Journal of Biomolecular NMR, **26:** 93–111, 2003.
KLUWER/ESCOM

PACES: Protein sequential assignment by computer-assisted exhaustive search

Brian E. Coggins & Pei Zhou*
Department of Biochemistry, Duke University Medical Center, Durham, NC 27710, U.S.A.

Received 13 November 2002; Accepted 21 February 2003

Key words: complexity of sequential assignment, complexity index of sequential assignment, computer-assisted sequential assignment, sequence-specific resonance assignment, sequential assignment

Abstract

A crucial step in determining solution structures of proteins using nuclear magnetic resonance (NMR) spectroscopy is the process of sequential assignment, which correlates backbone resonances to corresponding residues in the primary sequence of a protein, today, typically using data from triple-resonance NMR experiments. Although the development of automated approaches for sequential assignment has greatly facilitated this process, the performance of these programs is usually less satisfactory for large proteins, especially in the cases of missing connectivity or severe chemical shift degeneracy. Here, we report the development of a novel computer-assisted method for sequential assignment, using an algorithm that conducts an exhaustive search of all spin systems both for establishing sequential connectivities and then for assignment. By running the program iteratively with user intervention after each cycle, ambiguities in the assignments can be eliminated efficiently and backbone resonances can be assigned rapidly. The efficiency and robustness of this approach have been tested with 27 proteins of sizes varying from 76 amino acids to 723 amino acids, and with data of varying qualities, using experimental data for three proteins, and published assignments modified with simulated noise for the other 24. The complexity of sequential assignment with regard to the size of the protein, the completeness of NMR data sets, and the uncertainty in resonance positions has been examined.

Introduction

The successful study of protein structure or dynamics by nuclear magnetic resonance (NMR) spectroscopy often begins with the sequential assignment of backbone resonances. Although sequential assignment can be completed for small proteins with relative ease, it can be a time-consuming process for large proteins. The sequential assignment process has been greatly facilitated by the development of automated or computer-assisted methods, some of which have been summarized in a recent review by Moseley and Montelione (1999). Many of these algorithms assemble connectivity fragments in a deterministic manner based on inter-residue connectivity information, and then determine amino acid types via carbon chemical shifts (Zimmerman et al., 1997; Li and Sanctuary, 1997a,b; Atreya et al., 2000); others utilize simulated annealing methodologies (Buchler et al., 1997; Leutner et al., 1998; Bartels et al., 1997), a combination of the two methods (Lukin et al., 1997), or a methodology based on comparison to known assignments for homologous proteins (Gronwald et al., 1998). The MAPPER program also exists, which maps user-assembled connectivity fragments to the protein sequence using an exhaustive search and a scoring system derived from a probability model (Güntert et al., 2000). Recently, assignment based on residual dipolar coupling has also been proposed (Tian et al., 2001).

Several problems can arise during the sequential assignment process that may present difficulties for automated assignment algorithms, and these programs differ in the methods they use to address these issues.

*To whom correspondence should be addressed. E-mail: peizhou@biochem.duke.edu

Degeneracies in the resonance data present one of the most challenging problems, requiring automated programs to choose among multiple connectivity possibilities. In addition, successful algorithms must be able to address missing peaks, extra peaks from impurity or multiple conformers, uncertainty in the amino acid type determinations, and artificially broken connectivities due to the uncertainty of resonance positions caused by the distortion or overlap of peaks.

Many algorithms described to date utilize some form of 'best-first' reasoning to resolve ambiguities in the data. Best-first approaches establish criteria to select the best candidate from several potential connectivities. Unfortunately, wrong choices made early in the assignment process can lead to solutions which can be substantially incorrect (Zimmerman et al., 1997), and, as a consequence, many of these algorithms break down when faced with highly degenerate or incomplete data sets. Although simulated annealing approaches do not make decisions based on best-first logic, there is a drawback in that their energy functions can be trapped at local minima instead of reaching a global minimum.

Because all of these problems become more prevalent as the size of the protein under study increases, automated assignment approaches have faced difficulties with larger proteins. Although automated assignment has become practical for proteins with less than 100–200 amino acids, the automated assignment of proteins with more than 200 amino acids has been less successful, except in a few particular cases (Lukin et al., 1997; Moseley and Montelione, 1999; Atreya et al., 2000).

Theoretically one should be able to exhaustively enumerate all of the possible assignment solutions for a given protein, and the correct assignment would always be embedded among these solutions. Thus one is left only with the task of eliminating improbable outcomes, until only one remains. Although exhaustive enumeration of all possible solutions has been deemed exceedingly expensive in computational terms due to the large number of potential outcomes, we have found that for a typical protein, the intrinsic constraints imposed by the connectivity requirements derived from triple-resonance data reduce the number of possible solutions significantly, making such an approach practical. We have also found that this approach produces unambiguous 'consensus' assignments for the majority of residues in a protein. By using an iterative method of determining consensus assignments first and eliminating solutions that are not logically consistent with those results, it is possible to resolve almost all assignment ambiguities rapidly. Alternatively, because an exhaustive search develops multiple assignment solutions in parallel, one can apply this approach to determine assignments for multiple conformers of a protein present in solution at the same time.

Here we report the development of PACES, an interactive program for **P**rotein Sequential **A**ssignment by **C**omputer-Assisted **E**xhaustive **S**earch. PACES establishes sequential assignments based on the sequential connectivity and residue type information derived from C^{α}, C^{β}, carbonyl and H^{α} triple-resonance data, or any subset of these data. Additional information derived from other experiments, such as residue types or NOESY constraints, can also be introduced as PACES input. The efficiency and robustness of this program have been tested against data for 27 proteins ranging from 10 to 80 kDa in molecular weight – 76 to 723 residues in length – using either experimental data or data generated using assignment chemical shifts from BioMagResBank (BMRB) entries modified with simulated experimental error. We demonstrate that this approach yields rapid, accurate and complete assignments for proteins with high quality data, and accurate partial assignments for proteins with less complete data. We further explore various factors, such as the size of the protein, the completeness of data, and the uncertainty of peak position, that affect the complexity of sequential assignment.

Theory and methods

Input data

The exhaustive search algorithm accepts the intra- and inter-residue (*i* and *i*-1) chemical shifts of the alpha, beta and carbonyl carbons, as well as the alpha proton(s), for each spin system, or any subsets of those four pairs of data sets that are available through triple-resonance experiments (Olejniczak et al., 1992; Yamazaki et al., 1994a,b). Since the possible amino acid types of individual spin systems are normally derived from the chemical shift information for the C^{α}, C^{β} and carbonyl nuclei, it is important that these chemical shifts be properly referenced (Markley et al., 1998). PACES anchors the resonances of a spin system to its HSQC peak, and subsequently refers to spin systems by HSQC peak numbers.

For peak lists created using the XEASY program (Bartels et al., 1995), the assembly of spin systems

from spectral data may be accomplished in a semi-automatic manner. PACES imports the XEASY HSQC peak list to provide spin system anchors, and uses peak lists from triple-resonance experiments to fill in additional resonances; peaks are added to spin systems when their HN and N chemical shifts match a given spin system within a user-defined tolerance. In cases of ambiguity or amide degeneracy, PACES provides on-screen dialog boxes allowing the user to determine how the ambiguity is resolved.

In cases of severe amide degeneracy, in which it is not possible to separate the peaks belonging to two different spin systems, multiple spin systems should be provided to PACES comprising the full set of possible combinations of peaks. This method will ensure that the correct combinations of peaks for a particular spin system is always generated.

Additional information, such as residue types or sequential connectivity derived from NOE crosspeaks (Wüthrich, 1986), may also be provided as additional restraints. PACES accepts possible amino acid types derived from a great variety of techniques, including specific isotopic labeling by residue type (LeMaster and Richards, 1985) or residue-type-specific experiments based on side-chain topology (Schubert et al., 1999, 2001a,b), or based on the number of coupling partners of C^{β} nuclei (Dötsch et al., 1996a–c; Dötsch and Wagner, 1996). Residue type information can also be derived from sidechain experiments such as HCC(CO)NH-TOCSY (Tashiro et al., 1995) or CC(CO)NH-TOCSY (Farmer and Venters, 1996). Restrictions on connectivity, in the form of constraints specifying how close together two spin systems are on the protein primary sequence, may be derived from NOESY spectra and provided to the PACES program.

Assembly of connectivity fragments

Spin systems are connected on the basis of matching chemical shifts. The intra-residue resonances of every spin system are compared to the inter-residue resonances of all other spin systems to build a table of dipeptide connectivities. For two spin systems j and k, with experimentally measured chemical shifts

$$\begin{aligned} j &= \left\{ C^{\alpha}_{j}, C^{\beta}_{j}, C'_{j}, H^{\alpha}_{j}, C^{\alpha}_{j-1}, C^{\beta}_{j-1}, C'_{j-1}, H^{\alpha}_{j-1} \right\}, \\ k &= \left\{ C^{\alpha}_{k}, C^{\beta}_{k}, C'_{k}, H^{\alpha}_{k}, C^{\alpha}_{k-1}, C^{\beta}_{k-1}, C'_{k-1}, H^{\alpha}_{k-1} \right\}, \end{aligned} \tag{1}$$

a dipeptide segment jk will be established if

$$\begin{aligned} \left| C^{\alpha}_{j} - C^{\alpha}_{k-1} \right| &\leq \delta_{C\alpha}, \\ \left| C^{\beta}_{j} - C^{\beta}_{k-1} \right| &\leq \delta_{C\beta}, \\ \left| C'_{j} - C'_{k-1} \right| &\leq \delta_{C'}, \\ \left| H^{\alpha}_{j} - H^{\alpha}_{k-1} \right| &\leq \delta_{H\alpha}, \end{aligned} \tag{2}$$

where $\delta_{C\alpha}$, $\delta_{C\beta}$, $\delta_{C'}$ and $\delta_{H\alpha}$ represent user-defined chemical shift thresholds for C^{α}, C^{β}, carbonyl and H^{α}, respectively. The check is disabled for any type of data (such as H^{α}) that is not available. For spin systems with missing resonances, the following rules are used to establish a dipeptide fragment jk:

(1) If only one type of chemical shifts is available (such as C^{α}), the intra-residue resonance of spin system j must match the inter-residue resonance of spin system k to establish the connectivity jk.
(2) If two or more types are available, the connectivity of jk is established if there are at least two complete and matching sets of resonances.

