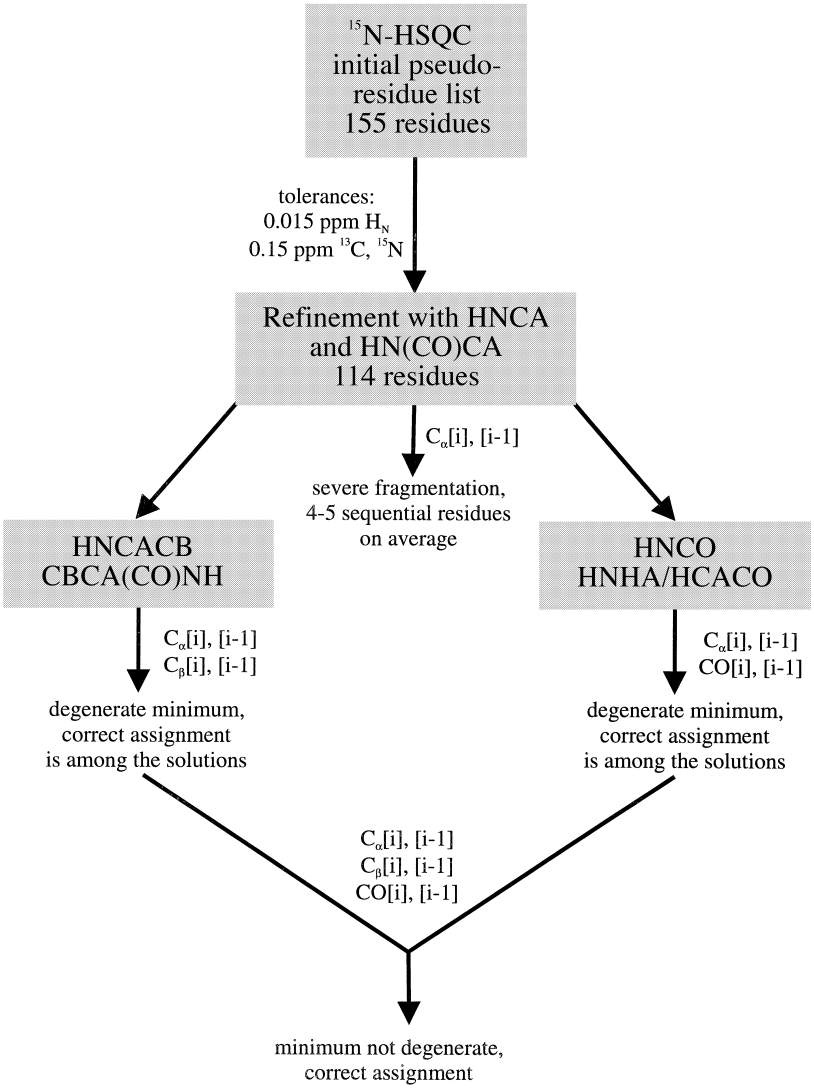

Figure 7. Diagram of the semiautomated assignment of hnps-PLA_2. The pseudo-residue list containing 114 residues was obtained from an initial ^{15}N-HSQC list by identifying noise and side-chain signals using HNCA and HN(CO)CA spectra. Using the C^α sequential information as a starting point, either the C^β or the CO data were added to get a sufficient amount of information to reach the correct assignment. However, an unequivocal assignment was only possible by using all three types of sequential information.

tral data were found for the residues Leu^2, Phe^{23}, Tyr^{24}, Cys^{77}, Tyr^{111}, Tyr^{112}, and Ser^{113} (Schwarz et al., 1997). Because the algorithm did not use any protein sequence information, at least the six fragments Val^3–Gly^{22}, Gly^{25}–Ser^{35}, Lys^{37}–Tyr^{76}, Ale^{78}–Gln^{110}, Asn^{114}–Thr^{121}, and Arg^{123}–Cys^{124} have to be contained in the solution. Therefore, an assignment containing exactly these six fragments has to be considered complete (corresponding to 100% in analogy to the definition given above).

