

SAGA: rapid automatic mainchain NMR assignment for large proteins

Gordon M. Crippen · Aikaterini Rousaki ·
Matthew Revington · Yongbo Zhang ·
Erik R. P. Zuiderweg

Received: 7 December 2009 / Accepted: 23 February 2010 / Published online: 16 March 2010
© Springer Science+Business Media B.V. 2010

Abstract Here we describe a new algorithm for automatically determining the mainchain sequential assignment of NMR spectra for proteins. Using only the customary triple resonance experiments, assignments can be quickly found for not only small proteins having rather complete data, but also for large proteins, even when only half the residues can be assigned. The result of the calculation is not the single best assignment according to some criterion, but rather a large number of satisfactory assignments that are summarized in such a way as to help the user identify portions of the sequence that are assigned with confidence,

vs. other portions where the assignment has some correlated alternatives. Thus very imperfect initial data can be used to suggest future experiments.

Keywords Automatic assignment · Generic spin system · Triple resonance · Large proteins

Electronic supplementary material The online version of this article (doi:10.1007/s10858-010-9403-2) contains supplementary material, which is available to authorized users.

G. M. Crippen (✉)
College of Pharmacy, University of Michigan,
Ann Arbor, MI 48109, USA
e-mail: gcrippen@umich.edu

A. Rousaki · M. Revington · Y. Zhang
LSA Biophysics, University of Michigan,
Ann Arbor, MI 48109, USA

Present Address:
M. Revington
Department of Chemistry and Biochemistry,
University of Windsor, Windsor, ON N9B 3P4, Canada

Present Address:
Y. Zhang
Department of Biochemistry, Molecular Biology
and Cell Biology, Northwestern University, Evanston,
IL 60208, USA

E. R. P. Zuiderweg (✉)
Department of Biological Chemistry, University of Michigan
Medical School, Ann Arbor, MI 48109, USA
e-mail: zuiderwe@umich.edu

Introduction

Presently, the backbone resonances of virtually all proteins in solution are assigned using a combination of 3D triple resonance NMR spectra pairs HNCA/HN(CO)CA, HNCA CB/HN(CO)CACB, HN(CA)CO/HNCO (for perdeuterated proteins) or HN(CA)HA/HN(COCA)HA for small proteated proteins (e.g., Cavanagh et al. 2007). Similar combinations of experiments have been devised for the assignments of the backbone resonances of proteins in the solid state (e.g., Baldus 2007). Smaller or unfolded proteins in solution are increasingly assigned using higher dimensional data with similar or extended coherence pathways, typically obtained by using projection methods on three frequency coordinates (e.g., Atreya and Szyperski 2004).

Here, we present an automated assignment algorithm for a combination of the “classical” 3D triple resonance experiments from proteins in solution. We focus on these 3D experiments because their relative simplicity allows the highest sensitivity per unit time, which is important for the assignment of the larger proteins we study in our laboratory (>300 residues), which are typically soluble at 300 μ M concentration or less.

In theory, the 3D experiments provide more than sufficient resolution to assign the spectra of such large proteins: all $^{13}\text{C}\alpha$ resonance frequencies of a 50 residue protein are likely to be unique within an achievable 0.1 ppm precision

(see Fig. S1); similar values hold for $^{13}\text{C}\beta$ frequencies (within an achievable 0.2 ppm precision) and ^{13}CO frequencies (within an achievable 0.1 ppm precision). Hence, if the sequence specific shifts of the $\text{C}\alpha$, $\text{C}\beta$ and CO resonance positions may be assumed to be uncorrelated, a combination of an HNCA/HN(CO)CA pair with HN(CA)CO/HNCO pair should suffice to assign a $50 \times 50 = 2500$ residue protein. Combined with $\text{C}\beta$ connectivity information from an HNCACB/HN(CO)CACB pair, the size limit increases further, and when the latter is combined with chemical shift statistics which allows the distinction of 6 different amino acid groups (A; G; P; S+T; D+F+I+L+N+Y; E+C+H+K+M+Q+R+V+W, see Fig. S2), the size limit of assignments for proteins with just triple resonance methods seems much larger than any protein that will likely be ever studied by solution NMR methods.