Spin systems with missing data that do not meet the criteria listed above are not used during the assembly process, but can be used to extend existing fragments using interactive tools. The complete list of possible dipeptide connectivities is stored as a raw connectivity table.

Larger fragments are then assembled based on this table. If the connectivity relationships present in the data are viewed as a directional network (e.g., Figure 1a), the process is to pick an arbitrary starting point in the network and trace out all possible downstream paths. Upstream nodes are added into the network as they are encountered. Each of these complete paths through the network is recorded as a connectivity fragment to be aligned with the protein sequence later.

The assembly process starts with the first spin system in the table, and its C-terminal connectivity fragments are traced out by following the linkages listed in the dipeptide table. Each of these downstream nodes is flagged as having been processed to prevent duplication. The algorithm then moves to the next spin system in the table that has not yet been flagged, and traces out all of its C-terminal paths. If one of these C-terminal paths connects onto the N-terminus of a spin system that has already been processed, the new fragment is merged with the existing one, and the new possible pathways are traced out.

If chemical shift degeneracy is present, several scenarios may be encountered (Figure 1b). When multiple

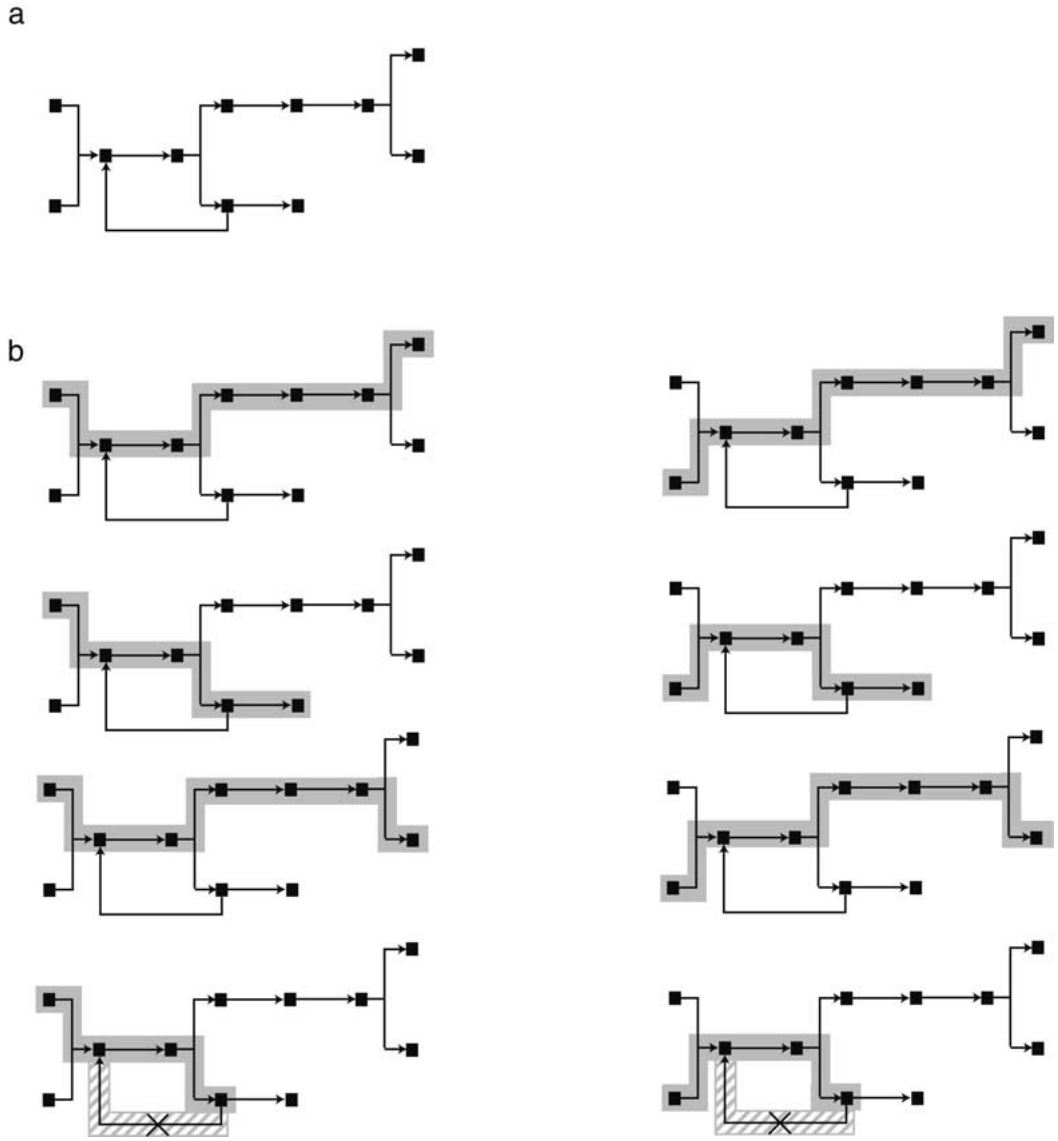

Figure 1. Assembly of connectivity fragments. (a) An example of a directional network representing interconnections between spin systems. Each square represents a spin system, and each arrow shows a sequential linkage that is possible based on the pairing of interresidue chemical shifts. Only one path through this network would be the correct path corresponding to a stretch of sequence in the protein. (b) All of the possible connectivity fragments in this network, each traced with gray shading. The bottom row shows how circular references are addressed in PACES; the connections shaded in solid gray are generated as fragments, while the hatched gray sections are not pursued.

C-terminal connectivities are present for a particular spin system, each of those connectivities is traced out in turn, generating multiple unique fragments with different C-terminal endings. When multiple spin systems show C-terminal connectivity converging to the same spin system, the different permutations are traced out for each variant N-terminal ending. If a circular reference is encountered – that is, if an upstream spin system in a fragment is also identified as a downstream branch – the algorithm generates a fragment that stops at the branch point (Figure 1b, bottom row).

The consequence of this exhaustive search and enumeration is the generation of spin system fragments containing all of the potential connectivities.

Identification of amino acid types

PACES determines the possible residue type information for each spin system by comparing its C^{α}, C^{β}

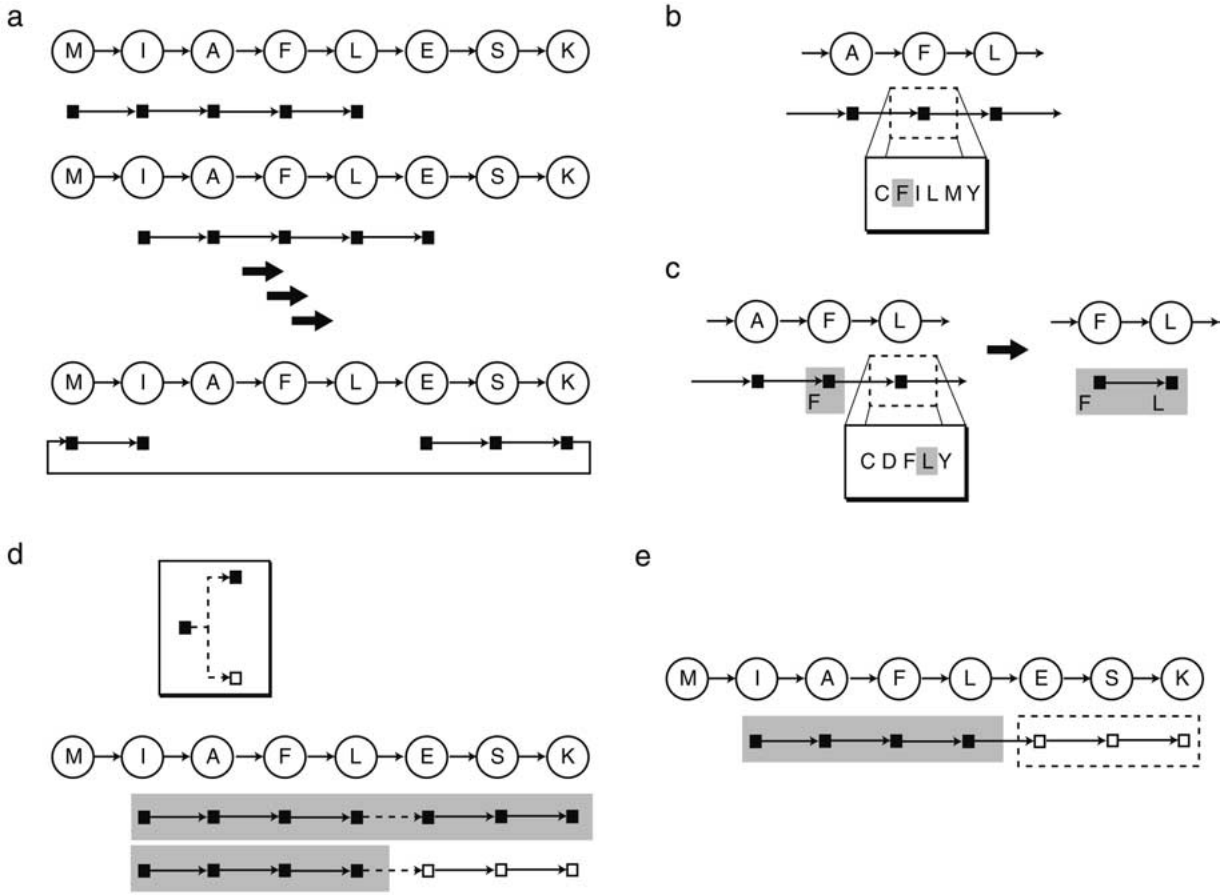

Figure 2. Mapping of fragments to the protein sequence. (a) The mapping process involves moving each fragment down the protein sequence, and testing for matching at each position. The string of circles represents a protein sequence, and the string of squares represents a fragment of connected spin systems. (b) A spin system is considered to match a residue on the protein sequence if the amino acid type of that residue is one of the possible amino acid types for that spin system. (c) A matching segment, indicated with gray shading, is identified when two or more consecutive spin systems match the protein sequence. (d) If degeneracy is present, it is possible for multiple fragments to be generated, which match to the same position on the protein sequence. In this example, the correct matches for the stretch of sequence are represented as filled squares, and incorrect matches as open squares. (e) Portions of fragments that do not match the protein sequence, such as the portions shown inside the dashed box, are recycled to the fragment pool for assignment elsewhere.

and carbonyl chemical shifts to the statistical chemical shift distribution of each amino acid type derived from the BioMagResBank. The chemical shift ranges used by PACES for determining amino acid types are provided in the Supplementary Information. All of the possible amino acid types for each spin system are recorded; the algorithm does not weight the probabilities or choose between them. If the protein under study is perdeuterated, PACES can be directed to adjust its amino acid ranges accordingly (Venters et al., 1996a; Farmer and Venters, 1999).