Even with tolerances as small as 0.15 ppm, the C^α chemical shifts strongly overlap in this mainly α-helical protein (Figures 8a and b). With only the C^α sequential information available, fragments with an average length of 4–5 correctly ordered residues were obtained, resulting in a total number of about 25 fragments. Therefore, a second sequential information had to be supplied. The assignment was repeated alternatively with the combinations C^α/C^β and C^α/CO. Now the program reproduced assignments without any real error for both data sets. Although the correct as-

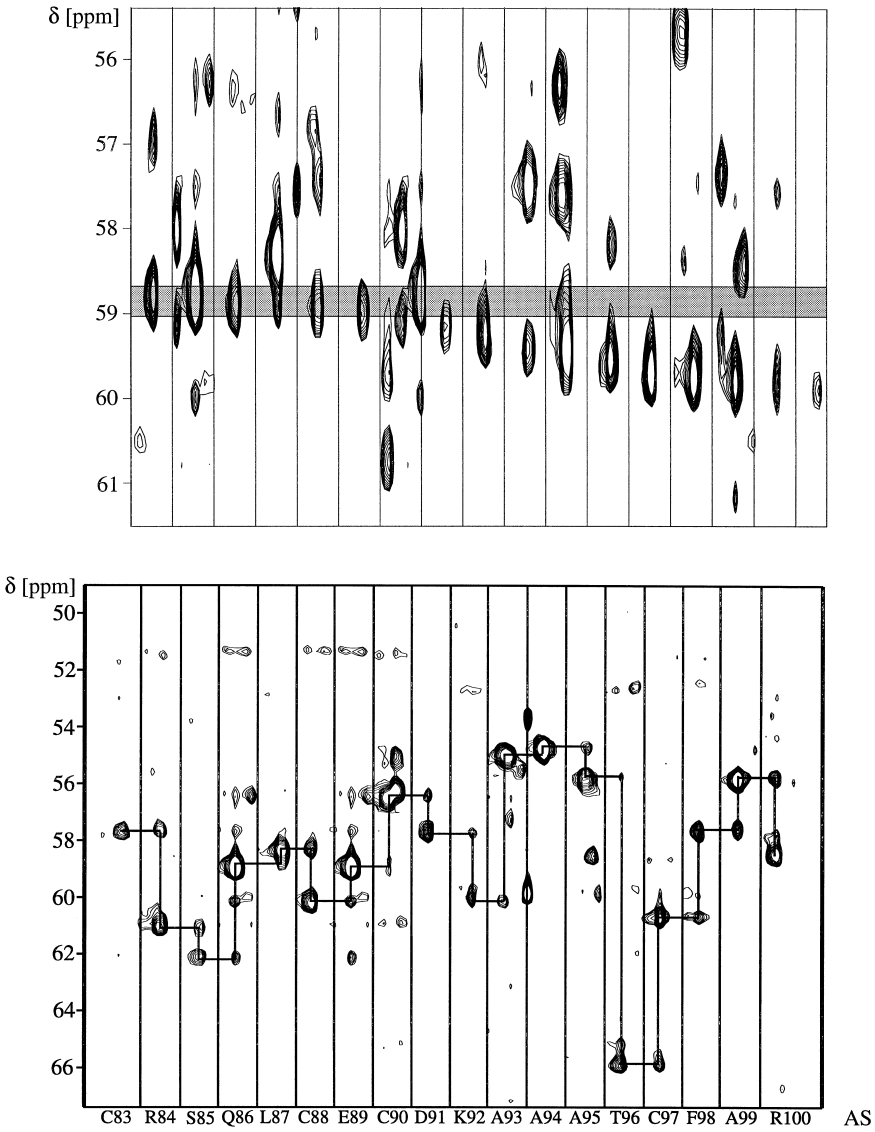

Figure 8. 600 MHz HNCA spectrum from hnps-PLA_2. (Top) The overlap problem is demonstrated by selected strips from the HNCA spectrum sorted by their C^{α} chemical shifts. The tolerance limit of 0.15 ppm is indicated by the shaded bar. It is evident that considerable overlap in the C^{α} region is still found in this well-resolved spectrum. (Bottom) Assignment walk from residues Cys^{83} to Arg^{100} according to the PASTA sequential alignment based on all three sequential informations.