While the information content of such spectra is sufficient, manual assignment of the spectra to the known amino acid sequence is extremely tedious for large proteins, and it has been hard to automate, especially for spectra of large proteins with incomplete data. Many computer algorithms have been devised (Friedrichs et al. 1994; Hare and Prestegard 1994; Hyberts and Wagner 2003; Leutner et al. 1998; Lukin et al. 1997; Meadows et al. 1994; Moseley et al. 2001; Olson and Markley 1994; Oschkinat and Croft 1994; Stratmann et al. 2010; Zimmerman et al. 1993, 1994, 1997; Zimmerman and Montelione 1995). See also recent reviews (Billeter et al. 2008; Williamson and Craven 2009; Güntert 2009).

Here we adopt the nomenclature of the AutoAssign program (Zimmerman et al. 1997; Moseley et al. 2001) to explain in broad terms common and distinguishing features. The general paradigm has been to first assemble peaks from the triple resonance experiments, and group these into generic spin systems (GS) that specify the chemical shifts of various nuclei of a particular residue and the sequentially previous residue. Next, the GSs are arranged into sequentially ordered groups (segments), where there is a unique match in chemical shifts between successive GSs in a segment. Finally, the segments are placed on the sequence in such a way as to optimize some scoring function based on amino-acid residue chemical shift statistics, such as shown in Fig. S2. A common theme is that the objective is the single best assignment according to some criteria, such as number of residues assigned and accuracy of agreement between chemical shifts in a GS and the sequence position to which it is assigned. Thus many different optimization procedures are employed by the various assignment programs to find a relatively good solution. The other common theme is to group GSs into segments before attempting to place them on the sequence. The motivation is that any individual GS may be quite compatible with dozens of sequence positions, whereas a

segment of several GSs may have only a few options, thus drastically reducing the combinatorial search.

Without making a comprehensive review of automatic sequential assignment methods, it is worth noting recent activity in this lively field. The CRAACK program (Benod et al. 2006) takes the usual input data and forms GSs, but then its objective is to assign them to residue types and secondary structure states, rather than assigning them to the sequence. Other methods aim to assign GSs to the sequence but employ additional inputs to the problem. For example, ABACUS (Lemak et al. 2008) uses the usual GSs plus NOESY cross peaks to help determine sequence separations in the assignment. NOE-net (Stratmann et al. 2010) uses largely NOEs plus residual dipolar couplings and chemical shifts. Xiong et al. (2008) start with a model of the three-dimensional structure of the protein, perhaps derived from homology modeling, plus NMR data from HSQC, TOCSY, and NOESY experiments. From this they deduce rough residue types and choose from the many possible NOESY cross peaks those that are consistent with the residue types and interresidue contacts in the model.

Another issue is the assessment of reliability of the one or more assignments produced by a method. In ABACUS (Lemak et al. 2008) the Monte Carlo search for assignments gives the probability of assigning a particular segment of GSs to a sequence position. The IDA method (Lin et al. 2006) gives exactly one assignment for a given set of GSs plus a score reflecting how good the assignment is. Gaussian perturbations of the rungs (chemical shifts) of the GSs produce different assignments and different scores, so an alternative assignment can be associated with a Z-score relative to the assignment from unperturbed GSs.

Another question is how best to structure the assignment search. Our earlier program, CASA (Wang et al. 2005), follows the standard organization of peaks into GSs into segments, but the subsequent branch-and-bound search for assignments was carefully structured for speed and robustness. In PISTACHIO (Eghbalnia et al. 2005) GSs are associated with tripeptides so that the assignment amounts to assigning GSs to mutually consistent overlapping tripeptides in the given sequence. In GASA (Wan and Lin 2007) the usual sequence of steps (peaks to GSs, GSs to segments, segments to sequence positions) are overlapped in order to improve performance. For example, the best way to link GSs together into a segment is judged by a score for how well such a segment might fit onto the sequence. The usual logic for joining GS_i to GS_j to form a segment GS_iGS_j is that (1) the two sets of rungs agree well enough in terms of numbers of rungs involved and similarity of corresponding chemical shifts, and (2) there is no other GS_k that would equally well form segments GS_iGS_k or GS_kGS_j . Vitek et al. (2005) argue that even though such a unique link seems too good to be coincidental, one can

never exclude the possibility that the GS that really follows GS_i in sequence is simply missing from the input data. Hence they structure the assignment search by breaking the whole sequence up into smaller non-overlapping windows and then considering what alternative assignments of GSs suit each window. Deeper in the branch-and-bound search, two sequentially adjacent windows having relatively few alternative assignments can be merged into a larger window, and so on.

