Journal of Biomolecular NMR (2005) 33: 261–279
DOI 10.1007/s10858-005-4079-8

Article

CASA: An efficient automated assignment of protein mainchain NMR data using an ordered tree search algorithm

Jianyong Wang[a], Tianzhi Wang[c], Erik R.P. Zuiderweg[c] & Gordon M. Crippen[b,c,*]
[a]*Department of Physics, University of Michigan, Ann Arbor, MI, 48109-1120, USA;* [b]*College of Pharmacy, University of Michigan, Ann Arbor, MI, 48109-1055, USA;* [c]*Biophysics Research Division, University of Michigan, Ann Arbor, MI, 48109-1055, USA*

Received 5 September 2005; Accepted 5 October 2005

Key words: automatic assignment, generic spin system, ordered tree search, triple resonance

Abstract

Rapid analysis of protein structure, interaction, and dynamics requires fast and automated assignments of 3D protein backbone triple-resonance NMR spectra. We introduce a new depth-first ordered tree search method of automated assignment, CASA, which uses hand-edited peak-pick lists of a flexible number of triple resonance experiments. The computer program was tested on 13 artificially simulated peak lists for proteins up to 723 residues, as well as on the experimental data for four proteins. Under reasonable tolerances, it generated assignments that correspond to the ones reported in the literature within a few minutes of CPU time. The program was also tested on the proteins analyzed by other methods, with both simulated and experimental peaklists, and it could generate good assignments in all relevant cases. The robustness was further tested under various situations.

Introduction

The assignment of the protein backbone resonances is necessary and sufficient for protein structure determination based on residual dipolar couplings, mapping of protein–protein interaction sites based on chemical shift mapping, and the determination of backbone dynamics on the subnanosecond and micro-second time scale. The backbone assignments form also the root for NOE-based NMR structure determinations (Wagner and Wüthrich, 1982; Wüthrich, 1986). Nowadays, main chain assignments are almost exclusively made from combinations of 3D (sometimes 4D) triple resonance experiments on isotopically labeled proteins introduced a decade ago (Ikura et al., 1990; Montelione and Wagner, 1990). Even though the potential for automation of assignments based on triple resonance data was immediately realized (Ikura et al., 1990), most NMR laboratories still carry out the assignment process essentially by hand. This is mostly due to the fact that available automated assignment programs require a certain set of spectra, cannot deal with noisy, incomplete spectra, or with spectra of proteins in multiple slowly interchanging conformations, and run into convergence problems for the assignment of larger proteins. (Moseley and Montelione, 1999; Moseley et al., 2001).

Here we introduce a new depth-first ordered tree search method of automated assignment, CASA (Combinatorial Automatic Sequential Assignment), which uses hand-edited peak-pick lists of a flexible number of triple resonance experiments. We show that it is capable of

*To whom correspondence should be addressed: E-mail: gcrippen@umich.edu assigning the spectra of very large proteins. With the advent of highly sensitive NMR systems, such as cryo-probes, and experiments, such as Triple-resonance-TROSY, high quality triple resonance spectra, in principle suitable for automated assignment, can be acquired on even very large proteins in reasonable time. We anticipate that our new method will contribute to the rapid assessment of protein structure, interactions, and dynamics of known and unknown structures by NMR.

Most implementations for automated assignment have employed optimization algorithms that minimize a pseudo-energy score function, using neural networks (Hare and Prestegard, 1994), simulated annealing (Bernstein et al., 1993; Kraulis, 1994; Morelle et al., 1995), mean-field simulated annealing (Buchler et al., 1997), genetic algorithms (Wehrens et al., 1993; Bartels et al., 1997) or Monte Carlo optimization (Lukin et al., 1997; Leutner et al., 1998; Hitchens et al., 2003). A key characteristic of these global optimization methods is the tendency to generate a complete assignment which is globally correct but may be locally ambiguous in preference to incomplete but high-quality assignments. Exhaustive and heuristic search (Atreya et al., 2000; Bailey-Kellogg et al., 2000a, b; Guntert et al., 2000; Coggins and Zhou, 2003; Jung and Zweckstetter, 2004) reduces the ambiguity by classifying residues into groups or mapping connected segments globally. The AutoAssign program by Zimmerman et al. (1997) uses, like CASA, best-first algorithms (Li and Sanctuary, 1997; Zimmerman et al., 1997; Moseley et al., 2001; Montelione, 2005, personal communication). However, it has not been tested on very large proteins with 300 residues or more.

The depth-first ordered tree search method presented here can use a flexible number of NMR peak pick lists. The computer program was tested on four experimental peak-pick data, and also on 13 artificially simulated peak pick lists for proteins up to 723 residues. In all cases, the program generated assignments that correspond to the ones reported in the literature. The program used only a few minutes of CPU time for even the largest data sets. We also tested CASA on the proteins analyzed by other methods using their tolerances. In almost all cases, we could get comparable or better assignment scores within minutes.

Many proteins show incomplete NMR spectra because of exchange broadening of resonances belonging to areas involved in milli- to microsecond conformational exchange. Also common are proteins that show multiple assignment pathways for regions that are in slow conformational exchange. While both types of proteins are often deemed less desirable for study in an NMR structural proteomics context, they are often very interesting in a biological context. We thus required our assignment procedure to be able to handle such cases. We simulated the spectra of proteins with intermediate exchange by eliminating resonances associated with some contiguous parts of the sequence. We could still get a high assignment score even when a substantial portion of chemical shifts were deleted. For proteins in slow conformational exchange, we simulated the spectra by duplicating and shifting the resonances corresponding to contiguous areas. The program could solve the problem unambiguously.

Methods

CASA uses a sequential approach for the assignment. First, HN roots are constructed from the peak picked data; second, the HN-roots are sorted into T-units; third, the T-units are sorted into generic spin systems (GS; Zimmerman et al., 1993, 1994); fourth, the GSs are linked into segments; and last, the segments are placed onto the sequence using an ordered tree search algorithm (see Figure 1).

Generating HN roots

The program is based on the well-known triple-resonance assignment scheme, and is currently set up to accept NMRPipe-format peak-picked data from HNCA, HN(CO)CA, HNCACB, HN(CO)-CACB (or CBCA(CO)NH), HN(CA)CO, HNCO, HN(CA)HA, HN(COCA)HA, or HA(CACO)NH (see Table 1). The different triple resonance spectra share the same backbone amide H^N–N resonances, which are called HN roots (Zimmerman et al., 1993, 1994). HN roots are used as a filter to group the resonance spectra from different experiments into "T-units" corresponding to pairs of residues along the sequence (Van Doren et al., 1993). HNCO spectra have the best sensitivity, and provide the information for generating HN roots (Zimmerman et al., 1997). If HNCO data are not provided, we can also work with HNCA spectra,

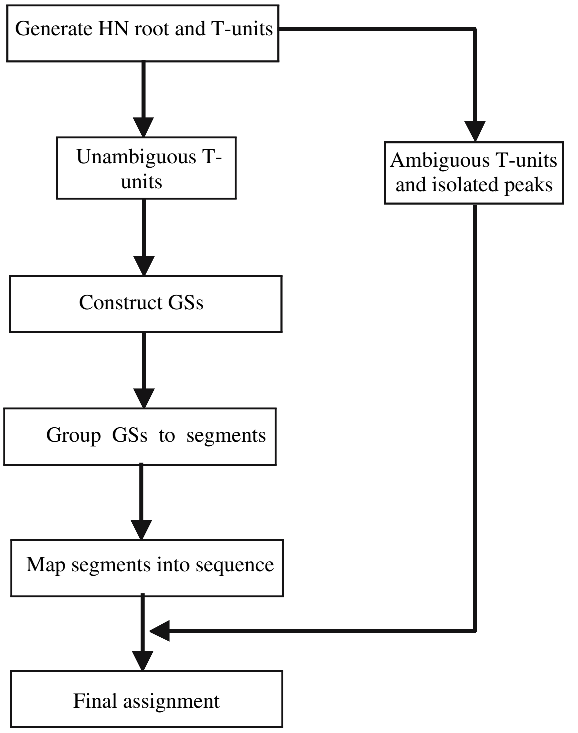
Generate HN root and T-units
Unambiguous T-units
Ambiguous T-units and isolated peaks
Construct GSs
Group GSs to segments
Map segments into sequence
Final assignment

Figure 1. Overview of the CASA assignment procedure.

Table 1. Types of experiments

Sequential	Intra-residue
HNCO	HN(CA)CO
HN(CO)CA	HNCA
HN(CO)CACB	HNCACB
HN(CO)(CA)H	HN(CA)H

which have lower but comparable sensitivity. HNCA spectra, with HN(CO)CA if provided, are used to obtain the chemical shift correlation between $H^N(i)$, $N(i)$, $C^\alpha(i)$ and $C^\alpha(i-1)$ for each residue (Van Doren et al., 1993) (see Figure 2). We calculate the distances of H^N and N chemical shifts between any two peaks. If these two distances are smaller than the tolerance of $H^N(\Delta H^N)$ and the tolerance of N (ΔN), respectively (see Table 2), these two peaks may form a T-unit. Sometimes, the respective H^N and N chemical shifts of four peaks are very close to each other, due to the overlap of two T-units. If the information of HN(CO)CA spectra is also provided, it is sometimes possible to separate the overlapping T-units when the two peaks of one T-unit are stronger than those of the other T-unit, because the *i*th (intra-residue) peak is usually stronger than the (*i* – 1)th (sequential) peak in the same T-unit. The H^N-N resonance from the separated C^α T-units can then be used as HN roots. Otherwise the overlapping T-units are set aside from the initial sorting set and will be treated at a later stage (see below).