signment is amongst the solutions obtained, it cannot be distinguished from other valid solutions produced by the algorithm due to the degeneracy of the minimum. This is also indicated by the fragmentation of the list for the runs without real errors, resulting in 9.2 fragments on average for the C^{α}/C^{β} combination and 11 fragments for the C^{α}/CO combination. In order to determine the correct solution, i.e., the global minimum, the CO shifts were included in further runs as additional independent constraints. In this way, it was possible for PASTA to reproduce the full assignment, which was independently obtained by manual evaluation (Figure 7).

Conclusions

It has been shown that PASTA reliably reproduces the correct assignment for the artificial data sets of interleukin 4, interferon γ, calmodulin, and FKBP without using the primary sequence or the proton side-chain

information. The algorithm is stable in its solutions; in all assignment runs it reaches solutions very close to the global minimum or even the minimum itself (100% correct assignment). The program is able to overcome up to 50% randomly missing signals and still aligns more than 90% of the residues correctly. Overlap of backbone information may lead to a fragmentation of the resulting data list. Nevertheless, when mapped onto the sequence, the correct assignment is easily obtained. The backbone resonances of hnps-PLA_2 were correctly assigned by PASTA, using automatically picked peak lists of several triple-resonance spectra. The results of the minimization based on the sequential information for C^{α}/C^{β} and C^{α}/CO both included the correct sequence among other possible solutions which are to be considered equally valid from the algorithm's point of view. Thus, additional information was necessary to enable the program to find the true global minimum. This can be accomplished when amino acid type determination (from C^{α} and C^{β} shifts) is used in combination with knowledge of the protein sequence. However, with *three* independent types of sequential information (C^{α}/C^{β}/CO), no knowledge of the primary structure was necessary to obtain the correct sequential assignment.

PASTA is a valuable tool in accelerating the tedious backbone assignment process using variable sets of backbone information. The core algorithm can be easily expanded to include new sequential information (such as proton spin systems) for a further reduction of the search space for the sequential assignment, in order to enhance the automatic assignment capabilities of the program.

Acknowledgements

Financial support of the Deutsche Forschungsgemeinschaft, the Sonderforschungsbereich 369, the Fonds der Chemischen Industrie and the Dr.-Ing. Leonhard Lorenz-Stiftung is acknowledged.

Note

A prototype of the program was presented at the 12th European Experimental NMR Conference in Oulu, Finland, 1994 (poster W118). In this early phase of the project, with only the minimization strategy implemented in the actual form, the program assigned IIB^{Glc} (Eberstadt et al., 1996), a small protein with 94 residues, based on the peak lists of HNCA/CBCA(CO)NH and HN(CA)CO/HNCO.

The ab initio assignment of two proteins, IIB^{Man} (Gschwind et al., 1997) with 168 residues and nusB (Berglechner et al., 1997) comprising 139 residues, was done by PASTA in our group, following a strategy similar to the assignment of hnps-PLA_2. Work is in progress to include side-chain and NOE information in order to fully automate the assignment process.

References

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

Bartels, C., Billeter, M., Güntert, P. and Wüthrich, K. (1996) *J. Biomol. NMR*, **7**, 207–213.

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

Berglechner, F., Richter, G., Fischer, M., Bacher, A., Gschwind, R.M., Huenges, M., Gemmecker, G. and Kessler, H. (1997) *Eur. J. Biochem.*, submitted.

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

Dueck, G. and Scheuer, T. (1990) *J. Comput Phys.*, **90**, 161–175.

Dueck, G., Scheuer, T. and Wallmeier, T. (1993) *Spektrum Wissenschaft*, **3**, 42–51.

Eberstadt, M., Golic-Grdadolnic, S., Gemmecker, G., Kessler, H., Buhr, A. and Erni, B. (1996) *Biochemistry*, **35**, 11286–11292.

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

Grzesiek, S. and Bax, A. (1992) *J. Magn. Reson.*, **96**, 432–440.