Methods

Here we present a new automatic backbone assignment procedure, SAGA (sequential assignment of GSs algorithm). The inputs are the amino acid sequence of the single polypeptide chain and at least some of the customary triple resonance peaks. Additional information is not used, such as NOEs, secondary structure, or homology models. As explained in detail below, the first step is to assemble the spectral data into GSs, as do most assignment programs. However, unlike most methods and our earlier work (Wang et al. 2005), the GSs are not subsequently assembled into apparently unambiguous segments, as suggested by Vitek et al. (2005). The second step is to assign the GSs to the sequence. Most methods define a quantitative measure of quality for an assignment and then use stochastic global optimization methods such as Monte Carlo or simulated annealing to search for the single best quality assignment. The second unusual feature of SAGA is that the user instead defines criteria that an acceptable assignment satisfies, and the output consists of a relatively broad sampling of perhaps very many acceptable assignments. In this way, we hope to learn what parts of the assignment are known with high confidence, what parts may be the result of slowly interconverting conformations, and what features suggest the need for scrutiny of the input data or additional experiments. In particular, one of the acceptability criteria is a lower bound on the fraction of the GSs that are assigned to residues. This allows one to focus only on rather complete assignments in the case of high-quality data, or to see what parts of the assignment have been well established at an early stage of a study where, e.g., only half the residues can be assigned. The third unusual feature of SAGA is that more than one algorithm is provided for searching for acceptable assignments. While most methods center around a particular search or optimization algorithm that works well for at least some test cases, we find that quite different algorithms are needed for thorough searches given a small protein and high-quality peak data, as opposed to broad samplings of assignments given a large protein with many peaks lost in the noise. The three search algorithms presented below are based on well-established

general combinatorial methods, such as clique finding, branch and bound tree searching, and greedy searches. However, tailoring these general approaches to the backbone sequential assignment problem constitutes the fourth unusual feature of SAGA, and this accounts for SAGA's pleasing performance, even on challenging data.

Peaks to GSs

The first step is to assemble the peaks gleaned from several triple-resonance experiments into data structures called generic spin systems (GS) (Zimmerman et al. 1997). A GS consists of the chemical shifts of several atoms in some unspecified amino acid residue i (intra-residue) and the sequentially previous residue $i-1$ (sequential). These always include the intra-residue HN root, H_i^N and N_i , and optionally various rungs on the i and $i-1$ sides for $C\alpha$, CO, and $C\beta$ atoms.

One way to input peak information is via NMRPipe-format (Delaglio et al. 1995) peak-pick lists derived from corresponding pairs of intra-residue and sequential experiments: HNCA and HN(CO)CA for $C\alpha$, HNCACB and HN(CO)CACB for $C\beta$, and HN(CA)CO and HNCO for CO, respectively. We require the first pair of experiments so that the GSs all have at least an intra-residue or sequential $C\alpha$ rung. For each available pair of experiments, allowance is made for a possible uniform additive shift between the two experiments of no more than 0.1 ppm in H^N and 0.75 ppm in N. Pairs of peaks are considered to match if their (shifted) H^N and N chemical shifts agree within a given tolerance while the third chemical shifts (e.g. $C\alpha$) differ by at least 0.25 ppm. Tolerances in H^N and N are varied from 0.005 to 0.03 and from 0.05 to 0.3 ppm, respectively, until the maximum number of uniquely matched pairs of peaks is achieved. In this way, each pair of available peak lists produces a set of so-called Ts, which are GSs having an HN root and one or both rungs for the intra-residue and sequential chemical shifts of either $C\alpha$, $C\beta$, or CO.

Exploring the same overall shift range and tolerance ranges used for constructing Ts, the $C\alpha$ Ts are combined with the available $C\beta$ and then CO Ts, choosing the shifts and tolerances that give the maximal number of unambiguous matches. If a resulting GS has no rungs on either the intra-residue or sequential side, it is still accepted.