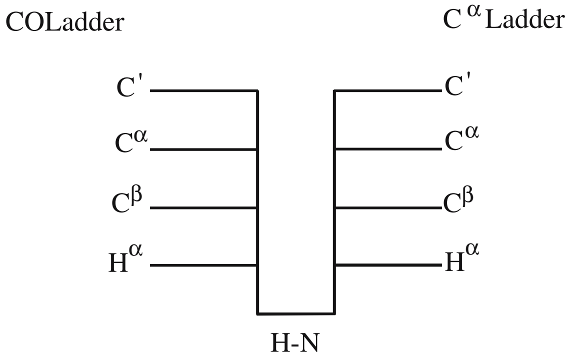

Figure 2. Definition of the Generic Spin System (GS; Zimmerman et al., 1993).

Table 2. Default tolerances (ppm)

	ΔH^N	ΔN	ΔC^α	ΔC^β	$\Delta C'$	ΔH^α
Condition I (ppm)	0.01	0.2	0.5	0.5	0.25	0.025
Condition II (ppm)	0.01	0.2	0.2	0.4	0.15	0.015

Constructing generic spin systems (GS)

We follow the general strategy as described in AutoAssign (Zimmerman et al., 1997; Moseley et al., 2001). Similar to forming C^α T-units from HNCA/HN(CO)CA spectra, we also form T-units from other triple resonance spectra, such as C^β T-units from HNCACB and HN(CO)CACB (if available). We group all available C^α T-units, C^β T-units, CO T-units and H^α T-units according to their HN roots to construct Generic Spin Systems (GS) (Zimmerman et al., 1997) (see Figure 2), the minimal assignable unit in our model. Each GS has two types of ladders, a C^α ladder and a CO ladder, containing chemical shifts of atoms of residue *i* in the sequence, and those of the residue *i* – 1, respectively. The number of rungs in each ladder may be different by nature of the residue, experimental artifacts, or missing experiments. For example, the C^{α} ladder of a glycine has a C^{α} rung, but no C^{β} rung by nature of the residue.

Link score

If the chemical shifts of all atoms in the C^{α} ladder of GS *i* match those of the corresponding atoms in the CO ladder of GS *j*, the link score $L(i, j)$ is 1; otherwise $L(i, j)$ is 0. Here a match means that the differences of corresponding chemical shifts are smaller than the corresponding tolerances (see Table 2). When the numbers of rungs of the corresponding ladders are different, we consider only the rungs appearing in both ladders. This strategy takes into account that rungs can be missing because of lack of signal to noise. If $L(i, j) = 1$, GS *i* and GS *j* can occupy sequentially adjacent sites along the sequence; if $L(i, j) = 0$, we should *not* assign GS *j* to a site following GS *i* (see Figure 3). Note that $L(i, j)$ does not equal $L(j, i)$ in general. It is possible that one GS may be linked favorably to several other GSs due to the degeneracy of spectra. This is one of the sources of ambiguity in assignment. The degeneracy of the spectra can be partially solved by including more rungs in the GSs. As we can see in the results, 3-rung links deal much better with degeneracy than 2-rung links.

If additional information on a sequential link is provided, such as that from NOE crosspeaks (Wüthrich, 1986), the user of CASA can also manually fix the link score between the corresponding GSs. By fixing the link between some specific GSs, the ambiguity of assignment can be greatly reduced, especially for spectra with severe degeneracy.

Occupation score

The characteristic distribution of chemical shifts associated with different residue types can be used to obtain typing scores for these chemical shifts. The binary typing score of a rung *r* in a GS for a residue of type *t* is evaluated as

$$S(r, t) = \begin{cases} 1 & \text{if} |p - \bar{p}_t| \leq R_p \cdot \sigma_t \\ 0 & \text{otherwise} \end{cases}$$

where $\bar{p}_t$ and σ_t are the mean and standard deviation of this type of chemical shift for residue type *t* obtained from the BioMagResBank (Seavey et al., 1991; http://www.bmrb.wisc.edu), and $R_p = 5$ for H chemical shifts, while $R_p = 4$ for those from other types of atoms, because the former have larger dispersion. The typing score equals 1 means that the chemical shift matches the corresponding residue type. Due to the overlap of chemical shifts of different residues, a chemical shift may match a few residue types simultaneously. This is another source of ambiguity of assignment. The typing scores of the rungs inside a GS allow it to be mapped onto some sites in the sequence while preventing it from being mapped onto other sites. The binary occupation score of a GS *j* to the *i*th site in the sequence $Occu(i, j)$ is 1 if all the CA rungs match the residue at the *i*th site and all the CO rungs match the residue at the $i - 1$st site. $Occu(i, j)$ is also set to be 1 if there is only one mismatch of all the rungs. This is

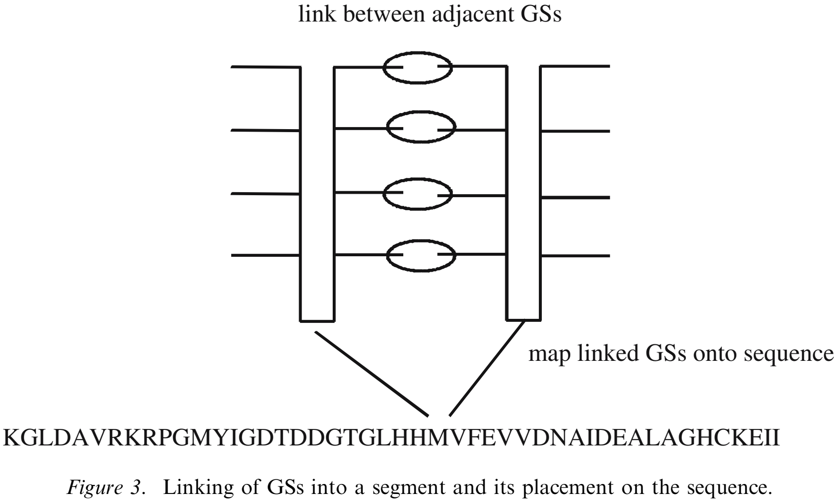
link between adjacent GSs
map linked GSs onto sequence
KGLDAVRKRPGMYIGDTDDGTGLHHMVFEVVDNAIDEALAGHCKEII

Figure 3. Linking of GSs into a segment and its placement on the sequence.

important to deal with the cases involving atypical chemical shifts. If there is more than one mismatch of the rungs, $Occu(i, j)$ is 0, which means that GS j should *not* be mapped to site i. On average, the typing score of a 3-rung (C^{α}, C^{β} and C') C^{α} ladder matches it to eight possible residue types, which is similar to the result of AutoAssign (Zimmerman et al., 1997). If the protein under study is perdeuterated, the chemical shifts can be adjusted accordingly (Venters et al., 1996; Moseley et al., 2004).

Additional constraints derived from a variety of sources can be utilized to manually fix the occupation score of some GSs (Coggins and Zhou, 2003; Jung and Zweckstetter, 2004). Residue-type information can be 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, b, c; Dötsch and Wagner, 1996; Schubert et al. 1999; Schubert et al., 2001a, b).

Forming Segments

In order to form a unique link between GS i and j, four criteria must be satisfied:

(1) $L(i, j) = 1$
(2) $L(i, k) = 0 \quad \text{for } k \neq j$
(3) $L(k, j) = 0 \quad \text{for } k \neq i$
(4) at least two rungs are matched when evaluating the link score $L(i, j)$.

When GS i and GS j are uniquely linked together, they form a segment $S(i, j)$. $S(i, j)$ can then be assigned as a unit, always occupying two sequentially adjacent sites. If GS j is also uniquely followed by another GS k, $S(i, j)$ can be updated to a longer segment $S(i,j,k)$. The segments are constructed based only on the link score between GSs. The occupation score of a segment j at the sequential sites starting from i is 1 only when each GS in the segment has a favorable occupation score at the corresponding site. Otherwise, the segment is forbidden to be mapped to those sites. Thus there is less ambiguity in mapping a segment than that in mapping a single GS, and the longer the segment, the smaller the number of assignable sites in the sequence. In general, a 5-GS or larger segment may be mapped uniquely onto the sequence, which is similar to the result of MARS (Jung and Zweckstetter, 2004). Note that the smallest segment contains only one GS which has no unique favorable link to any other GS.