Mapping of fragments to the protein sequence

The mapping process (Figure 2) involves aligning a connectivity fragment at the beginning of the protein sequence, and then moving it down the sequence, one residue at a time. When the C-terminus of the protein is encountered, spin systems are looped back to align with the N-terminus of the protein, until every spin system in the fragment has been examined at every position on the protein sequence (Figure 2a). If at any position, the list of possible amino acid types for a spin system includes the amino acid type of the paired residue in the sequence, that pairing is considered a match (Figure 2b). If multiple sequentially-connected spin-systems match, the algorithm identifies the result-

ing segment as a matching fragment (Figure 2c). For the first spin system in a matching fragment, the inter-residue $(i-1)$ chemical shifts are also checked against the amino-acid type of the residue on the protein sequence that precedes the matched fragment.

Ideally, a 'correct fragment' should match the protein sequence completely if it is aligned at the right position. It is possible, though, for chemical shift degeneracy to cause fragments other than the correct one to be generated, possessing only a certain consensus portion of the correct fragment, but with different N-terminal or C-terminal overhangs joined at the point of degeneracy (Figure 2d). When aligned against the protein sequence, the 'correct' connectivity usually stands out as the longest contiguous matching fragment (Figure 3). Nonmatching portions of fragments are recycled into the fragment pool for independent consideration at other positions (Figure 2e). It is important to recycle the non-matching portions of the fragments because in many cases these segments are real and should be assigned at other positions on the protein sequence. If a matching fragment is found to be a subset or superset of a previously identified fragment at the same position on the protein sequence, only the longer one will be retained.

Because proline residues are not directly observed as spin-systems in triple-resonance experiments (Olejniczak et al., 1992; Yamazaki et al., 1994a,b), they must be treated differently. During the mapping process, spin systems are not allowed to match at prolines, meaning that matching fragments must be broken on either side of such a residue.

Filtering of matching fragments

Although the procedure described above is both 'inclusive' and 'exhaustive' – that is to say, all possible combinations of connectivities are considered and the correct solution is always embedded within these possible fragments – this process may generate such quantities of potential alignments that further examination by the user would be prohibitive. We have thus imposed a mask that filters out fragments with high uncertainties before presentation. Short matching fragments (mostly di- and tripeptides) are filtered out unless they contain a spin system with less than five possible amino acid types. This filter may be disabled by the user, ordinarily at the latter stages of assignment, to allow the remaining short fragments to be assigned.

Incorporation of user constraints

Additional constraints derived from a variety of sources can be utilized during the procedure described above to reduce the number of possible solutions that must be considered. Residue-type information obtained from selective labeling experiments (LeMaster and Richards, 1985), side-chain assignment data or from amino-acid-type-specific NMR experiments (Tashiro et al., 1995; Farmer and Venters, 1996; Dötsch et al., 1996a–c; Dötsch and Wagner, 1996; Schubert et al., 1999; Schubert et al., 2001a,b) can be used by the program directly, and are supplied as a list of possible types. These constraints will override residue-type determinations based on chemical shifts. In most cases, sequential or medium range connectivity can also be established from NOESY data (Wüthrich, 1986), complementing the connectivity information available from scalar couplings. These spatially defined sequential and medium range connectivities can be used to eliminate improperly assembled connectivity fragments.

Implementation

We have implemented this algorithm and provided appropriate tools for analyzing results in the PACES package for Microsoft Windows. Core processing occurs in a module written in Microsoft Visual Basic 6.0 and encapsulated in a Component Object Model object class, which was compiled as a Win32 executable for x86 systems running Windows NT/Windows 95 or later. The user interface was developed in Microsoft Visual Basic 6.0 as an add-in module to Microsoft Excel 2000 and later.

Chemical shift data and additional constraints are provided by the user on a properly-formatted Excel worksheet, in a text file or from XEASY peak lists. The program is operated from a drop-down menu added to Excel automatically when a PACES-compatible data file is open. Connectivity thresholds, the protein sequence, and flags governing program operation – such as one for use with perdeuterated proteins, directing that the chemical shift ranges be adjusted to account for deuterium isotope effects – are selected through a dialog box. Results are presented on a formatted Excel worksheet, with one column for each residue (Figure 3). Each matching fragment is presented as a horizontal row, with the spin system numbers for each possible assignment positioned in the appropriate column for the corresponding residue. The matching fragments are sorted first by starting

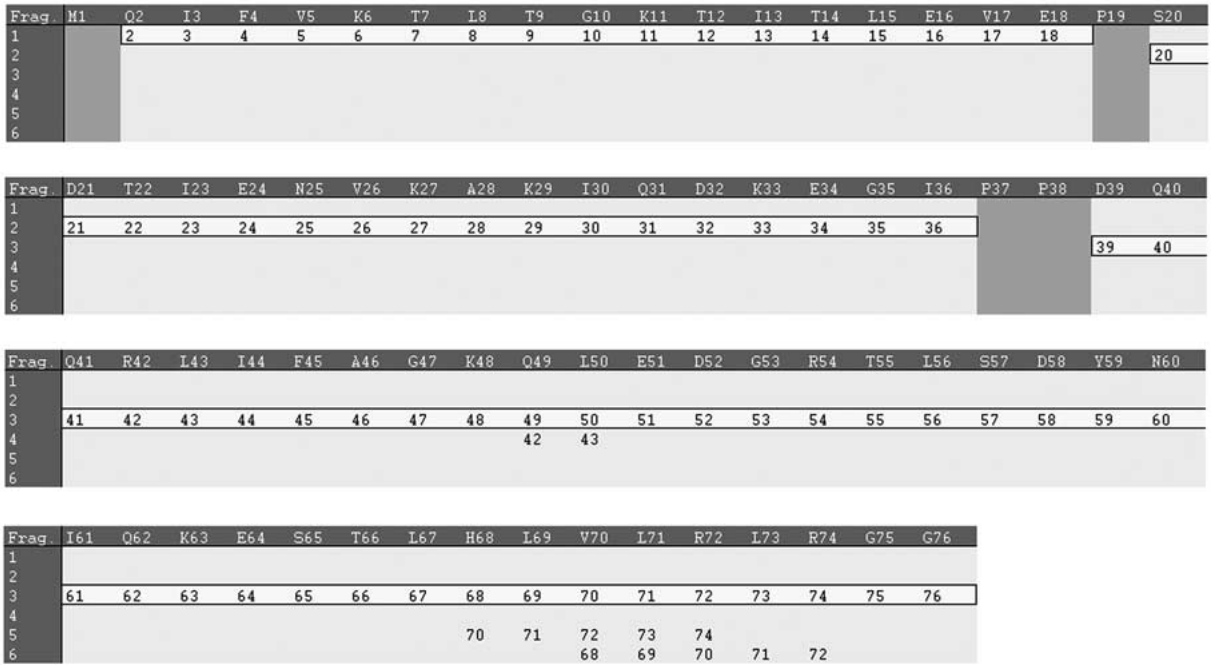

Figure 3. PACES output for human ubiquitin. The ubiquitin sequence is shown across the top, and matching fragments of spin systems are shown as rows underneath. Each number in a row is the reference number for that spin system; here, for clarity of display, spin systems have been labeled with the residue number they were assigned to in the published assignments. The most likely assignment for each position is boxed, as determined by the PACES program based on fragment length.

position on the amino acid sequence, from N- to C-terminus, and then by length for each fragment, from longest to shortest. The longest fragment for each region of the sequence is highlighted in color. If two long fragments overlap at an end, or if two long fragments are identical except for conflicting possibilities at particular residues, the conflicting residues are either not highlighted or highlighted in a cross-hatched pattern at the user's option. User-assigned spin systems are colored in red.

PACES is able to make use of all available memory during analysis, but typically requires only three megabytes for overhead, and five to ten megabytes while processing. If, on account of excessive degeneracy, PACES encounters an abnormally large number of fragments, it will cease processing at a programmed limit. This limit is set to 10 000 by default, which prevents PACES from running for more than about ten minutes, and limits memory consumption. The limit may be changed or removed at the user's option, however, to allow PACES to analyze complex situations with large numbers of fragments. Increasing the limit an order of magnitude may require as much as 200 megabytes of memory for processing, depending upon the lengths of the fragments PACES is generating. In any event, the program will automatically stop execution when memory is exhausted, without crashing the system.