An alternative format for peak information is Sparky (Goddard and Kneller 2003), where one Sparky file substitutes for one of the three pairs of NMRPipe files because the intra-residue and sequential pairings are already specified by the user through the use of matching tag pairs in the file. However, the resulting Ts need not always have a matching i and $i-1$ rung, due to experimental noise considerations. Missing rungs are denoted by entries larger than 500 ppm. This input style is especially suited for noisy

and incomplete data from large proteins, where an investigator needs to curate peak-pick files by hand. The HN root chemical shifts are taken to be those of the intra-residue peak when both peaks are available. Associated with each T is the common tag specified by the user. Using Sparky files instead of peak pick lists affords the spectroscopist greater control over building GSs. Ts from more than one Sparky file are joined together into a GS by the program by matching up their tags, where there must be a C α T, and the resulting GS has its HN root. That is, in the Sparky input mode, the ^{15}N – ^1H frequencies are not used to construct GSs, which relieves some of the criteria on the quality of the input NMR data. In this input mode, the investigator can also easily assess the precision of the peak matching. The Sparky program allows hand-adjustment of the center of the peak-pick positions. After these optimizations are carried out, we find that in properly zero-filled data, the center of $^{13}\text{C}\alpha$ peaks can be defined with a 0.1 ppm, $^{13}\text{C}\beta$ with a 0.2 ppm and ^{13}CO with a 0.1 ppm precision, even for data sets of large proteins. The ^{15}N – ^1H frequencies data are taken from the C α T and are kept by the program for amino acid identification purposes (significant for the ^{15}N frequencies of glycyls only), and are returned at the output stage.

If the GSs are constructed from NMRPipe files, they don't have the unique identifier tags seen in GSs built from Sparky files. As such tags are helpful later in communicating the results to the user, the program automatically constructs tags for any GSs that lack them.

GSs to matching graph

Only after the GSs are prepared is the sequence considered at all. This is read from a FASTA format file, where in addition to the standard 20 single character symbols for residue types, B is recognized as either D or N, and Z is recognized as either E or Q. Each GS may occupy residue i in the sequence if its intra-residue rungs are in agreement with the chemical shifts expected for the type of that residue, and if its sequential rungs agree with the type of residue $i-1$ in the sequence. A GS that fits nowhere in the sequence is deleted, but most are compatible with multiple sequence positions. In any case, compatibility of a GS with a residue is treated as a qualitative yes/no decision, rather than assigning a quantitative compatibility score.

As in our previous algorithm (Wang et al. 2005), GS occupancy is based on chemical shift statistics from the BioMagResBank (Seavey et al. 1991; <http://www.bmrb.wisc.edu>). For each atom type in each residue type they provide the chemical shift a = lower bound, b = upper bound, μ = mean, and σ = standard deviation. Then a chemical shift from a GS rung is in agreement with the corresponding data bank statistics if it falls in the

interval = $[\max(a, \mu-s\sigma), \min(b, \mu+s\sigma)]$, where by default $s = 4.0$ except $s = 5.5$ for H because of the larger dispersion. The user of SAGA can optionally choose other values for s (see Fig. S2). The data bank values of σ generally are much smaller than the standard deviation for a uniform distribution over the interval $[a,b]$, so accepting chemical shifts that are four standard deviations from the mean is not so permissive as it might seem. In our tests, however, we used $s = 6$ for ^1HN , $s = 6$ for ^{15}N , $s = 4$ for $^{13}\text{C}\alpha$, $s = 2.5$ for $^{13}\text{C}\beta$ and $s = 3$ for ^{13}CO shift ranges, based on our own experience with assigning spectra of large proteins.

Each GS is tested for possible occupancy at each assignable residue position, excluding the N-terminus and prolines, of course. The chemical shifts of the N and all available intra-residue and sequential C α , C β , and CO rungs must fall within the corresponding allowed intervals. The amide proton chemical shift is not considered in determining occupancy. As a special case, if a GS has an intra-residue C α chemical shift less than 50 ppm and no intra-residue C β rung, then it must occupy a glycyl residue, as long as the other sequential rungs match satisfactorily. Another special case is a GS having an intra-residue C β rung less than 20 ppm, in which case it must occupy an alanyl residue where the sequential rungs match.

SAGA allows the user to optionally include further constraints on occupancy derived from other experiments. Each constraint specifies a GS by giving its tag and HN root chemical shifts. The constraint requires it to occupy certain residue types and/or certain sequence positions. Ordinarily this is used to narrow down the list of possible occupancies already determined, but in exceptional circumstances, the constraints can completely override them.