Assigning segments with ordered tree search

In this approach, the assignment consists of mapping segments onto assignable sites in the sequence via a depth-first ordered tree search algorithm. For a protein with chain length N, the maximum number of assignable sites is

$$N_{\mathrm{a}} = N - 1 - N_{\mathrm{pro}}$$

where the site of the N-terminal residue (no HN root because of exchange) and sites occupied by prolines (no HN roots) cannot be assigned to any GS.

Briefly, the procedure is as follows. The root of the tree corresponds to no segment assigned, and the various leaves to all segments assigned, all available sites occupied, or the unoccupied sites unassignable to any unassigned segments. Each branching node corresponds to certain segments assigned to certain sites in the sequence. Each allowed node has no unfavorable occupation score or link score. Segments are sorted by increasing order of numbers of assignable sites, and the segment with the smallest number of available sites is placed first on the sequence for the best possible occupation score, followed by placing the segment with the second smallest number of available sites on the remaining sites in the sequence, and so on until one of the leaves is reached (i.e. depth-first). If all segments are assigned or all available sites are occupied, this leaf is viewed as an assignment, and the process will stop unless it is set up to find alternative assignments (see below). If there are still some segments and sites unassigned but no available sites in the sequence for the unassigned segments, that leaf is viewed as a dead end. The process backtracks up the tree to explore alternative placements of the previous segment until these are exhausted, backtracks up another level in the tree, and so on, until an assignment is found or the tree has been explored.

Assigning overlapping T-units

After the segments are assigned, it is possible to separate and assign the overlapping T-units.

Consider an empty site in the sequence (assignable site with missing spectra) adjacent to one or two assigned GSs. The favorable link score required by the assigned GSs picks out specific C$^\alpha$ and CO ladder chemical shifts, which helps retrieve correct T-units and enlarge the assigned segments (Atreya et al., 2000).

Ambiguity score of an assignment

The reliability of an assignment strongly depends on the quality of the spectra, such as completeness, correctness, the number of rungs for connection, etc. For a specific assignment, the ambiguity score of a segment is defined to be the number of assignable sites on the sequence given the assignment of all the other segments. Then, the ambiguity score of the assignment is taken to be the mean segment ambiguity score. The smaller the ambiguity score, the less likely a segment is mapped onto a wrong site. The smallest assignment ambiguity score is one, corresponding to all the segments being uniquely assigned. If the spectra are very incomplete, or the number of rungs for connection is not big enough, the ambiguity score could be very large. In this case, there are many assignments with similar assignment score, and it is very hard to distinguish the correct assignment from the alternative assignments unless additional constraints from occupation score or link score are available.

Comparison of CASA with other automatic assignment methods

For the purpose of comparing CASA with other automatic assignment methods, we downloaded MARS from the corresponding website (version 1.1.3, http://www.mpibpc.gwdg.de/abteilungen/030/zweckstetter/_links/software.htm), used the web servers for Redpoll (http://redpoll.pharmacy. ualberta.ca/~shan/cgi-bin/ssass.cgi) and PISTACHIO (Eghbalnia et al., 2005) (http://bija.nmrfam.wisc.edu/PISTACHIO), and obtained an authorized copy of AutoAssign (version 1.1.5, July 2005). For the web servers of Redpoll and PISTACHIO, we input data with the required format and used the default tolerances on these servers. For MARS and AutoAssign, we tested the example data distributed with these programs, and all of them generated correct assignments on their example data.

Results

Overview

It is necessary to test CASA rigorously on proteins with various sizes, degrees of data completeness, and degrees of data degeneracy. Here we simulated the data of proteins that were commonly tested by other methods (TATAPRO, PACES and MARS) (see Table 3). Of these proteins, the size ranges

Table 3. CASA assignments of proteins with simulated data and data quality

Protein	BMRB code	Number of residues	Number of pro/gly	C$^\alpha$(%)[a]	C$^\beta$(%)[a]	C′ (%)[a]	H$^\alpha$(%)[a]
Malate synthase G	5471	723	31/51	96	96	96	N/A[b]
Dnak-Tth	6229	381	18/35	92	89	91	N/A[b]
Maltose binding protein	4354	370	21/29	95	95	88	N/A[b]
GluR2 extracellular ligand-binding domain	5182	263	7/25	97	96	96	N/A[b]
Rous sarcoma virus capsid	4384	262	23/20	95	90	93	N/A[b]
Human carbonic anhydrase I	4022	260	17/16	99	98	93	N/A[b]
E-cadherin domains II and III	4457	227	14/12	73	72	N/A[b]	65
Human prion protein	4402	210	15/43	98	98	N/A[b]	78
Thiopurine methyltransferase	5820	203	9/18	98	98	93	N/A[b]
Superoxide dismutase	4341	192	8/14	73	72	57	N/A[b]
Calmodulin/M13	547	148	2/11	100	N/A[b]	100	92
Profilin	4082	139	4/16	99	99	N/A[b]	N/A[b]
E. *coli* EmrE	4136	110	5/12	83	56	72	N/A[b]

[a] Percentage of available chemical shifts.

[b] Data not available.

from *E. coli* EmrE with 110 residues to malate synthase G with 723 residues. The degree of completeness varies from superoxide dismutase, with data for only 55% of the protein's residues spread out intermittently over its sequence, to nearly 100% of others. We also tested CASA on the human prion protein, which is a challenge to automated assignment because of its narrow chemical shift dispersion and severe degeneracy. The original chemical shifts of these proteins were taken from the BMRB database. If H^N and N chemical shifts were available for a certain residue *i*, the intra-residue chemical shifts were constructed with the chemical shifts of this residue. The inter-residue chemical shifts were constructed by the combination of H^N and N chemical shifts of residue *i* and the carbon (or H^α) chemical shifts of residue $i - 1$, if available.

CASA was tested on these proteins using different numbers of spectra, according to the size of the proteins. For small proteins (chain length less than 200), fewer spectra were used so that we constructed and assigned the GSs with only 2 rungs (C^α and C^β). For large proteins (chain length larger than 200), it is necessary to include C' or H^α chemical shifts to guarantee fast and reliable assignment. Similar to MARS, we tested each protein under two tolerance conditions, namely 0.5, 0.5 and 0.25 ppm (condition I) and 0.2, 0.4 and 0.15 ppm (condition II) for C^α, C^β and C', respectively (see Table 2). For human prion protein and calmodulin/M13, the chemical shifts of H^α were introduced, and the corresponding tolerance is 0.025 ppm (condition I) or 0.015 ppm (condition II).

It is difficult but necessary to simulate the effects of overlap, line broadening, missing resonances and spectral artifacts with simulated peak list data. In CASA, only unambiguous paired rungs are included into a GS. Applying a tight tolerance for pairing and aligning, the probability of introducing paired noise or artifacts is very low. On the other hand, the number of GSs and/or the number of rungs inside GSs may be smaller than expected, because some experimental error may distort some of the H^N and N chemical shifts to be outside the tight pairing tolerance. As a result, some rungs in the GSs or even entire GSs are missing. In order to simulate the experimental errors introduced into realistic data sets, CASA was also tested with random deletion of GSs as well as random deletion of certain rungs within GSs. We also tested the case when the distortion of certain chemical shifts brought about the loss of sequential links between GSs.

Furthermore, we tested CASA on the experimental chemical shifts provided by other automatic assignment methods, such as mt0895 (thioredoxin like protein) (Bhattacharyya et al., 2002) from Redpoll; Z domain (Tashiro et al., 1997), RNase A wildtype (Shimotakahara et al., 1997) and fgf (basic fibroblast growth factor) (Moy et al., 1995) distributed with the AutoAssign software (see Table 5).

We also tested CASA on the experimental chemical shifts of four proteins collected in Prof. Zuiderweg's lab, namely, ubiquitin, calmodulin (without peptide), GRPE and CTD (see Table 6). All these experiments were performed on a Bruker Avance 500 MHz NMR spectrometer, and generated data of average sensitivity. The spectra were processed using the NMRPipe program (Delaglio et al., 1995). Peak-pick lists generated by NMRPipe were then edited by hand to erase obvious noise and side-chain signals.

The CASA assignment was carried out in MOE (http://www.chemcomp.com) using the SVL computer language. CASA and AutoAssign ran on a Sun Ultra 10 workstation with a 333 MHz CPU, while MARS ran on a SGI O2 workstation with a 180 MHz CPU.

Test on simulated chemical shifts

A. Disordered protein with severe degeneracy

The N-terminal half (residues 1–125) of the human prion protein (210 residues) is completely disordered, which results in a very narrow chemical shift dispersion. Furthermore, a large proportion of the residues are glycines (41 out of 210 residues). Thus severe degeneracy characterizes the spectra of this protein and poses a significant challenge to sequential assignment. Using only data corresponding to the structured C-terminal domain (residues 126–230) (Coggins and Zhou, 2003), the assignment was rapid by either using 2-rung GSs (C^α and C^β) under condition II or 3-rung GSs (C^α, C^β and H^α) under condition I with at least 98% of the residues assigned correctly. When the full-length protein was tested, the assignment score decreased greatly, but there were still 132 GSs assigned correctly using 3-rung connection under condition II (see Table 4).