Interactive tools are provided for analyzing PACES results, and for assigning spin systems (Figure 4). Context-sensitive popup menus enable the user to access information about or issue commands regarding a particular spin system (Figures 4a and 4b), enabling one to access quickly that spin system's chemical shifts and possible amino acid types, the other positions on the protein's primary sequence at which it is suggested for assignment, and the user constraints supplied for it, and enabling one to assign it at that position, or to add other constraints. Fragment extension tools (Figures 4c and 4d) allow users to extend or join existing fragments. These tools enable one to locate spin systems that have been eliminated during the assembly process because (1) their chemical shifts are outside the connectivity thresholds, due to peak distortion, overlap, or uncertainty in the peak position; (2) their chemical shifts are outside the statistical range for the corresponding amino acid, as would occur with residues that bind metals, for example; or (3) insufficient data are present to establish connectivity, as in the case of missing inter- or intra-residue

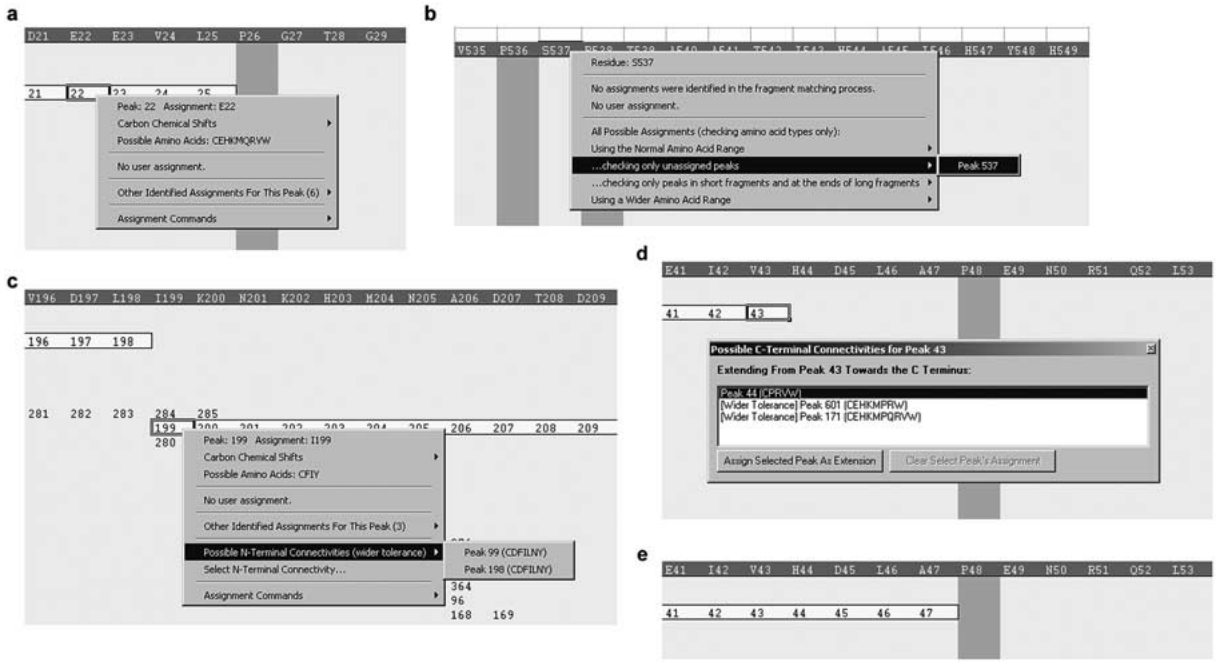

Figure 4. Screen shots of PACES tools for examining assignment possibilities. In all examples here, spin systems are labeled by their published assignments for clarity of display. (a) Clicking on a spin system label with the right mouse button prompts the program to display a popup window with information about that spin system. (b) Assigning an isolated residue between two prolines. This example was taken from malate synthase G after all assignable residues elsewhere in the protein had been assigned. (c) Connecting two spin systems with connectivity outside the set thresholds. In this example from maltose binding protein with simulated error, the error in the connectivity data between spin systems 198 and 199 is greater than the threshold for establishing connectivity. The PACES display enables the user to determine that the two would be connected using a wider threshold. (d) Residue H44 of malate synthase G has a C^{α} chemical shift which is ~0.5 ppm higher than the PACES program's normal range for histidine, and as a consequence, its assignment at that position was not suggested. Using the PACES fragment extension tool it is possible to determine that the two show connectivity. (e) If the PACES program is subsequently run with extended amino acid ranges, spin system 44 and those that follow are suggested for assignment.

resonances. Popup information is also available at the residue headings, often enabling one to assign individual amino acids or small fragments between prolines once the remainder of the protein has been assigned (Figure 4b).

To facilitate the checking of multiple assignment possibilities against the original NMR spectra, PACES can interface directly with spectrum analysis programs, such as XEASY (Bartels et al., 1995). PACES reads in XEASY sequence and atom list files and writes out peak lists and strip files with assignments. PACES can also export completed assignments to the TALOS program for creating torsion angle restraints (Cornilescu et al., 1999).

The PACES package is available upon request from the authors.

Assignment procedure

The procedure for interpreting PACES results is derived from the fact that the longest matching segments identified by our algorithm have the greatest certainty of being the correct assignments. Once these spin systems have been assigned, the PACES program will exclude solutions that are not logically consistent with these assignments during future runs. This reduces the number of possible solutions at other positions on the protein sequence and enables the user to make assignments by an iterative process of elimination.

For data sets that are relatively complete (i.e. that contain data for more than 80% of residues), fragments of seven spin systems or longer can usually be assigned immediately in their entirety, except for portions that conflict with other long fragments. Sometimes several long fragments are produced that cover the same residues and are identical in content except for one or two residues – usually this variation occurs at an end, but sometimes it appears in the interior. In these cases all consensus residues may be assigned except for those that conflict. Occasionally a situation is encountered in which a region of the protein has

several short fragments that are adjacent to each other on the sequence, but have not been automatically connected by PACES. These may be examined using the popup displays to determine whether or not they can be connected using a wider connectivity tolerance, or to conclude that they have not been connected simply due to insufficient data. If several of these short fragments can be connected, they may be treated as one long fragment and subsequently assigned.

The PACES output for most parts of the protein sequence will normally be simplified considerably after the second run, as solutions that are inconsistent with the initial assignments will have been eliminated. The fragment extension tools may then be used to extend long fragments. Independent fragments that match a stretch of four or more amino acids may be assigned if the solution is unique. The program may be run additional times after making additional assignments to allow PACES to remove suggested solutions that are no longer logically consistent with the established assignments. It may be useful to run the PACES program at this point with extended amino acid ranges (2 ppm wider than the default ranges), to find any connectivities that have been eliminated due to residues with slightly atypical chemical shifts (Figures 4d and 4e).

When most of the protein has been assigned, the filter mask can be removed to display fragments as small as dipeptides, and the residue-heading popup displays can be used to locate possible assignments for isolated positions, such as a single residue between two prolines (Figure 4b). In most cases, almost the entirety of the protein will eventually be assigned in this manner.

Occasionally, there will still exist a few irresolvable instances of spin systems that can not be assigned because they appear at multiple positions. These issues must be resolved by reference to the original spectra, by using lineshape analysis, for example, or by using other types of information, such as NOE crosspeaks that establish short-range sequential connectivity (Wüthrich, 1986).

With less complete data, using a more conservative assignment procedure is advisable. Nonconflicting interior portions of fragments with more than ten connected spin systems may be assigned when no alternatives are presented in other long fragments. These fragments may be extended in subsequent runs using the fragment extension tools. Shorter fragments should be considered as suggestions only and might not be assignable with certainty.

If it is not possible, due to excessive degeneracy, to analyze a particular data set at the normal chemical shift thresholds suggested by the spectral resolution, it may be possible to assign the protein using reduced chemical shift thresholds during the initial PACES run. To do this, thresholds are reduced until the program is able to run within the limit of the computer's available memory. The results are then analyzed using a conservative procedure, assigning only nonconflicting interior portions of fragments with ten or more spin systems. In subsequent runs, the standard thresholds are restored, and assignments can be made following the previously described procedure.

Testing procedures

The algorithm as implemented in the PACES package was tested for the 27 proteins with assignment data listed in Table 1. In the cases of apoptotic protease activating factor I (Zhou et al., 1999), CIDE (Lugovskoy et al., 1999) and human carbonic anhydrase II (Venters et al., 1996b), original data sets with extraneous or missing peaks, and with the experimentally measured chemical shifts for all residues, were obtained from the authors and used for testing. These data sets were received from the authors as lists of possible spin systems with their resonance chemical shifts, which included multiple possible chemical shifts for some resonances, where it was not possible for the authors to determine the proper peak corresponding to that resonance prior to assignment. In these cases, the affected spin systems were duplicated, with the alternative values tested independently for the relevant resonance positions. For the other test proteins, published assignment data were obtained from the BMRB or from original publications, with all residues that had reported HN and N chemical shifts entered as spin-systems, and with the carbon chemical shifts of the preceding residue entered as inter-residue chemical shifts. In all of these tests, H^{α} chemical shifts were not used for analysis. Additional tests using H^{α} chemical shifts are included as Supplementary Material, available from the authors. Spin system index numbers were randomized before testing. For the calmodulin/M13 complex (Ikura et al., 1991), C^{β} chemical shifts were not available in the BMRB entry or the original publication, and were therefore extracted from the chemical shift database in the program TALOS (Cornilescu et al., 1999).