Once the possible occupancies of all GSs have been established, the links between GSs must be determined. That is, for each GS $_j$, can GS $_k$ immediately follow it in the sequence? One necessary condition is that there is at least one sequence position i that GS $_j$ can occupy, and GS $_k$ can also occupy position $i + 1$. The other necessary condition is that all available intra-residue C α , C β , and CO rungs of GS $_j$ must match the corresponding available sequential rungs of GS $_k$ within default tolerances of 0.1, 0.2, and 0.1 ppm, respectively. The user may optionally choose different tolerances. In the vacuous case of no corresponding rungs between the two GSs, the link is still considered possible.

Matching graph

One way to visualize the sequence, GSs, occupancies of GSs, and links between GSs is as a bipartite graph of GS vertices vs. residue vertices. There may be more or fewer GSs than assignable residues, depending on factors such as

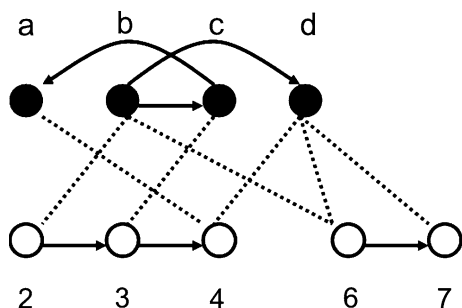


Fig. 1 A simple bipartite assignment graph. *Open circles* are assignable residues with corresponding sequence positions, and *arrows* between them indicate sequentially adjacent residues. *Filled circles* are GSs with links between them shown as *arrows*. Edges between GSs and assignable residues are *dashed lines*

the noise floor for detecting peaks, peaks rendered unobservable due to conformational exchange broadening, and possible multiple conformations in slow exchange. Each assignable residue may be occupied by zero or more different GSs, each GS may occupy at least one residue, and each GS may be sequentially linked to zero or more other GSs. Before describing the assignment algorithms in general, consider a very simple example shown in Fig. 1. Suppose the polypeptide chain consists of only seven residues, sequence being ADREPLE, so that five residues are assignable, $i = 2, 3, 4, 6,$ and 7 . Suppose there are four GSs having tags $a, b, c,$ and d , where a and c have only a single possible occupancy, b may occupy residues 2 or 6 , and d may occupy residues $4, 6,$ or 7 . As for linking, b may be followed by c or d , but none follows a . Converting the initial bipartite graph to an assignment amounts to removing some of the GS to residue edges until each GS vertex and each residue vertex has zero or one edge. If two GSs are thus assigned to sequentially adjacent residues, there must be a corresponding link from the first to the second GS. As a shorthand notation for assignments, write five characters for the five assignable residues, where “#” means “unassigned” and comes first in the alphabet. Then there are 35 possible assignments in this simple example, ranging from ##### to bcad#, of which five are maximal in the sense that no additional residue can be assigned. The maximal assignments are ##db#, b#d##, bcad#, bca#d, and #cabd, the last three being shown in Fig. 2.

Clique algorithm

There is a very long history of algorithms for bipartite matching, where the initial graph looks like Fig. 1 with only the GS vertices and residue vertices joined by dashed edges between a GS and a residue, but without the directed edge arrows between residue vertices for sequential

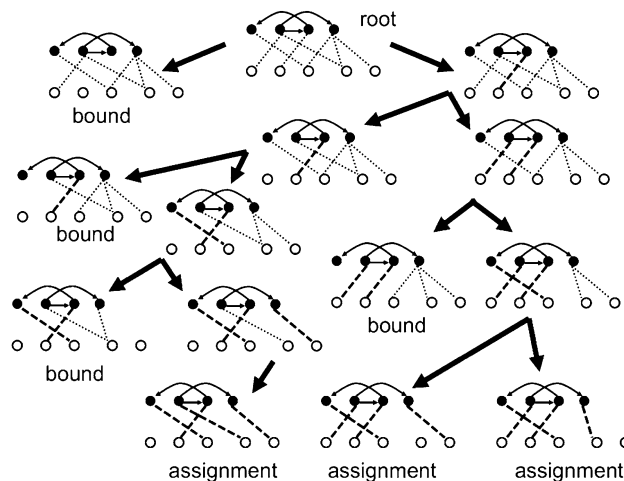


Fig. 2 An example branch-and-bound binary search tree starting with the bipartite graph in Fig. 1. Each branch involves either adding some GS to residue assignment, *highlighted as a bold dashed line*, or eliminating that edge from the bipartite graph. A bounding condition requiring that all GSs eventually be used eliminated from further consideration those tree vertices marked “bound”