Grzesiek, S., Döbelli, H., Gentz, R., Garotta, G., Labhardt, A.M. and Bax, A. (1992) *Biochemistry*, **31**, 8180–8190.

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

Gschwind, R.M., Gemmecker, G., Leutner, M., Kessler, H., Gutknecht, R., Lanz, R., Flükiger, K. and Erni, B. (1997) *FEBS Lett.*, **404**, 45–50.

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

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

Kirkpatrick, S., Gelatt, C.D. and Vecchi, M.P. (1983) *Science*, **220**, 671–680.

Kramer, R.M., Hession, C., Johanson, B., Hayes, G., McGray, P., Chow, E.-P., Tizard, R. and Pepinsky, R.B. (1989) *J. Biol. Chem.*, **264**, 5768–5775.

Lawler, E.L., Lenstra, J.K., Rinnooy, G., Kan, A.H. and Shmoys, D.B. (1985) *The Traveling Salesman Problem. A Guided Tour of Combinatorial Optimization*, Wiley/Interscience, New York, NY.

Logan, T.M., Olejniczak, E.T., Xu, R.X. and Fesik, S.W. (1992) *FEBS Lett.*, **314**, 413–418.

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

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

Montelione, G.T., Lyons, B.A., Emerson, S.D. and Tashiro, M. (1992) *J. Am. Chem. Soc.*, **114**, 10974–10975.

Morales, L.B., Garduno-Juarez, R. and Romero, D. (1992) *J. Biomol. Struct. Dyn.*, **9**, 951–957.

Morelle, N., Brutscher, B., Simorre, J.-P. and Marion, D. (1995) *J. Biomol. NMR*, **5**, 154–160.
Neidig, P., Geyer, M., Görler, A., Antz, C., Saffrich, R., Beneicke, W. and Kalbitzer, H.R. (1995) *J. Biomol. NMR*, **6**, 255–270.
Olson Jr., J.B. and Markley, J.L. (1994) *J. Biomol. NMR*, **4**, 385–410.
Powers, R., Garret, D.S., March, C.J., Frieden, E.A., Gronenborn, A.M. and Clore, M.G. (1992) *Biochemistry*, **31**, 4334–4356.
Press, W.H., Teukolsky, S.A., Vetterling, W.T. and Flannery, B.P. (1992) *Numerical Recipes in C*, 2nd ed., Cambridge University Press, Cambridge, p. 283.
Sattler, M. and Fesik, S.W. (1996) *Structure*, **4**, 1245–1249.
Schwarz, C., Schmidt, J., Diercks, T., Planker, E., Leutner, M., Kelly, M. and Kessler, H. (1997), manuscript in preparation.
Seilhamer, J.J., Pruzanski, W., Vadas, P., Miller, J.A., Kloss, J. and Johnson, L.K. (1989) *J. Biol. Chem.*, **264**, 5335–5338.
Shan, Xi., Gardner, K.H., Muhandiram, D.R., Rao, N.S., Arrowsmith, C.H. and Kay, L.E. (1996) *J. Am. Chem. Soc*., **118**, 6570–6579.
Wishart, D.S., Bigham, C.G., Holm, A., Hodges, R.S. and Sykes, B.D. (1995) *J. Biomol. NMR*, **5**, 67–81.
Wittekind, M.G. and Mueller, L. (1993) *J. Magn. Reson.*, **B101**, 201–205.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York, NY.
Xu, J., Strauss, S.K., Sanctuary, B.C. and Trimble, L. (1994) *J. Magn. Reson.*, **B103**, 53–58.
Xu, R.X., Nettesheim, D., Olejniczak, E.T., Meadows, R., Gemmecker, G. and Fesik, S.W. (1993) *Biopolymers*, **33**, 535–550.
Yamazaki, T., Lee, W., Arrowsmith, C.H., Muhandiram, D.R. and Kay, L.E. (1994) *J. Am. Chem. Soc.*, **116**, 11655–11666.
Zimmerman, D. and Montelione, G.T. (1995) *Curr. Opin. Struct. Biol.*, **5**, 664–673.