Table 4. CASA assignment results for proteins with simulated data

Protein	Number of GSs[a]	Rungs in GSs	Condition I assignment number		Condition II assignment number		Simulated errors[d]	
			All[b]	Correct[c]	All[b]	Correct[c]	All[b]	Correct[c]
Malate synthase G	653	$C^{\alpha}C^{\beta}C'$	519	451	653	649	556	513
Dnak-Tth[e]	324	$C^{\alpha}C^{\beta}C'$	295	279	324	317	276	254
Maltose binding protein	328	$C^{\alpha}C^{\beta}C'$	328	326	328	326	279	242
GluR2 extracellular ligand-binding domain	245	$C^{\alpha}C^{\beta}C'$	245	245	245	245	209	207
Rous sarcoma virus capsid	220	$C^{\alpha}C^{\beta}C'$	218	208	218	208	186	164
human carbonic anhydrase I	242	$C^{\alpha}C^{\beta}C'$	242	240	242	242	206	196
		$C^{\alpha}C^{\beta}$	N/A[h]	N/A[h]	242	242	218	214
E-cadherin domains II and III	133	$C^{\alpha}C^{\beta}H^{\alpha}$[f]	133	88	133	118	N/A[i]	N/A[i]
		$C^{\alpha}C^{\beta}$	133	57	133	79	N/A[i]	N/A[i]
Human prion protein	189	$C^{\alpha}C^{\beta}H^{\alpha}$[f]	145	121	149	132	151	91
		$C^{\alpha}C^{\beta}$	96	65	132	119	113	73
Thiopurine methyltransferase	189	$C^{\alpha}C^{\beta}C'$	189	189	189	189	161	161
Superoxide dismutase	117	$C^{\alpha}C^{\beta}C'$	117	107	117	108	N/A[i]	N/A[i]
		$C^{\alpha}C^{\beta}$	117	83	117	100	N/A[i]	N/A[i]
calmodulin/M13	144	$C^{\alpha}C'H^{\alpha}$[g]	129	89	144	144	130	86
		$C^{\alpha}C'$[g]	N/A[h]	N/A[h]	132	131	123	59
profilin	132	$C^{\alpha}C^{\beta}$	132	132	132	132	113	113
E. *coli* EmrE	79	$C^{\alpha}C^{\beta}C'$	79	50	79	70	68	55
		$C^{\alpha}C^{\beta}$	79	30	79	48	68	45

[a] Counting only those residues with HN root and at least one T-unit.
[b] The number of GSs mapped onto the sequential sites.
[c] The number of GSs mapped onto the correct sequential sites.
[d] Simulated errors were introduced by random deletion of 15% of GSs as well as random deletion of rungs within 15% of the remaining GSs under condition II.
[e] Peak list data provided by E.R.P. Zuiderweg.
[f] C' chemical shifts are missing in the original BMRB entry,H^{α} chemical shifts were introduced to construct GS.
[g] C^{β} chemical shifts are missing in the original BMRB entry,H^{α} chemical shifts were introduced to construct GS.
[h] No reliable assignment could be generated within a reasonable time scale in this case.
[i] Original spectra are far from complete. It is not necessary to delete additional GSs or rungs.

B. Large proteins

The assignment of six large proteins with more than 250 residues were tested by CASA, namely human carbonic anhydrase I (260 residues), Rous sarcoma virus capsid (262 residues), GluR2 extracellular ligand-binding domain (263 residues), maltose binding protein (370 residues), Dnak-Tth (381 residues), and malate synthase G (723 residues). The peak pick lists were constructed from the available BMRB databank. Except for human carbonic anhydrase I, which was assignable using only 2-rung GSs (C^{α} and C^{β}) under condition II, all the other large proteins required 3- rung GSs (C^{α}, C^{β} and C') for CASA to generate results with significant assignment scores. For the 723-residue malate synthase G, the assignment was accomplished within 200 s and only four out of 653 GSs were not assigned to the correct sequential sites. Three out of these 4 GSs, corresponding to residues 94, 159, and 537, were surrounded by either empty sites or sites corresponding to prolines; and the fourth GS, corresponding to residue 456, had only a C^{α} rung. Considering the larger number (38) of empty sites in the sequence, it's reasonable that there is flexibility to assign these residues to alternative sites due to the insufficient constraints from link and/or occupation score. This leads to some alternative assignments which have similar assignment scores (all GSs mapped into the sequence with no unfavorable occupation or link score) but

Table 5. Tests on the chemical shifts provided by other methods

		Mt0895[a]	Z domain[b]	RNase A wildtype[b]	fgf[b]	Maltose binding protein[c]
Number of residues		77	71	124	154	370
Number of pro/gly		1/7	3/0	4/3	9/16	21/29
Number of GSs		72	65	119	141	328
Spectra available		HNCACB	HSQC, HNCO, HNCA, HN(CO)CA, HNCACB, HN(CO), CACB, HNHA, HN(CO)HA	HSQC, HNCO, HN(CA)CO, HNCA, HN(CO)CA, HNCACB, HN(CO), CACB, HNHA, HN(CO)HA	HSQC, HNCO, HN(CA)CO, HNCA, HN(CO)CA, HNCACB, HN(CO), CACB, HNHA, HN(CO)HA	HNCO, HN(CA)CO, HNCA, HN(CO)CA, HNCACB, HN(CO), CACB
Default tolerances		H^N: 0.025 N: 0.4 CA: 0.3 CB: 0.3	H^N: 0.025 N: 0.35 CA: 0.5 CB: 0.75 H^α: 0.05	H^N: 0.025 N: 0.35 CA: 0.8 CB: 0.8 C′: 0.5 H^α : 0.13	H^N: 0.02 N: 0.25 CA: 0.4 CB: 0.6 C′: 0.2 H^α: 0.05	H^N: 0.01 N: 0.2 CA: 0.2 CB: 0.4 C′: 0.15
CASA	I(%[d])[e, f]	94[k]	50	100	61	85
	II(%[d])[e, g]	97[k]	100	100	99	99
MARS	I(%[d])[e, f]	92[k]	52	100	60	94
	II(%[d])[e, g]	97[k]	100	100	99	99
Redpoll(%[d])[h]		100	0	0	8	5
PISTACHIO(%[d])[i]		N/A[m]	72	N/A[m]	80	82
AutoAssign(%[d])[j]		N/A[m]	100	100	100	1[l]

[a] Experimental peaklist was provided by redpoll.

[b] Experimental peaklists were provided by Autoassign.

[c] Simulated peaklist from BMRB data.

[d] Assignment scores.

[e] CASA used the default tolerance of H^N and N to construct GSs, which were also used by MARS. For both CASA and MARS, 2-rung GSs were assigned for mt0895, Z domain, RNase A and fgf; 3-rung GSs were assigned for maltose binding protein.

[f] Link tolerance of condition I was used.

[g] Link tolerance of condition II was used.

[h] Default tolerances of redpoll were used for all proteins.

[i] Default tolerances of PISTACHIO were used for all proteins.

[j] Default tolerances were used for all proteins.

[k] CASA and MARS could not assign the peaklist, but assigned the GSs simulated from the assignment of redpoll.

[l] Since this is a simulated peaklist for a large protein, we also tested it using the parameter *no_referencing* and/or *deuterated* in Autoassign, but found no improvement in the assignment score.

[m] No assignment was generated.

Table 6. Test on experimental data and assignments of four proteins

Protein		Ubiquitin	Calmodulin	GRPE	CTD
Number of residues		76	148	137	163
Number of pro/gly		3/6	2/11	7/13	12/15
Spectra available		HNCA, HNCOCA, HNCACB	HNCA, HNCACB	HNCA, HN(CO)CA, HNCACB, HN(CO), CACB, HNCO, HN(CA)CO	HNCA, HN(CO)CA, HNCACB, HN(CO), CACB, HNCO, HN(CA)CO
Number of GSs assigned	CASA[a]	69 (69)[e]	104 (75)[e]	98 (71)[e]	114
	MARS[a]	69 (68)[e]	94 (70)[e]	66 (65)[e]	96
	Redpoll[b, f]	N/A[g]	N/A[g]	57 (12)[e]	96
	PISTACHIO[c]	N/A[g]	N/A[g]	80 (42) [e]	111
	AutoAssign[d]	N/A[g]	N/A[g]	27 (20)[e]	107

[a] CASA and MARS used the GSs constructed by CASA. For both CASA and MARS, link tolerances of condition II were used, and 2-rung GSs were assigned for ubiquitin and calmodulin, 3-rung GSs for GRPE and CTD.

[b] Default tolerances of redpoll were used.

[c] Default tolerances of PISTACHIO and all spectra were used.

[d] Tolerance of H^N: 0.02 and N: 0.2 were used for constructing GSs. Link tolerances of condition II were used for all proteins. AutoAssign was not applicable for ubiquitin and calmodulin. All spectra were used for GRPE and CTD.

[e] In parenthesis is the number of GSs assigned correctly.