adjacency or the directed edges between GS vertices for link compatibility. The standard bipartite matching algorithms try to find the largest subset of the dashed edges such that no more than one edge connects to each vertex, or given weights associated with each such edge, they try to maximize the sum of the weights in the final matching graph (Kuhn 1955; Hopcroft and Karp 1973; Kao et al. 2001). Our sequential assignment problem is qualitatively different because of the link compatibility feature between GSs, which excludes certain *pairs* of dashed edges in the matching graph. This problem was treated in the equivalent context of calculating ways to dock a small molecule (represented as one set of vertices) to a binding site (represented as another set of vertices), where the edges correspond to energetically favorable ligand-receptor interactions, but geometric constraints exclude certain pairs of edges from the solutions (Kuhl et al. 1984). In that work, it was shown that for arbitrary pair exclusions, the problem in general mapped to an NP-complete one, so that any algorithm would require exponentially increasing computer time as the problem’s size increased. However, for the assignment problem the situation is by no means hopeless. First, if even a non-polynomial algorithm runs fast enough for practical application to the largest proteins NMR can handle (around 800 residues), then that is adequate, and we need not worry about the algorithm’s behavior in the limit of infinite proteins. Second, in the assignment problem the pair exclusions are by no means arbitrary, so for that restricted class of problems there may be a polynomial time algorithm.

As in Kuhl et al. (1984), the matching graph, such as in Fig. 1, can be transformed into an assignment graph, where every vertex in the assignment graph represents an edge of the matching graph, and there are edges between vertices of the assignment graph whenever the two vertices are mutually compatible. In this way, we can encode the constraint that each GS and each residue is used at most once, and GSs assigned to sequentially adjacent residues must be link compatible. Then an assignment of the GSs to the sequence corresponds to a clique of the assignment graph, where a clique is defined to be a maximal, completely connected subgraph.

More specifically, the transformation of the matching graph into the assignment graph is quite straightforward, but the assignment graph tends to be rather large. If GS_j can occupy m_j residues in the sequence, the assignment graph has $\sum_j m_j = k$ vertices and has nearly $k(k-1)/2$ edges because assigning one GS to a particular residue is compatible with most other assignments of another GS to its possible residues. There are a number of different general clique finding algorithms available, and we used that of Bron and Kerbosch (1973), which performs well for large graphs. Without going into the intricate details of the algorithm, we did not attempt to find all cliques, as that would include many small ones, corresponding to few assigned GSs. Instead, we used the option to guide the algorithm's branch-and-bound search for cliques toward those that would be at least as large as the largest so far found, and in the end we retain only the ten largest cliques. Converting a clique into an assignment involves nothing more than assigning the GS of each vertex to the residue for that vertex, and then automatically each GS and each residue is used at most once, and sequentially adjacent assigned residues are paired with link-compatible GSs.

Greedy assignment algorithm

One standard paradigm for solving combinatorial problems is the so-called greedy approach, where the best available choice is made at each stage according to some criterion regardless of the ultimate consequences, and the choice at an earlier stage is never revised. Obviously there is no guarantee that the best solution will be found, but it is a fast and reliable way of finding a fairly good one, even in the face of a vast number of possibilities. This general approach has long been applied to many different combinatorial or discrete optimization problems, so it is discussed in many standard textbooks (Parker and Rardin 1988). To apply the greedy approach to a particular class of problems, one must specify what constitutes a choice and how to measure the quality of a choice. How well it performs on that class of problems depends critically on these two definitions.

For the sequential assignment problem, the starting point amounts to a bipartite graph such as Fig. 1, where there are many edges (shown as dashed lines) between GS vertices and the residue vertices they *may* occupy. Even if a GS has only one such edge to a residue that has no alternative edges, the final assignment might not involve that edge. The definition of a choice in our tailored greedy algorithm amounts to selecting one such edge, marking it as “required” (shown as a bold dashed line in Fig. 2), and then simplifying the graph by deleting any alternative edges from the GS and its required assigned residue. In Fig. 2, different choices are illustrated as the right-hand branches in the tree.