All testing was conducted on a 1.6 GHz Intel Pentium system running Windows 2000 with Microsoft

Table 1. Proteins used for PACES testing

Protein	Reference (BMRB Accession Number)	Data types available	Number of residues[a]	Number of residues with data[b]
Class I				
Malate synthase G	Tugarinov et al., 2002	Cα, Cβ, C′	723	654
Maltose binding protein	Gardner et al., 1998 (4354)	Cα, Cβ, C′	370	329
HIV-1 gag protein	Tang et al., 2002 (5316)	Cα, Cβ, C′	288	264
6-Phosphogluconolactase	Miclet et al., 2002 (5468)	Cα, Cβ, C′	266	239
Human carbonic anhydrase II	Venters et al., 1996	Cα, Cβ, C′	265	230
Rous Sarcoma Virus capsid	Campos-Olivas et al., 1999 (4384)	Cα, Cβ, C′	262	220
Human carbonic anhydrase I	Sethson, Ingmar, et al. 1996 (4022)	Cα, Cβ, C′	260	241
β-Lactamase	Scrofani, S.D.B. et al., 1998 (4102)	Cα, Cβ, C′	232	212
Peptide methionine sulfoxide reductase	Beraudi et al., 2001 (4844)	Cα, Cβ, C′	221	197
Hepatitis A 3C protease	Bjorndahl et al., 2001 (4836)	Cα, Cβ, C′	217	205
Peptide deformylase	Scahill et al., 2001 (4834)	Cα, Cβ, C′	183	165
Calmodulin/M13 complex	Ikura et al., 1991	Cα, Cβ, C′	148	144
Interleukin-4	Powers et al., 1992 (4094)	Cα, Cβ C′	133	128
Lysozyme	Kumeta et al., 2002 (5142)	Cα, Cβ, C′	130	126
Ferredoxin	Schweimer et al., 2000 (4444)	Cα, Cβ, C′	128	105
Bovine pancreatic ribonuclease A	Shimotakahara et al., 1997 (4032)	Cα, Cβ, C′	124	118
CIDE	Lugovskoy et al., 1999	Cα, Cβ, C′	116	106
Human ubiquitin	Wang et al., 1995	Cα, Cβ, C′	76	72
Class II				
Adenylate kinase	Burlacu-Miron et al., 1999 (4152)	Cα, Cβ	214	196
Human prion protein	Liu et al., 2000 (4402)	Cα, Cβ	210	189
Calmodulin/M13 complex	Ikura et al., 1991	Cα, C′	148	143
Profilin	Metzler et al., 1993 (4082)	Cα, Cβ	139	131
CIDE	Lugovskoy et al., 1999	Cα, Cβ	116	106
Apoptotic protease activating factor I (APAF I)	Zhou et al., 1999	Cα, Cβ	96	90
TFIIE core domain	Okuda et al., 2000 (4722)	Cα, Cβ	81	73
Human ubiquitin	Wang et al., 1995	Cα, Cβ	76	72
Yeast ubiquitin	Hamilton et al., 2000 (4769)	Cα, Cβ	76	66
Class III				
Epithelial cadherin (E-cadherin) domains II and III	Allatia et al., 2000 (4457)	Cα, Cβ	227	161
Superoxide dismutase	Vathyam et al., 1999 (4341)	Cα, Cβ, C′	192	105
E. coli EmrE	Schwaiger et al., 1998 (4136)	Cα, Cβ, C′	110	72

[a]Includes prolines and the N-terminal residue.
[b]Includes only those spin systems that were observed directly, and have reported HN and N chemical shifts (i.e. excludes prolines).

Excel 2000. For all proteins except human carbonic anhydrase II, the PACES program was run initially with chemical shift tolerances of 0.2 ppm for C^{α}, 0.4 ppm for C^{β} and 0.15 ppm for carbonyl, with the perdeuteration flag set for those proteins that were perdeuterated originally (Venters et al., 1996a; Farmer and Venters, 1999). The human carbonic anhydrase II data (Venters et al., 1996b) had been collected with somewhat lower spectral resolution, and therefore C^{α}, C^{β} and carbonyl chemical shift tolerances of 0.3 ppm, 0.5 ppm and 0.3 ppm, respectively, were used for it. For proteins that did not include either C^{α}, C^{β} or carbonyl chemical shifts, analyses of the corresponding chemical shifts were turned off. Assignments were completed following the procedure described above. If, due to excessive degeneracy, analysis could not

be conducted for a particular protein using the thresholds listed above, under the default fragment limit of 10 000, the thresholds were decreased by 25–50%, until the program could run, and assignments were completed using the reduced threshold procedure described earlier. In the case of calmodulin, analyzed without C^{β} data, it was also necessary to raise the fragment limit to 20 000 in order to complete processing. With the full length prion protein, it was not possible to complete processing prior to exhausting available memory.

In real data sets, resolution of the NMR spectrum, experimental noise, and the distortion or overlap of resonance peaks create uncertainty in the measured chemical shift values. This uncertainty must be reflected in the threshold used to establish sequential connectivity of spin systems. If the thresholds are set to encompass three standard deviations of the noise distribution on each side of the center point (roughly equal to, or slightly larger than, the spectral resolution), such uncertainty will have little effect on the process of fragment assembly because all connectivity possibilities falling within these thresholds will be considered. Occasionally an outlier pairing *jk* is observed in which the intra-residue chemical shift of spin system *j* and the inter-residue chemical shift of spin system *k* do not match within the specified threshold, although in reality they are sequentially connected. To assess the effects of these outliers on the assignment process, additional tests were conducted on the proteins that lacked experimental data, in which simulated noise was introduced, so that approximately 1% of each pairings would fall outside the chemical shift tolerances, thereby affecting 2–3% of overall connectivities. For each inter-residue resonance in a spin system, a noise deviation d was added, with the deviation determined as $d = N(0, \delta/2.5)$, where the function $N(\mu, \sigma)$ represents a random variable of normal probability density with mean μ and standard deviation σ, and where δ represents the chemical shift threshold for that spin system type. $N(\mu, \sigma)$ random variables were generated by the Polar Marsaglia method from random variables with uniform probability density (Morgan, 1984).

In order to compare the effects of having varying numbers of data sets on the assignments for a given protein, three of the proteins with C^{α}, C^{β} and carbonyl data were also tested using only a subset of that data. CIDE and human ubiquitin were tested with C^{α} and C^{β} only in these tests, while calmodulin was tested with C^{α} and carbonyl only, as was available in the original BMRB entry from Ikura et al. (1991).

Results

Overview

We have tested this algorithm, as implemented in the PACES package, with data from 27 proteins with very different characteristics, ranging in size from ubiquitin, with 76 residues, to malate synthase G, with 723 residues, and in the degree of data completeness from superoxide dismutase, with data for only 55% of the protein's residues spread out intermittently over its sequence, to nearly 100% in other cases. All testing was conducted using carbon chemical shifts only. Some of these test sets contained data for all three types of carbon chemical shifts, while others contained only two of the three. Thresholds of 0.2 ppm for C^{α}, 0.4 ppm for C^{β} and 0.15 ppm for carbonyl have been used during the tests unless stated otherwise. The detailed results of these tests are listed in Table 2; a summary of the results is given in Table 3.

Generally speaking, when data were present for the majority of a protein, the first runs of PACES typically took between 10 and 75 s using a 1.6 GHz Intel Pentium system, and provided unambiguous assignments for about 80% of the protein. The remaining residues could then be assigned by conducting additional runs according to the assignment protocol described above. These runs typically required ten seconds or less of processing time. When less data were available, the results varied somewhat depending upon how the missing residues were distributed across the protein sequence, and depending upon the quality of the data for the remaining residues. An example of PACES program output is shown in Figure 3. When data are plentiful the interpretation of PACES output is straightforward and proceeds rapidly; in situations of poor data, however, assignments are less certain, and the utility of the PACES output lies in the fact that it provides useful advice for manual spectral analysis.

The 27 proteins used for testing can be divided into three distinct classes, as indicated in Tables 1 and 2. Class I proteins had data for 80% or more of residues and contained, for most residues, all three types of carbon chemical shifts used by PACES. Class II proteins only had data for two types of carbon chemical shifts. Finally, three proteins with less than 80% of residues represented were designated as class III proteins.

Table 2. Detailed results of PACES testing

Protein	Number of residues with data	Original data				Data with simulated error			
		Number assigned	Fraction assigned	Number incorrect	Iterations needed[a]	Number assigned	Fraction assigned	Number incorrect	Iterations needed[a]
Class I									
Malate synthase G	654	640	98%	0	4	629	96%	0	5
Maltose binding protein	329	323	98%	0	2	310	94%	0	3
HIV-1 gag protein	264	263	100%	0	3	263	100%	0	3
Human carbonic anhydrase I	241	240	100%	0	3	224	93%	0	3
6-Phosphogluconolactase	239	236	99%	0	2	235	98%	0	4
Human carbonic anhydrase II	230	218	95%	0	6	N/A[b]	N/A[b]	N/A[b]	N/A[b]
Rous Sarcoma Virus capsid	220	207	94%	0	3	206	94%	0	2
β-Lactamase	212	203	96%	0	3	201	95%	0	3
Hepatitis A 3C protease	205	205	100%	0	2	198	97%	0	2
Peptide Met sulfoxide reductase	197	194	98%	0	3	195	99%	0	3
Peptide deformylase	165	164	99%	0	3	164	99%	0	2
Calmodulin/M13 complex	144	144	100%	0	2	144	100%	0	2
Interleukin-4	128	128	100%	0	1	128	100%	0	2
Lisozyme	126	126	100%	0	2	126	100%	0	2
Bovine pancreatic ribonuclease A	118	118	100%	0	1	118	100%	0	2
CIDE	106	106	100%	0	2	N/A[b]	N/A[b]	N/A[b]	N/A[b]
Ferredoxin	105	105	100%	0	2	105	100%	0	2
Human ubiquitin	72	72	100%	0	1	72	100%	0	1
Class II									
Adenylate kinase	196	193	98%	0	3	190	97%	0	5
Human prion protein (full length)	189	0[c]	0[c]	0	0[c]	0[c]	0[c]	0	0[c]
Human prion protein (126–230)	97	94	97%	0	3	95	98%	0	3
Calmodulin/M13 complex	143	143	100%	0	2	78	54%	0	1
Profilin	131	131	100%	0	2	130	99%	0	3
CIDE	106	91	86%	0	3	N/A[b]	N/A[b]	N/A[b]	N/A[b]
APAF I	90	87	97%	0	3	N/A[b]	N/A[b]	N/A[b]	N/A[b]
TFIIE core domain	73	69	95%	0	2	71	95%	0	1
Human ubiquitin	72	72	100%	0	1	72	100%	0	1
Yeast ubiquitin	66	64	97%	0	2	58	88%	0	2
CIass III									
E-cadherin domains II and III	161	113	70%	0	2	28	17%	0	1
Superoxide dismutase	105	77	73%	0	2	54	51%	0	1
E. coli EmrE	72	67	93%	0	4	59	82%	0	2

[a]One iteration consists in running the PACES program one time, analyzing the results manually, and making assignments.
[b]For this protein, experimental data was available, and it was not necessary to simulate experimental error.
[c]This protein could not be analyzed by PACES due to the high level of degeneracy present in the data.