[f] HNCA and HNCACB spectra were used for ubiquitin, HNCA, HNCACB and HN(CO)CACB were used for GRPE and CTD.

[g] No assignment was generated.

differ in detailed placing of the GSs. The correct assignment corresponds to one of the alternative assignments, but it's not distinguishable from the others without additional constraints.

C. Proteins with incomplete chemical shift data
EmrE, superoxide dismutase and E-cadherin are missing H^N/N chemical shifts for a substantial portion of their residues (Coggins and Zhou, 2003; Jung and Zweckstetter, 2004), resulting in a large number of empty sites in the sequence. For EmrE, the residues with missing data were concentrated into a specific region of the protein sequence (residue 32–76). Using only 2-rung (C^α, C^β) GSs for linking, all the 79 GSs were assigned, but only a small proportion of the GSs were assigned to the correct sites (38% for condition I, 61% for condition II). However, if the additional C′ information was included into the GSs, the assignment score was improved dramatically (63% for condition I, 89% for condition II).

For superoxide dismutase, only 61% of the expected pseudoresidues were listed in the data bank entry, and about half of the available GSs are scattered throughout the length of the protein. Nevertheless, even using only 2-rung GSs (C^α, C^β) in condition I, more than 70% of the GSs were assigned correctly, but 3-rung GSs performed much better than 2-rung, resulting in nearly complete assignment (97% in condition II).

For E-cadherin, the available data covers long segments in domain II, but only isolated residues and short segments scattered sporadically throughout the unstructured domain III. Again, the assignment was much better for the 3-rung GSs (60% for condition I, 89% for condition II) than the 2-rung case (42% for condition I, 59% for condition II).

D. Assignment without C^β rungs
The original BMRB entry for calmodulin did not contain C^β chemical shifts, so GSs with C^α, C′ and/or H^α rungs were constructed. The assignment of this small protein was possible for 2-rung GSs only under condition II. Using 3-rung GSs under condition II, the assignment was rapid and complete (assignment score was 100%). However, deletion of only 10% of the GSs and rungs caused a great decrease in assignment score (66%). At the same time, a large portion (80%) of the sequential links were conserved in this assignment. This interesting result corresponded to the fact that most of the GSs linked correctly to each other and formed segments, but quite a few of the segments were assigned to the wrong sequential sites due to insufficient constraints from the occupation score. Therefore, the chemical shifts of C^β play an important role in providing sufficient constraints on typing score (Atreya et al., 2000). Missing C^β chemical shifts may result in uncertainty in the assignment when the number of empty sites is not negligible.

E. Complexity of assignment
The complexity of the assignment depends on the size of the proteins, the degeneracy and the completeness of the spectra data, the sizes of the tolerances, and the number of rungs in the GSs. So it is very hard to compare the complexity among different proteins. Figure 4 shows the relation of execution time and the number of segments vs. the

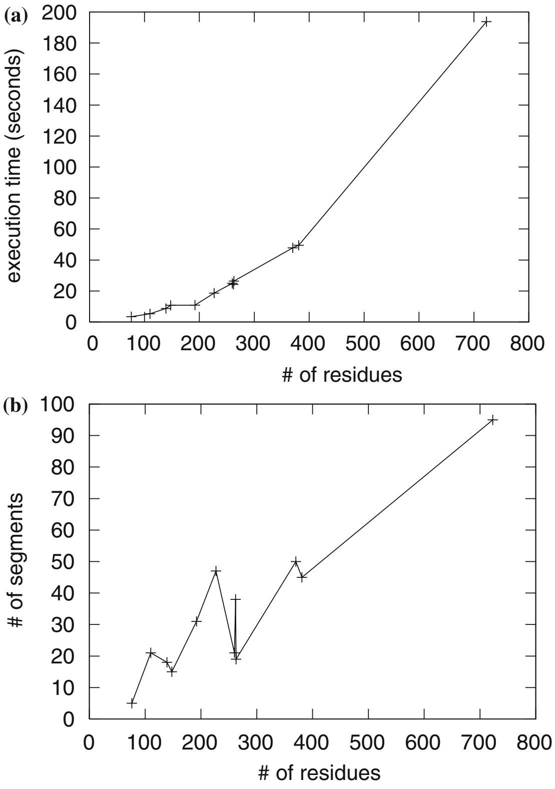

Figure 4. (a) Completion time of the CASA assignment for different proteins using 3-rung GSs under condition II. All the tests were carried out in MOE (http://www.chemcomp.com) using the SVL computer language on a Sun Ultra 10 workstation. (b) the number of segments vs. the number of residues using 3-rung GSs under condition II.

chain length using 3-rung GSs under condition II. In this situation, the increase of running time was monotonic, and the assignment was still rapid for a protein as large as malate synthase G, and the assignment was nearly complete. However, when using fewer rungs or larger tolerances, the substantial increase of the degeneracy of the chemical shift data may result in a larger number of shorter segments, and the constraints of typing and linking for these segments will be much looser than those for longer ones. The execution time increases exponentially, eventually making the assignment impossible.

F. Assignment score vs. ambiguity score

Figure 5 is a scatter plot of assignment score vs. ambiguity score, tested on the simulated data of proteins in Table 3 under different link conditions (2-rung/3-rung connection, link condition I and II). Although the test proteins differ greatly in chain length, data completeness and data degeneracy, and the test was conducted under different link conditions, there is an obvious correlation between assignment score and ambiguity score. When the ambiguity score was small (≤ 3), the assignment scores were always high (> 90%); as the ambiguity score increases, the assignment score drops. We can also see that the dispersion of assignment scores at small ambiguity scores was smaller than that at large ambiguity scores. When the ambiguity score is small, the constraint from occupation and link score dominates the ambiguity, and thus a correct assignment is guaranteed regardless of the difference in proteins and link conditions. However, there are insufficient constraints at a large ambiguity score, while the assignment score depends more on data quality and link conditions.

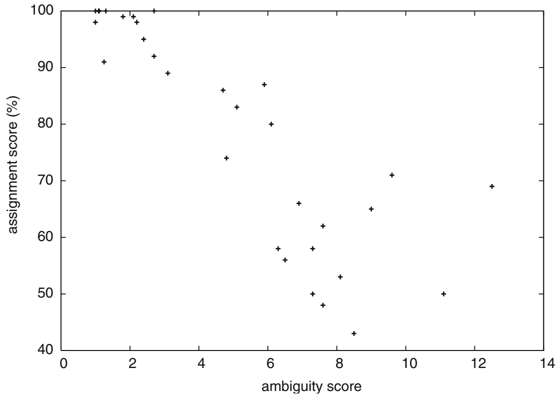

Figure 5. Assignment score vs. ambiguity score.

G. Comparison to other methods

Table 3 covers all the proteins with simulated data tested by MARS (Jung and Zweckstetter, 2004), except the N-terminal domain of enzyme I (EIN) whose N chemical shifts were missing in the original BMRB entry. Using 3-rung GSs under condition II, CASA could get comparable assignment scores for all the proteins in MARS, but it ran much faster (200 s compared to 2 h for malate synthase G). Using 2-rung GSs with large tolerances for large proteins, the execution time increases exponentially, and CASA could not get reliable assignments within a reasonable time scale. Our conclusion is similar to that of AutoAssign (Moseley et al., 2001), namely that using a 2-rung strategy has a great impact on the robustness of the assignment, and introducing more rungs in GSs increases the speed and reliability of assignment dramatically.

Part of the proteins in Class I and II, as well as all of the Class III proteins in PACES (Coggins and Zhou, 2003) were included in Table 3. We further tested all the remaining proteins in PACES with simulated data using their tolerances (same as condition II). For all these proteins, CASA could get comparable or better assignment scores without any intervention. For Class II and Class III proteins, it's not necessary to introduce reduced tolerances at the beginning.

The five proteins tested by TATAPRO (Atreya et al., 2000) with simulated data were also tested by CASA using their tolerances (0.5, 0.2, and 0.25 ppm for C^{α}, C^{β} and C′, respectively). All the proteins were assignable on the order of seconds without any intervention or reduced tolerances, and our assignment scores were comparable to theirs.

Tests on experimental chemical shifts

A. Tests on the data provided by other methods

For the protein mt0895 provided by Redpoll, only a HNCACB peaklist was available (see Table 5). Note that there are 560 peaks in the HNCACB peaklist for this 77-residue protein, considerably more than the maximum of 308. In this crude form, it is not assignable for CASA, or PISTACHIO (Eghbalnia et al., 2005). It is also not assignable for AutoAssign because that method requires at least additional HN(CO)CACB information. However, we can artificially construct GSs from the assignment of Redpoll, by using the peaks of H^N and N of one assigned residue as the HN root of a GS, the peaks of the C^α and C^β of this residue as the CA rungs in the GS, and the peaks of the C^α and C^β of the previous residue as the CO rungs in the GS. If the assignment is correct, all the GSs constructed should also be correct. Using these GSs, both CASA and MARS could assigned most of the GSs (>90%) correctly (compared to the assignment of Redpoll). PISTACHIO and AutoAssign do not accept GSs constructed by other programs (or by hand); hence these programs could not be further tested for mt0895.