Class I

Eighteen of the test proteins had C^{α}, C^{β} and carbonyl data for 80% or more of residues, and in each of these cases, assignment using the PACES package was rapid and straightforward – both with and without simulated experimental error. Proteins in this category varied in size from 76 to 723 amino acids. With all of these data sets it was possible to assign 95% or more of the residues. Although in a large protein the amount of degeneracy is substantial even with three types of carbon data present for each residue, this degree of de-

generacy does not prevent analysis by PACES, which can still be accomplished in a short amount of time. The amount of degeneracy does influence the degree to which unambiguous assignments could be made. In several situations, such as with malate synthase G, a very small fraction of the spin systems could not be assigned because two or more assignment possibilities remained for certain spin systems even after all others had been assigned. Additional constraints such as NOE information would be needed to resolve these issues (Tugarinov et al., 2002).

Class II

Proteins in this category contained only two pairs of carbon chemical shift data, consisting of either C^{α} and C^{β} or C^{α} and carbonyl. The results of PACES analysis varied somewhat depending upon the degree of degeneracy that was present in the data. With six of the proteins in this category, processing was rapid and assignment was straightforward for almost all residues. Analysis of adenylate kinase and profilin required using reduced thresholds during the first run. In the case of full-length human prion protein, the degeneracy was so severe that PACES analysis was not possible. This is not entirely surprising considering that the N-terminal half (residues 1–125) of the protein is completely disordered and has a very narrow chemical shift dispersion. When only data corresponding to the structured C-terminal domain (residues 126–230) of human prion protein were used (Zahn et al., 2000), the analysis by PACES was rapid and complete. To measure how the quality of results varies for a single protein using different numbers of data sets, three of the Class I proteins were also tested as Class II proteins, with one of the three data sets removed. For human ubiquitin and CIDE, testing was conducted with carbonyl data removed. The assignment of human ubiquitin proceeded without difficulty whether or not carbonyl data were present. Analysis of CIDE without carbonyl chemical shifts required using reduced chemical shift thresholds, but was straightforward. With calmodulin the original BMRB entry from Ikura et al. (1991), which did not contain C^{β} chemical shifts, was used for reduced data testing. Analysis with this data missing was possible, but required several hours of processing time, and was unable to yield assignments for substantial portions of the protein.

Class III

Proteins in Class III were missing data for a substantial portion of their residues. For EmrE, the residues with missing data were concentrated into a specific region of the protein sequence (residues 32–76), and our algorithm was easily able to supply unambiguous assignments for the remainder of the sequence, with or without simulated experimental noise. For the regions with mostly missing data, assignment suggestions were provided that included the correct assignments for the isolated data points, although it would not be possible to choose which of these highly ambiguous suggestions are correct on the basis of the PACES program output alone. With superoxide dismutase, however, about half of the available data points were scattered throughout the length of the protein, separated by numerous small gaps, as a result of paramagnetic relaxation of residues in the vicinity of an Fe^{3+} ion, while the other half existed as part of extended segments. Our algorithm had no difficulty assigning these extended segments, but the short segments and isolated residues that comprised the remainder of the data could not be assigned. The available data for E-cadherin covered long segments in domain II, and only isolated residues and short segments scattered sporadically throughout the unstructured domain III. A large amount of degeneracy coupled with having only C^{α} and C^{β} data meant that reduced thresholds were required for the initial PACES run. Assignment was straightforward using the original data, but the addition of simulated noise severely affected the ability to assign spin systems. If the thresholds were reduced enough to allow processing to proceed, a substantial portion of the connectivities in domain II were missed, leaving only very short fragments.

These three cases represent some of the most difficult situations encountered with sequential assignment. EmrE and E-cadherin both were missing data for a large number of residues in particular regions of their protein sequences, with the former due to an unusual T_2 relaxation behavior, and the latter to the missing residues in an unstructured domain. Superoxide dismutase belongs to a class of metalloproteins where signals were not observed for residues in the vicinity of a paramagnetic Fe^{3+} ion. PACES was able to handle these three situations effectively.

Testing with simulated noise outside the connectivity thresholds

In real data sets, the uncertainty of the peak positions for a small number of resonances may be larger than the connectivity threshold expected based on spectral resolution, leading to broken connectivities at these positions. To test the robustness of the PACES program in terms of experimental uncertainty, we calibrated our noise distribution so that about 2–3% of sequentially-connected spin systems would fall out of the defined connectivity thresholds. Although connectivity would be broken at these spin systems, we have designed interactive software tools for dealing with this situation that enable users to extend existing fragments, to merge broken fragments, or to fill in assignment gaps (Figures 4c and 4d). Sometimes, however, these poor pairings had more substantive global effects on the assignment process, by cutting long fragments into very short pieces, for example. As a result, for some of the proteins studied, placing 3% of connectivity pairings outside the connectivity thresholds disrupted assignments for more than 3% of residues. The effect of having noise larger than the connectivity threshold is most influential for proteins with a reduced number of data sets (class II) or incomplete data (class III), sometimes leading to a dramatic drop in the completeness of the assignment. However, for proteins with all three sets of connectivity data (class I), there is little difference in the final statistics for the completeness of the assignments (Tables 1 and 2).

Discussion

Sequential assignment is a prerequisite for solution structure determination, and efforts to automate and accelerate this process have continued to receive substantial attention. Although exhaustive search methods have been proposed as optimal solutions to many obstacles encountered with sequential assignment, it has been thought that the implementation of such methods would be impossible due to the astronomical number of potential outcomes. Here we show that the restrictions imposed by connectivity requirements between spin systems significantly reduce the number of possible solutions, making it feasible to analyze connectivities exhaustively. By aligning the fragments produced by such an analysis at every position along a protein sequence, the result is the set of all possible assignments – that is, all that are consistent with the requirements for establishing connectivity and mapping to the sequence. An interactive, iterative procedure involving assigning spin systems with certainty first, will reduce the number of solutions at other positions on the sequence, allowing most assignments to be completed quickly and enabling the user to focus on those few points of degeneracy that must be resolved by manual examination of the spectra.

We have implemented this approach in the PACES program, a flexible tool that is able to reduce substantially the time and effort required of an NMR spectroscopist to complete sequential assignments of even very large proteins. When high quality data are provided, sequential assignment with this method is a rapid and essentially automated procedure. However, this tool is generally able to make at least some use out of whatever data can be provided. Its ability to generate useful information even from data sets with very poor quality, and to integrate a wide variety of information from numerous sources to arrive at the correct assignment solution, make it a flexible tool for data analysis.

Errors vs. completeness of the assignment

The fact that our algorithm performs an exhaustive search of connectivity and assignment possibilities means that the correct assignment will always be presented as one of the possible assignments for any given position on the protein sequence *so long as* the data are complete, the intra- and inter-residue chemical shifts between sequentially connected spin systems match up within the user-defined thresholds, and the chemical shifts fall within the normal ranges for the amino acid types. Because of degeneracy, the correct solution might be one of several presented in some cases, but it would always be presented. Naturally these ideal conditions are rarely met in practice, and it is therefore theoretically possible for this algorithm to yield incorrect results for certain residues under extreme situations. Whether one is determining sequential assignments by hand or with a computer, however, there are always a handful of situations in which one could not tell that an assignment might be incorrect – such as when a spin system is missing or has atypical chemical shifts, and another spin system happens to fill the position instead. These situations are quite rare: After testing 27 proteins we have yet to encounter one. Nevertheless they are theoretically possible, and it is always wise to check assignments generated by this

method against other information, such as the cross-peaks observed in NOESY spectra. It is important to point out that the extremely low error rate of our approach is sometimes a trade-off with the completeness of the assignment. This is especially the case for class II and class III proteins with reduced data sets or incomplete data, where only a low percentage of the data can be unambiguously assigned.

Consequences of missing residues or atypical chemical shifts

The most common result encountered when spin systems are missing from the data is that a gap is left in the assignments where those spin systems should have been assigned. In some cases one or two spin systems that would be connected to a missing spin system will not appear as suggested assignments in PACES because they are now isolated as single spin systems or very short segments and are thus filtered out before presentation. PACES provides tools to locate and assign isolated residues that would often allow these assignments to be made later in the process (Figure 4b).

Residues that have unusual chemical shifts – as would result from binding a metal, for example – will likewise leave a gap in the suggested assignments (Figure 4d). It is frequently straightforward to assign these residues once 90% or more of the protein is assigned, as they will show connectivity to the spin systems assigned on each side of the gap position, and alternative connectivities at that position will usually have been eliminated. It is important to note that although using tighter chemical shift ranges to derive residue type information sometimes causes gaps in the initial assigned fragments, it improves the accuracy and reduces the uncertainty in the overall assignments (Andrec and Levy, 2002). To compensate for the effects of using tighter ranges, PACES also includes an option to extend all amino acid ranges 1 ppm higher and 1 ppm lower, which often allows residues with atypical chemical shifts to be located and assigned automatically after making an initial round of assignments (Figure 4e).

Thresholds for establishing connectivity

The thresholds used to establish connectivities must be set correctly for the algorithm to give accurate results. Ideally the thresholds should be set slightly larger than the uncertainty in peak positions expected on the basis of the digital resolution of the NMR spectra, ensuring that most correct pairings between sequentially connected spin systems are represented among the assembled fragments. It is important to remember, however, that the computational complexity of conducting an exhaustive search varies substantially depending upon the degree of degeneracy present in the data, and thresholds that are slightly too large can result in an observed degeneracy orders of magnitude beyond what is feasible for analysis. Choosing the appropriate thresholds for a given situation requires balancing the concerns about observing all of the real connectivities with the need to keep degeneracy within limits.

During our testing of the PACES program we used a C^{α} threshold of 0.2 ppm, a C^{β} threshold of 0.4 ppm, and a carbonyl threshold of 0.15 ppm, which reflect the level of spectral resolution most commonly used for triple resonance experiments. When these thresholds are used with all three sets of carbon resonances, degeneracy was not an issue even for the largest proteins tested (Figure 5). With data for C^{α} and C^{β} or C^{α} and carbonyl only (class II), however, our results varied substantially – four proteins could be analyzed without difficulty, three required reduced tolerances, two required several hours of processing, and one could not be analyzed at its full length.

The effects of a reduced number of data sets on the robustness of the program

Our results suggest that collecting triple-resonance data for C^{α}, C^{β} and carbonyl carbon nuclei would be highly recommended when conducting sequential assignment using the PACES program. With three sets of data, the percentage of residues that can be assigned is higher, the number and length of program runs are decreased, and most importantly, the robustness of the program against experimental uncertainties is significantly improved over the results with two sets (Tables 2 and 3). However, the PACES program was designed to be flexible such that it is possible to generate useful results without meeting these conditions, as we have demonstrated with eight proteins (class II, Tables 2 and 3). How difficult this may be, both in terms of the ambiguity in the suggested assignments and also the computational complexity of running the program, varies considerably depending upon the particular protein. Whenever possible, the program should be run with tolerances equal to or greater than the spectral resolution; it is possible, however, to employ a strategy of using reduced thresholds initially to work around

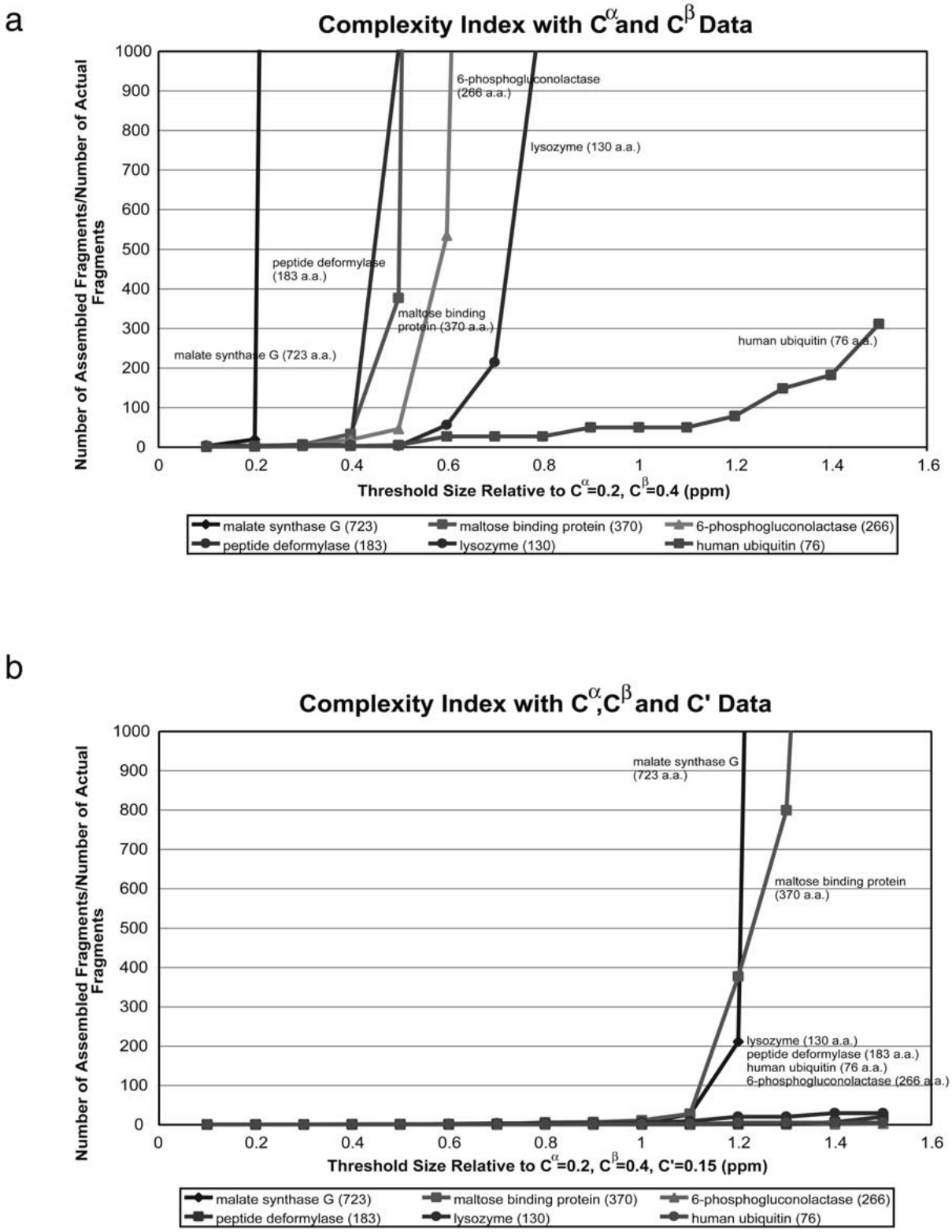

Figure 5. Analysis of the complexity of sequential assignment using PACES. (a) Results with data for C^{α} and C^{β} only. (b) Results with data for C^{α}, C^{β} and C'. In each graph, the *y*-axis represents the complexity index, the ratio of fragments generated by PACES over the expected minimal number of fragments separated by prolines in a protein. A complexity index of 1.0 indicates a single solution with no degeneracy. The *x*-axis represents the relative threshold sizes as a proportion to the thresholds used in assignment testing ($\delta_{C\alpha}$ of 0.2 ppm, $\delta_{C\beta}$ of 0.4 ppm and $\delta_{C'}$ of 0.15 ppm).

Table 3. Summary of PACES test results

Protein class/ test category	Number of proteins analyzed	Number of residues with data	Number of assigned residues	Fraction of data points assigned	Number of incorrect assignments	Error rate
Class I	18					
Original data		3755	3692	98.3%	0	0.0%
Data with simulated error		3419	3318	97.0%	0	0.0%
Experimental data		336	324	96.4%	0	0.0%
Class II	9					
Original data		974	944	96.9%	0	0.0%
Data with simulated error		778	694	89.2%	0	0.0%
Experimental data		196	178	90.8%	0	0.0%
Class III	3					
Original data		338	257	76.0%	0	0.0%
Data with simulated error		338	141	41.7%	0	0.0%
All Proteins	30					
Original data		5067	4893	96.6%	0	0.0%
Data with simulated error		4535	4153	91.6%	0	0.0%
Experimental data		532	502	94.4%	0	0.0%

problems with excessive degeneracy. The combination of C^{α} and C^{β} data works better than that of C^{α} and carbonyl, because C^{β} chemical shifts are more dispersed and enable more precise determinations of amino acid types than carbonyl chemical shifts.

Naturally, having an additional set of connectivity constraints from H^{α} (Olejniczak et al., 1992) would make the assignment process faster and more reliable, and we have provided the capability to use this information. Testing results using H^{α} connectivity are provided as Supplementary Material.

The complexity of the sequential assignment problem

The complexity of sequential assignment is directly related to the degree of degeneracy present in the resonance data, and thus is greatly influenced by the size of the protein, the number of paired data sets used to deduce connectivity, and the threshold used to establish such connectivity. In order to obtain a better understanding of the complexity of the sequential assignment problem, we employed the PACES program to investigate the number of connectivity fragments assembled for proteins of varying size, containing different numbers of data sets, and with different thresholds. For each protein there are a minimum number of fragments that one would expect to find, separated by proline residues. Any fragments that are generated beyond this number are extraneous, resulting from degeneracy. Thus, the ratio of the number of fragments generated during the fragment assembly process over the expected number of fragments would be a good index of the complexity of the sequential assignment problem. A complexity index of 1.0 indicates that there is a single solution for the assignments, with no degeneracy and thus no alternative outcomes. Figure 5 shows plots of such a complexity index as a function of the chemical shift threshold used, for six proteins, with C^{α} and C^{β} data in one plot and C^{α}, C^{β} and carbonyl data in the other. To reduce the number of parameters, the thresholds have been defined as a ratio of those being used to determine complexity to those used for the assignment testing (0.2 ppm for C^{α}, 0.4 ppm for C^{β} and 0.15 ppm for carbonyl). When three sets of data for C^{α}, C^{β} and carbonyl carbons were available, degeneracy was only an issue for the two largest proteins tested – maltose binding protein and malate synthase G – and even for these it was not a problem so long as the tolerances were held to within

10% of the values used during the assignment testing (0.22 ppm, 0.44 ppm, and 0.17 ppm for C^{α}, C^{β} and carbonyl, respectively). For others, thresholds as large as 0.3 ppm, 0.6 ppm and 0.23 ppm for C^{α}, C^{β} and carbonyl, respectively, do not cause problems. When only C^{α} and C^{β} data are available, however, the complexity of sequential assignment can quickly become a problem for all but the smallest of proteins, even at very narrow thresholds.

Other factors, such as the completeness of the spin system data and the uncertainty of the measured chemical shifts also add to the complexity of the sequential assignment by generating even more connectivity fragments during the assembly process. For a given protein, however, these factors only represent a linear portion of the complexity of the sequential assignment problem, which otherwise grows exponentially as a function of the chemical shift thresholds used to establish the connectivity fragments. We have therefore excluded these considerations from the above discussions.

Multiple conformers

Stemming from the fact that it retains multiple assignment possibilities, our algorithm has the unique capability of being able to track assignments for multiple conformers simultaneously. A multiple conformer mode is included in the PACES program, in which assigning a spin system to a residue prevents that spin system from being assigned elsewhere in the protein, but does not prevent additional spin systems from being positioned at that point on the protein sequence. This should enable one to assemble alternative assignments for sections of protein sequence, corresponding to different conformers.

Conclusions

We have developed a novel approach to protein sequential assignment based on the methods used by NMR spectroscopists to assign proteins manually, but taking advantage of the abilities of a computer to work through all of the possibilities for connecting peaks and mapping them to the protein sequence. Our exhaustive search algorithm generates a set of spin system segments anchored at positions on the protein sequence, providing numerous assignment possibilities. By assigning those that have the most certainty first – those that are longest and have the most contiguous residue-to-spin-system matches – and gradually working through to less certain assignments, it is possible to progress rapidly through almost all of the assignments for a protein with relatively complete data. For data of lower quality, this method is frequently able to yield assignments for those parts of the protein represented in the data. When multiple assignments are possible at a position on the sequence, it can provide them all for consideration. We have tested PACES using data from 27 proteins and found that it has been able to produce correct results in every situation examined, including with proteins as large as 723 amino acids. We believe that this method has great potential to accelerate sequential assignment, and to simplify the work involved in determining protein solution structure by NMR.