The experimental chemical shifts of Z domain, RNase A wildtype, and fgf were provided by AutoAssign. CASA and MARS could assign all these proteins using 2-rung connection, and we can see that the assignment scores were much lower under condition I than those under condition II for both CASA and MARS. PISTACHIO could assign Z domain and fgf using 2-rung connection with high assignment scores. Redpoll used simulated experimental HNCACB data in NMRPipe format, but the assignment scores were low for all these proteins.

We used simulated data from BMRB results for the last protein, maltose binding protein. It provides a test of the assignability of all these methods for large proteins. Due to the great chain length of this protein, we tested all the methods, except Redpoll, with 3-rung connections. CASA, MARS and PISTACHIO could assign more than 80% of the GSs correctly. AutoAssign failed on this simulated data. The number of GSs constructed was smaller than expected, and some of the GSs had incorrect rungs assembled to their HN roots, especially incorrect C^β. Redpoll can accept up to three spectra, HNCACB, HNCA, CBCA(CO)NH, so it can only use 2-rung connection, which made it difficult for redpoll to assign such a large protein.

B. Tests on our experimental data

Ubiquitin is a small protein with 76 residues and 72 assignable sites (see Table 6). Using a tolerance of 0.01 and 0.1 ppm for H^N and N, respectively, CASA constructed 69 C^α T-units from HNCA and HN(CO)CA spectra; subsequently 61 C^β T-units from HNCACB were aligned to these C^α T-units. Altogether 69 GSs were constructed. Using a link tolerance of 0.4 ppm for both C^α and C^β, 22 segments were formed, with only one 8-GS, 7-GS, or 6-GS segment, two 4-GSs segments, three 5-GS or 2-GS segments, four 3-GS segments, and seven 1-GS segments. CASA completed the entire assignment within 4 s. Comparison of this assignment with the published one (http://www.bmrb.wisc.edu) showed that all 69 GSs were assigned to the correct sites, while the chemical shifts of residues 24(glu), 28(ala) and 53 (gly) were missing in the original spectra. Using the GSs constructed by CASA and the same link tolerances, MARS assigned 68 GSs correctly. Redpoll, PISTACHIO, and AutoAssign could not generate any assignment with our experimental ubiquitin data.

We then tested CASA on our experimental data of calmodulin (without peptide) with only HNCA and HNCACB spectra available. 106 C^α T-units could be picked out unambiguously from the HNCA spectrum under the tolerance of 0.01 and 0.2 ppm for H^N and N, respectively. Since the HN(CO)CA spectrum was not available, it was impossible to extend the list of C^α T-units by separating overlapping C^α T-units or matching a single intra-residue peak in HNCA to the corresponding sequential peak in HN(CO)CA. At the same time, the quality of the HNCACB data was much worse. Using the same pairing tolerance, only 47 unambiguous C^β T-units could be picked out and aligned to the corresponding C^α T-units. As a result, there are 39 empty sites in the sequence and 59 GSs (more than 50% of all GSs) with only C^α rungs. Compared to the manual assignment, only 35 GSs were assigned correctly under condition II due to the great ambiguity in the data. This suggests that the quality of picked peaks and assembled GSs is crucial for successful assignment, and manual inspection of the original spectra may be necessary to guarantee such quality (Jung and Zweckstetter, 2004). The spectra were then manually analyzed using the Xeasy program (Bartels et al. 1995), and degeneracy could often be resolved when the information on peak shapes was taken into account. In total, 122 C^α T-units and 79 C^β T-units were picked out to construct 122 GSs. Under condition II for the link tolerance, CASA constructed 80 segments from these GSs, with one 5-GS, one 4-GS, nine 3-GS segments, 17 2-GS segments and 52 1-GS segments. Although there was still considerable ambiguity, 104 GSs were assigned, out of which 75 GSs were assigned correctly (assignment score 61%) within 100 s (as shown in Table 6). Compared to the ideal spectra, our assignment was based on spectra where 15% of the GSs were missing and 35% of the available GSs had only C^{α} rungs. This assignment score (61%) is consistent with the test result on the simulated data with random deletion of GSs and rungs (see Discussion). Using the 122 manually constructed GSs, MARS assigned 94 GSs under condition II, with 70 GSs assigned correctly. However, Redpoll, PISTACHIO and AutoAssign could not generate any assignment in this case.

The third protein we tested was GRPE, with 137 residues and 7 prolines. Due to the limited quality of the spectra, we had difficulty in constructing C^{β} Ts and aligning them to the HN roots. With the help of manual inspection, 98 GSs were constructed, but only 47 out of them had C^{β} rungs. Under condition II, 42 segments were constructed, with two 9-GS, one 6-GS, three 5-GS and 4-GS, five 3-GS, four 2-GS, and 24 single GS segments. With such incomplete data, CASA assigned 71 GSs correctly. Noticing that there were 27 GSs assigned incorrectly, we further checked the ambiguity score of each segment. One group of 15 segments, mostly single GSs segments, had ambiguity scores no smaller than 25, while the scores of all the other segments were no larger than 13. Discarding the assignment of the 15 most ambiguous segments, CASA assigned 82 GSs out of which 69 GSs were assigned correctly; the number of wrong assignments was greatly reduced. Using the same GSs and link condition, MARS assigned 65 GSs correctly. PISTACHIO and AutoAssign assigned 42 and 20 GSs correctly, respectively. Using three spectra, HNCACB, HNCA, and HN(CO)CACB, Redpoll assigned 12 GSs correctly using 2-rung connection.

We finally tested CASA with protein CTD, a 163-residue protein with 12 prolines, for which currently we do not know the correct assignment. Using a tolerance of 0.02 and 0.15 ppm for H^{N} and N, respectively, 116 C^{α} Ts, 72 C^{β} Ts, and 82 C' Ts were picked out and aligned to construct 116 GSs. Under condition II, 59 segments were formed, with one 10-GS, 7-GS and 5-GS segments, two 6-GS, five 3-GS, 12 2-GS, and 35 1-GS segments. Within 100 s, CASA assigned 114 GSs onto the sequence. We also tested these data on MARS, PISTACHIO, AutoAssign and Redpoll using the same tolerances. For MARS, the GSs constructed by CASA were used as input. For the assignment results from PISTACHIO, AutoAssign, and redpoll, we artificially constructed GSs using the methods as described for mt0895. Since we do not know the correct answer, we can only make comparison of the assembly and assignment of the GSs from different methods. The numbers of GSs assigned by the five methods are quite similar among each other, ranging from 96 GSs (MARS) to 114 GSs (CASA), and most of the HN roots of those GSs are common for all the methods. However, the rungs in the GSs with the same HN roots might be different for different methods, possibly because the spectra of different experiments may shift a little, and different methods have different criteria and techniques to align them. This is not surprising since the manually assembled GSs may also depend on the experience and preference of the experimentalist, resulting in different results from person to person. As we have discussed, the difference in rungs of GSs may change the link score and occupation score greatly, so the difference in the assignment of these GSs could be even larger, as we can see in Table 7. In general, CASA,

Table 7. Comparison of assignments of different methods for protein CTD

	CASA	MARS	Redpoll[c]	PISTACHIO	AutoAssign
CASA	114[a]	96[a,b]	66	69	87
MARS	75	96	66	69	87
Redpoll[c]	28	26	96	43	70
PISTACHIO	63	66	20	111	76
AutoAssign	72	85	28	74	107

[a] In the diagonal of the table are the numbers of GSs assigned by the corresponding methods; in the upper-right of the table are the numbers of GSs constructed in common between the corresponding two methods; in the lower-left of the table are the numbers of GSs assigned in common by the corresponding two methods.

[b] Mars used the GSs constructed by CASA.

[c] Redpoll used only three spectra: HNCA, HNCACB, HN(CO)CACB; while all the others used all six spectra.

MARS, PISTACHIO and AutoAssign had a large number of GSs assembled and assigned in common. Since CASA and MARS accept assembled GSs as input, we fed these two methods with the GSs from the assignment of PISTACHIO and AutoAssign to compare the assignment of different methods using the same GSs. For the 111 GSs from PISTACHIO, CASA assigned all of them, with 105 GSs assigned the same as PISTACHIO; while MARS assigned 101 GSs, with 97 GSs assigned the same as PISTACHIO, and these commonly assigned 97 GSs were assigned the same by CASA. For the 107 GSs from AutoAssign, CASA assigned all of them, with 104 GSs assigned the same as AutoAssign; while MARS assigned 106 GSs with 104 GSs assigned the same as AutoAssign. All the 106 GSs assigned by MARS were assigned in common by CASA.

Discussion

Robustness against missing chemical shifts

As seen in Table 4, the assignment was nearly 100% correct when the spectra were nearly complete and 3-rung GSs under condition II were introduced. However, in reality there are always missing GSs, or overlapping GSs which cannot be separated initially. The number of GSs may be smaller than the number of assignable sites in the sequence, so a certain number of empty sites are introduced into the sequence. These empty sites break segments into shorter ones which are assigned under looser constraints from their typing and link score. There is uncertainty particularly in assigning isolated GSs or short segments (Coggins and Zhou, 2003) such as the case of malate synthase G. Certain errors may be introduced into the assignment, and as the number of empty sites increases, a reliable assignment could eventually become impossible.