Acknowledgements

We thank Drs Ronald Venters and Leonard Spicer for stimulating discussions and for critical reading of the manuscript, and we thank R.V. further for graciously allowing us to work with some of his unpublished data while developing this program. We thank Dr Christian R.H. Raetz for suggesting the name 'PACES'. B.E.C. was supported in part by NIH grant GM51310 to C.R.H.R. P.Z. was supported in part by the Duke University Junior Faculty Start-up Fund and the Whitehead Institute.

References

Alattia, J., Tong, F.K., Tong, K.I. and Ikura, M. (2000) *J. Biomol. NMR*, **16**, 181–182.

Andrec, M. and Levy, R.M. (2002) *J. Biomol. NMR*, **23**, 263–270.

Atreya, H.S., Sahu, S.C., Chary, K.V. and Govil, G. (2000) *J. Biomol. NMR*, **17**, 125–136.

Bartels, C., Güntert, P., Billeter, M. and Wüthrich, K. (1997) *J. Comp. Chem.*, **18**, 139–149.

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

Beraud, S.B., Chambost, J.B.B., Gans, P.J.R., Barras, F. and Marion, D. (2001) *J. Biomol. NMR*, **20**, 97–98.

Bjorndahl, T.C., Watson, M.S., Slupsky, C.M., Spyracopoulos, L., Sykes, B.D. and Wishart, D.S. (2001) *J. Biomol. NMR*, **19**, 187–188.

Buchler, N.E., Zuiderweg, E.R., Wang, H. and Goldstein, R.A. (1997) *J. Magn. Reson.*, **125**, 34–42.

Burlacu-Miron, S., Perrier, V., Gilles, A., Mispelter, J., Barzu, O. and Craescu, C.T. (1999) *J. Biomol. NMR*, **19**, 93–94.

Campos-Olivas, R., Newman, J.L., Ndassa, Y. and Summers, M.F. (1999) *J. Biomol. NMR*, **15**, 267–268.

Cornilescu, G., Delaglio, F. and Bax, A. (1999) *J. Biomol. NMR*, **13**, 289–302.

Dötsch, V. and Wagner, G. (1996) *J. Magn. Reson. B*, **111**, 310–313.

Dötsch, V., Matsuo, H. and Wagner, G. (1996a) *J. Magn. Reson. B*, **112**, 95–100.
Dötsch, V., Oswald, R.E. and Wagner, G. (1996b) *J. Magn. Reson. B*, **110**, 107–111.
Dötsch, V., Oswald, R.E. and Wagner, G. (1996c) *J. Magn. Reson. B*, **110**, 304–308.
Farmer, 2nd, B.T. and Venters, R.A. (1996) *J. Biomol. NMR*, **7**, 59–71.
Farmer, 2nd, B.T. and Venters, R.A. (1999) In *Biological Magnetic Resonance*, Vol. 16, pp. 75–120.
Gardner, K.H., Zhang, X., Gehring, K. and Kay, L.E. (1998) *J. Am. Chem. Soc.*, **120**, 11738–11748.
Gronwald, W., Willard, L., Jellard, T., Boyko, R.F., Rajarathnam, K., Wishart, D.S., Sonnichsen, F.D. and Sykes, B.D. (1998) *J. Biomol. NMR*, **12**, 395–405.
Güntert, P., Salzmann, M., Braun, D. and Wüthrich, K. (2000) *J. Biomol. NMR*, **18**, 129–37.
Hamilton, K.S., Ellison, M.J. and Shaw, G.S. (2000) *J. Biomol. NMR*, **18**, 319–327.
Ikura, M., Kay, L.E., Krinks, M. and Bax, A. (1991) *Biochemistry*, **30**, 5498–5504.
Kumeta, H., Kobashigawa, Y., Miura, K., Nishimiya, Y., Oka, C., Nemoto, N., Miura, A., Nitta, K. and Tsuda, S. (2002) *J. Biomol. NMR*, **22**, 183–184.
LeMaster, D.M. and Richards, F.M. (1985) *Biochemistry*, **24**, 7263–7268.
Leutner, M., Gschwind, R.M., Liermann, J., Schwarz, C., Gemmecker, G. and Kessler, H. (1998) *J. Biomol. NMR*, **11**, 31–43.
Li, K.-B. and Sanctuary, B.C. (1997a) *J. Chem. Inf. Comput. Sci.*, **37**, 467–477.
Li, K.-B. and Sanctuary, B.C. (1997b) *J. Chem. Inf. Comput. Sci.*, **37**, 359–366.
Lugovskoy, A.A., Zhou, P., Chou, J., McCarty, J.S., Li, P. and Wagner, G. (1999) *Cell*, **99**, 747–755.
Lukin, J.A., Gove, A.P., Talukdar, S.N. and Ho, C. (1997) *J. Biomol. NMR*, **9**, 151–166.
Markley, J.L., Bax, A., Arata, Y., Hilbers, C.W., Kaptein, R., Sykes, B.D., Wright, P.E. and Wüthrich, K. (1998) *Pure Appl. Chem.*, **70**, 117–142.
Metzler, W.J., Constantine, K.L., Friedrichs, M.S., Bell, A.J., Ernst, E.G., Lavoie, T.B. and Mueller, L. (1993) *Biochemistry*, **32**, 13818–13829.
Miclet, E., Duffieux, F., Lallemand, J. and Stoven, V. (2002) BioMagResBank, accession number 5468.
Morgan, B.J.T. (1984) *Elements of Simulation*, Chapman and Hall, Ltd., London.
Moseley, H.N.B. and Montelione, G.T. (1999) *Curr. Opin. Struct. Biol.*, **9**, 635–642.
Okuda, M., Watanabe, Y., Okamura, H., Hanaeoka, F., Ohkuma, Y. and Nishimura, Y. (2000) *EMBO J.*, **19**.
Olejniczak, E.T., Xu, R.X., Petros, A.M. and Fesik, S.W. (1992) *J. Magn. Reson.*, **100**, 444–450.
Powers, R., Garret, D.S., March, C.J., Frieden, E.A., Gronenborn, A.M. and Clore, G.M. (1992) *Biochemistry*, **31**, 4334–4346.
Scahill, T.A., Kloosterman, D.A., Cialdella, J.I., Deibel, Jr., M.R., Marshall, V.P. and Yem, A.W. (2001) *J. Biomol. NMR*, **19**, 81–82.
Schubert, M., Oschkinat, H. and Schmieder, P. (2001a) *J. Magn. Reson.*, **153**, 186–192.
Schubert, M., Oschkinat, H. and Schmieder, P. (2001b) *J. Magn. Reson.*, **148**, 61–72.
Schubert, M., Smalla, M., Schmieder, P. and Oschkinat, H. (1999) *J. Magn. Reson.*, **141**, 34–43.
Schwaiger, M., Lebendiker, M., Yerushalmi, H., Coles, M., Groeger, A., Schwarz, C., Schuldiner, S. and Kessler, H. (1998) *Eur. J. Biochem.*, **254**, 610–619.
Schweimer, K., Marg, B., Oesterhalt, D., Roesch, P. and Sticht, H. (2000) *J. Biomol. NMR*, **16**, 347–348.
Scrofani, S.D.B., Wright, P.E. and Dyson, J.H. (1998) *J. Biomol. NMR*, **12**, 201–202.
Sethson, I., Edlund, U., Holak, T.A., Ross, A. and Jonsson, B. (1996) *J. Biomol. NMR*, **8**, 417–428.
Shimotakahara, S., Rios, C.B., Laity, J.H., Zimmerman, D.E., Scheraga, H.A. and Montelione, G.T. (1997) *Biochemistry*, **36**, 6915–6929.
Tang, C., Ndassa, Y. and Summers, N.F. (2002) *Nat. Struct. Biol.*, **9**, 537–543.
Tashiro, M., Rios, C.B. and Montelione, G.T. (1995) *J. Biomol. NMR*, **6**, 211–216.
Tian, F., Valafar, H. and Prestegard, J.H. (2001) *J. Am. Chem. Soc.*, **123**, 11791–11796.
Tugarinov, V., Muhandiram, R., Ayed, A. and Kay, L.E. (2002) *J. Am. Chem. Soc.*, **124**, 10025–10035.
Vathyam, S., Byrd, R.A. and Miller, A. (1999) *J. Biomol. NMR*, **14**, 293–294.
Venters, R.A., Farmer, 2nd, B.T., Fierke, C.A. and Spicer, L.D. (1996a) *J. Mol. Biol.*, **264**, 1101–1116.
Venters, R.A., Farmer, B.T.I., Fierke, C.A. and Spicer, L.D. (1996b) *J. Mol. Biol.*, **264**, 1101–1116.
Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids,* John Wiley and Sons, New York, NY.
Yamazaki, T., Lee, W., Arrowsmith, C.H., Muhandiram, D.R. and Kay, L.E. (1994a) *J. Am. Chem. Soc.*, **116**, 11655–11666.
Yamazaki, T., Lee, W., Revington, M., Mattiello, D.L., Dahlquist, F.W., Arrowsmith, C.H. and Kay, L.E. (1994b) *J. Am. Chem. Soc.*, **116**, 6464–6465.
Zahn, R., Liu, A., Luhrs, T., Riek, R., von Schroetter, C., Lopez Garcia, F., Billeter, M., Calzolai, L., Wider, G. and Wüthrich, K. (2000) *Proc. Natl. Acad. Sci. USA*, **97**, 145–150.
Zhou, P., Chou, J., Olea, R.S., Yuan, J. and Wagner, G. (1999) *Proc. Natl. Acad. Sci. USA*, **96**, 11265–11270.
Zimmerman, D.E., Kulikowski, C.A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C., Powers, R. and Montelione, G.T. (1997) *J. Mol. Biol.*, **269**, 592–610.