For the proteins with nearly complete spectra, we simulated experimental error by randomly deleting 15% of GSs and randomly deleting certain rungs in 15% of the remaining GSs under condition II. As one sees, the assignment score using CASA was still quite high (> 87%) for all of the proteins except calmodulin, which has no C^{β} chemical shifts to provide sufficient typing constraints.

The robustness of CASA against missing data was further tested by random deletion of up to 30% of the GSs for maltose binding protein using 3-rung GSs under condition II (see Figure 6a). The more GSs deleted, the smaller was the proportion of correctly assigned residues. However, even

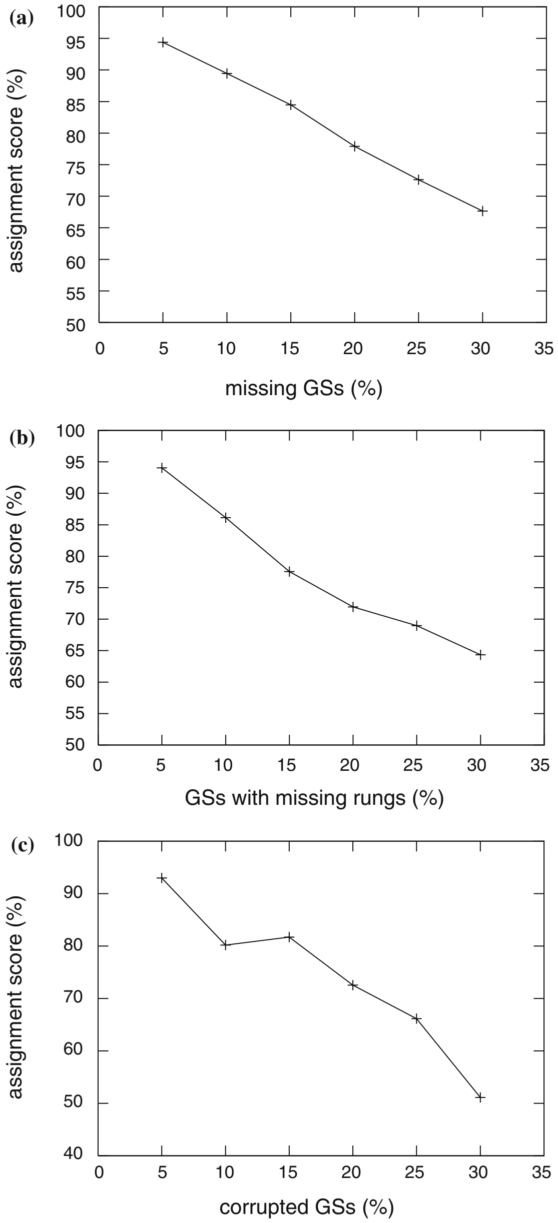

Figure 6 Robustness of CASA against (a) missing entire GSs, (b) missing rungs in GSs (1 rung in the corresponding GSs after deletion), (c) percentage of chemical shifts outside the link tolerances. All these tests were performed for maltose binding protein using 3-rung GSs under condition II.

when 30% of the GSs were deleted, the assignment score is still better than 67%. And the robustness for 3-rung GSs under smaller tolerances (condition II) is much better than that for 2-rung GSs under larger tolerances (condition I).

We also tested the robustness against missing data by random deletion of certain rungs in GSs (only C^{α} rungs were kept for those GSs). Deletion of rungs in GSs introduces ambiguity in occupation and linking for the corresponding GSs, which increases the number but decreases the sizes of segments. As a result, the size of the search tree increases greatly, and so does the execution time. It is possible that the assignment could not be finished within a reasonable time scale, so when the rungs of more than 15% of GSs were reduced, we stopped the assignment after 200 s of running time. As we can see, more than 64% of GSs were still assigned correctly by CASA within 200 s in the worst case. (see Figure 6b)

Robustness against chemical shifts outside the link tolerances

The linking information provided by inter- and intraresidue chemical shifts is an essential component of the assignment process (Jung and Zweckstetter, 2004). In CASA, we construct segments for the uniquely linked GSs and assign them by the increasing order of the number of assignable sites. The size of the ordered searching tree is greatly reduced so that the assignment is completed quickly. However, it is possible in experiment that some peaks are distorted and the corresponding chemical shifts fall outside the tolerances. Breaking the link between GSs corresponding to sequentially adjacent residues prevents those GSs from being assigned to the correct sites simultaneously, and it is even worse when these two GSs were originally located in the middle of a long segment. We tested CASA in this situation on maltose binding protein using 3-rung GSs under condition II by randomly breaking the favorable link between GSs. This may cut off the path from the root of the search tree to the leaf corresponding to correct assignment, and the assignment always stops at some dead end. However, as we can see from Figure 6 (c), a large proportion of GSs could still be assigned correctly even when 30% of the links were broken.

Robustness against extra peaks

The robustness should also be tested when the number of GSs provided is larger than that of the assignable sites. The extra GSs may be due to noise, experimental artifacts, or erroneous peaks when looser constraints are applied in peak picking and/or peak aligning, so they may have little relevancy to the sequence. Here the extra peaks were introduced by combining the peaks of the test protein (maltose binding protein) with those from another protein, namely, ubiqiutin, thiopurine methyltransferase, and human carbonic anhydrase I, respectively (see Table 8); while the sequence of the test protein was unchanged. The robustness of CASA was then tested under condition II. Since the extra GSs do not overlap with the original GSs too much, most of the segments constructed from the original GSs were conserved and assigned onto the correct sites, which resulted in high and stable assignment scores (~85% for all cases) regardless of the percentage of extra GSs. The extra segments (constructed from the extra GSs) were not com-

Table 8. Robustness of CASA tested on extra peaks

Test protein	Maltose binding protein		
Number of residues of test protein	370		
Number of GSs from test protein	328		
Protein providing extra peaks	Ubiquitin	Thiopurine methyltransferase	Human carbonic anhydrase I
Number of extra GSs	70	189	242
% of extra GSs[a]	17	36	42
completion time (s)[b]	60	96	120
Number of GSs assigned correctly	280	280	284

[a] The percentage of extra GSs in the total number of GSs.

[b] The completion time of assignment using 3-rung GSs under condition II.

petitive, because it was hard to map the extra segments onto the sequence of the test protein.

Multiple conformers

Sometimes the resonances for some segment in the sequence may be duplicated because the protein has two slowly interconverting conformations. We tested the robustness of CASA for multiple conformers by generating a shifted duplication of the peaks of a segment of 20 residues starting from residue 38 of protein thiopurine methyltransferase. That is, the chemical shifts of the H nuclei in the duplicated peaks are 0.05 ppm larger than those in the original peaks, and the chemical shifts of all the other nuclei are 0.5 ppm larger than those in the original peaks. As a result these duplicated peaks still have favorable occupation scores at the corresponding sites and favorable link scores to the GSs before and after the segment. CASA could generate two alternative assignments. For each of these two assignments, the unduplicated segments and one of the duplicated segments were assigned correctly, and the other duplicated segment was not assigned because the corresponding sites were already occupied. By checking the assignable sites of the unassigned segment and comparing this segment to those assigned at these sites, the multiple conformers can be easily identified.

Typing of chemical shifts

The complexity of sequence specific assignment is effectively determined by the constraints from the typing and linking of GSs. Thus, a tighter constraint from typing of GSs is very desirable. However, the value of a chemical shift is determined not only by the residue type, but also by the secondary structure the residue is involved in and the types of neighboring residues, which enlarges the dispersion region of the corresponding chemical shifts and results in big overlap among different residues.

AutoAssign uses a probability score based on the statistical mean and standard deviation from the BMRB database. The Gaussian-like typing score of assigning a chemical shift p to a residue of type t is

$$Occu\,(p,t) \propto \exp(-(p-\bar{p}_t)^2/\sigma_t^2)$$

where $\bar{p}_t$ and σ_t are the statistical mean and standard deviation of the chemical shifts of the residue of type t. However, the mean chemical shifts of most residues are very close to each other, and it is highly probable that the chemical shift of a certain residue is far away from the mean chemical shift in a specific experiment. For example, there are 15 amino acids whose mean chemical shifts of C^{α} are between 50 and 60 ppm, and the corresponding standard deviations are typically 2–3 ppm. If the C^{α} chemical shift of an alanine is 2 ppm (1 standard deviation) larger than its average, it has a higher score to be assigned to a leucine residue. The meaning of the size of a Gaussian score is thus dubious, and we elected to use the binary form described above.

Based on the distribution of C^{α} and C^{β} chemical shifts, TATAPRO classified the 20 amino acids into eight groups. Instead of being assigned a probabilistic typing score, each GS is assigned deterministically to one of the eight groups, which reduces the ambiguity of typing. However, new ambiguity is introduced when mapping segments into sequence because the original 20-letter sequence is converted into an 8-letter sequence. It is less likely to map a segment uniquely onto the sequence for larger proteins. Furthermore, the grouping of residues strongly demands complete C^{β} chemical shifts (for all the proteins tested by TATAPRO, C^{β} chemical shifts are nearly 100% complete), which limits its general application.

It is very attractive that MARS introduced an improved typing score by considering corrections due to secondary structure and neighboring residue effects. A tighter constraint with general application could be achieved using this updated typing score. It requires a highly accurate prediction of chemical shifts based on the prediction of secondary structure. However, we have seen no differences in the assignment scores from CASA and MARS on the protein sets provided. For now the programs appear competitive, with an advantage for CASA in computational efficiency.

Link tolerances and number of rungs for linking

Constraints from link scores between GSs also play an important role in reducing the ambiguity of assignment, which determines the depth of the search tree as well as the number of branches in each level. In general, this constraint is determined by the size of the tolerances and the number of rungs in GSs to evaluate the link score. Small tolerances plus a large number of rungs for linking can reduce the ambiguity substantially, resulting in fast and reliable assignment. However, the tolerances should be set slightly larger than the uncertainty in peak positions based on the digital resolution of the NMR spectra, since using too narrow tolerances may break some links between sequentially adjacent GSs. Furthermore, the number of triple resonance experiments could be limited due to some practical reason such as the stability of the protein, so it is necessary for an automated program to be able to use reduced numbers of rungs for linking. Thus there is a tradeoff between the theoretical requirement and experimental feasibility. In our experience, it is enough to use 2-rung GSs in condition I for small proteins with less than 150 residues. However, for large proteins or proteins with severe degeneracy, 3-rung linking is required and tighter tolerances (condition II) are desired to guarantee rapid and reliable assignment. It is mentioned in some methods (Atreya et al., 2000; Coggins and Zhou, 2003) that they use tighter tolerances at the beginning of assignment and enlarge the tolerances when a large proportion of GSs have been assigned. However, it is possible that the chemical shifts of sequentially adjacent GSs are outside the tight tolerances due to peak distortion, while at the same time they may be linked favorably to other GSs due to spectral degeneracy. Thus the incorrect links are formed while the correct links are ruled out, and the partially assigned result may be a dead end with errors already introduced.

Conclusion

We have introduced a new algorithm, CASA, for the automatic assignment of the NMR spectra of proteins. It follows the same general paradigm used in other methods of combining peaks from different spectra into GSs, linking these together into segments, and mapping the segments onto the amino acid sequence. CASA performs at least as reliably as any method, and much more reliably than some, both on experimental data and on simulated data with added errors. CASA runs extremely fast, taking only minutes of CPU time for even a protein with 723 residues. For incomplete data, high chemical shift degeneracy, or very large proteins, CASA may produce some wrong assignments. We further introduced a new assignment ambiguity score by which the confidence in an assignment can be assessed and the number of wrong assignments can be reduced. We observed substantial differences among the assignments produced by different methods starting with the same data, especially experimental data. This state of affairs is surprising, and calls for further improvements in all assignment algorithms. The greatest source of these disagreements appears to lie in the formation of GSs, rather than the linking or mapping stages. Our method is more flexible than most, accepting data from many different NMR experiments, constructing GSs with a variety of different rungs under the tolerances suitable to the quality of the data. Not only can CASA use peak lists to assemble GSs automatically, but it can also use GSs that are constructed either manually or by other methods. CASA is relatively robust upon deletion or addition of extraneous data. With the extreme speed of CASA, one can consider exploring many assignments generated from the same spectra under different peak-picking conditions.

The results so far indicate that future improvements may arise from better assembly of GSs and their subsequent linking. Closer attention to the original spectra would enable the use of peak shape information to reduce ambiguity. Work in these directions and the construction of a user-friendly GUI for CASA are in progress. The software in its current raw form (version 1.0, in MOE SVL language) is available from G.M. Crippen.

Acknowledgments

This work was supported in part by a grant from the University of Michigan Bioinformatics Program and the Howard Hughes Medical Institute (J. Wang and G.M. Crippen) and NSF MCB 0135330 (T. Wang and E.R.P. Zuiderweg).

References

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

Bailey-Kellogg, C., Widge, A., Kelley, III J.J., Berardi, M.J., Bushweller, J.H. and Donald B.R (2000a) *The 4th Int'l Conf. On Computational Molecular Biology (RECOMB)*, 33–44.

Bailey-Kellogg, C., Widge, A., Kelley, J.J. III, Berardi, M.J., Bushweller, J.H. and Donald, B.R. (2000b) *J. Comp. Biol.*, **7**, 537–558.

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

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

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

Bhattacharyya, S., Habibi-Nazhad, B., Amegbey, G., Slupsky, C.M., Yee, A., Arrowsmith, C. and Wishart, D.S. (2002) *Biochemistry*, **41**, 4760–4770.

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

Coggins, B.E. and Zhou, P. (2003) *J. Biomol. NMR*, **26**, 93–111.

Delaglio, F., Grzesiek, S., Vuister, G.W., Zhu, G., Pfeifer, J. and Bax, A. (1995) *J. Biomol. NMR*, **6**, 277–293.

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.

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

Eghbalnia, H.R., Bahrami, A., Wang, L., Assadi, A. and Markley, J.L. (2005) *J. Biomol. NMR*, **32**, 219–233.

Farmer, B.T. II and Venters, R.A. (1996) *J. Biomol. NMR*, **7**, 59–71.

Guntert, P., Saltzmann, M., Braun, D. and Wüthrich, K. (2000) *J. Biomol. NMR*, **18**, 129–137.

Hare, B.J. and Prestegard, J.H. (1994) *J. Biomol. NMR*, **4**, 35–46.

Hitchens, T.K., Lukin, J.A., Zhan, Y., McCallum, S.A. and Rule, G.S. (2003) *J. Biomol. NMR*, **25**, 1–9.

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

Jung, Y.S. and Zweckstetter, M. (2004) *J. Biomol. NMR*, **30**, 11–23.

Kraulis, P.J. (1994) *J. Mol. Biol.*, **243**, 696–718.

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. (1997) *J. Chem. Inf. Comput. Sci.*, **37**, 467–477.

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

Molecular Operating Environment (MOE), Chemical Computing Group, Inc., http://www.chemcomp.com.

Montelione, G.T. and Wagner, G. (1990) *J. Magn. Reson.*, **83**, 183–188.

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

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

Moseley, H.N.B., Monleon, D. and Montelione, G.T. (2001) *Meth. in Enzymol.*, **339**, 91–108.

Moseley, H.N.B., Sahota, G. and Montelione, G.T. (2004) *J. Biomol. NMR*, **28**, 341–355.

Moy, F.J., Seddon, A.P., Campbell, E.B., Bohlen, P. and Powers, R. (1995) *J. Biomol. NMR*, **6**, 245–254.

Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J. (1991) *J. Biomol. NMR*, **1**, 217–236.

Schubert, M., Oschkinat, H. and Schmieder, P. (2001a) *J. Magn. Reson.*, **148**, 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.

Shimotakahara, S., Rios, C.B., Laity, J.H., Zimmerman, D.E., Scheraga, H.A. and Montelione, G.T. (1997) *Biochemistry*, **36**, 6915–6929.

Tashiro, M., Rios, C.B. and Montelione, G.T. (1995) *J. Biomol. NMR*, **6**, 211–216.

Tashiro, M., Tejero, R., Zimmerman, D.E., Celda, B., Nilsson, B. and Montelione, G.T. (1997) *J. Biomol. NMR*, **272**, 573–590.

Van Doren, S.R., Kurochkin, A.V., Ye, Q., Johnson, L.L., Hupe, D.J. and Zuiderweg, E.R.P. (1993) *Biochemistry*, **32**, 13109–13122.

Venters, R.A., Farmer, B.T., Fierke, C.A. and Spicer, L.D. (1996) *J. Mol. Biol.*, **264**, 1101–1116.

Wagner, G. and Wüthrich, K. (1982) *J. Mol. Biol.*, **155**, 347–366.

Wehrens, R., Lucasius, C., Buydens, L. and Kateman, G. (1993) *J. Chem. Inf. Comput. Sci.*, **33**, 245–251.

Wuthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, John Wiley and Sons, New York, NY.

Zimmerman, D.E., Kulikowski, C.A., Huang, Y., Feng, W., Tashiro, M., Shimotakahara, S., Chien, C.Y., Powers, R. and Montelione, G.T. (1997) *J. Mol. Biol.*, **269**, 592–610.

Zimmerman, D.E., Kulikowski, C.A. and Montelione, G.T. (1993) *Proc. First Intl Conf. Intel. Sys. Mol. Biol.*, **1**, 447–455.

Zimmerman, D.E., Kulikowski, C.A., Wang, L.L., Lyons, B. and Montelione, G.T. (1994) *J. Biomol. NMR*, **4**, 241–256.