The crucial second definition is a measure of how good a choice is. A good choice amounts to assigning some GS to a residue where the GS to residue compatibility is relatively certain, and where there are adjacent residues which have been or could be assigned to other GSs that link together relatively certainly to form a contiguous assigned segment of the chain. Note that choices involving a residue near a proline or near a chain terminus tend to be worse than for a residue in the middle of a five-residue assignable segment. As a measure of the certainty of assigning GS i to residue r , we use w_i = the total number of intra-residue and sequential rungs for GS_i , all of which must match the types of residues r and $r-1$ for any allowed occupancy of that GS. As a measure of the certainty a link from GS_i to GS_j , we use the number of matching rungs available, $v_{i,j}$. Then the total score $s(i,r)$ for choosing GS i and residue r looks at the best combination of link-compatible GSs that may occupy the previous and subsequent two residues, if available. In order to favor a choice that extends a segment having some already assigned residues, let $x_{j,r+1} = 2$ if the GS_j to residue $r + 1$ edge is required, or 0 otherwise.

$$s(i,r) = w_i + \max_{\substack{j,r+1 \\ k,r+2 \\ m,r-1 \\ n,r-2}} (w_j + v_{i,j} + x_{j,r+1} + w_k + v_{j,k} + x_{k,r+2} + w_m + v_{m,i} + x_{m,r-1} + w_n + v_{n,m} + x_{n,r-2})$$

The greedy assignment algorithm is then very simple. (1) Start with the initial bipartite graph, where no edges are required. (2) Calculate the score for all edges that are not required. (3) Generally, more than one edge has the maximum score. Choose at random one of these. (4) Mark the chosen edge as required and remove any other edges from that GS and that residue. (5) If all edges are required, an assignment has been found. Otherwise, return to step 2.

As a simple example, consider the bipartite graph in Fig. 1. Suppose $w_i = 1$ for all GSs, and all links are equally strong, $v_{i,j} = 1$. Then $s(b,2) = s(c,3) = s(a,4) =$

5 because for each of these edges, the other two are link compatible to sequentially adjacent residues. In contrast, $s(d, 6) = 1$ because no other edge could be required that would have a link compatible GS to a sequentially adjacent residue. Thus the greedy algorithm first requires c to 3, then b to 2, and then a to 4. At this point $s(d, 6) = s(d, 7) = 1$, and either can be chosen. In our shorthand notation, the assignment proceeds from ##### to #c### to bc### to bca## to either bca#d or bcad#. This sequence of events is illustrated in the rightmost branch of Fig. 2.

As shown in the results section, for sequential assignment problems involving large proteins and GSs having many missing rungs, the greedy approach is a recommendable way to explore the different possible assignments. Because of the random selection in step 3, the different assignments found tend to be a broad sampling of the possibilities.

Branch-and-bound assignment algorithm

Another combinatorial optimization approach discussed in standard textbooks (Parker and Rardin 1988) is branch-and-bound. As illustrated in Fig. 2, the idea is to search all possibilities arranged like a family tree, starting at the initial problem (the root) drawn at the top, and working down the branches toward solutions (leaves of the tree). To be efficient for a particular class of problems, one must choose a definition for branching and a test for bounding. We also choose a depth-first search, as opposed to a breadth-first search, so that relatively promising branches are explored downward in search of solutions in case the full examination of the tree is infeasible.

For the sequential assignment problem, a branching procedure we have found to be suitable is just the assignment choice used in the greedy algorithm above. Thus a search tree vertex splits into two children vertices: one where the graph edge between the chosen residue r and GS i has been removed, and the other where the i to r assignment has been required and all other edges to vertices i and r have been removed. The edge that is required and deleted in the two children vertices of the search tree is simply the first edge with the maximum score calculated as in the greedy procedure. This tends to require relatively promising edges early in the search tree that will lead to segments of assigned residues. However, the branch-and-bound exhaustive search tree examines all the possible choices if time permits, so the ultimate results do not depend critically on the ordering of edge choices.

The bounding test amounts to an upper bound on the number of GSs that could possibly be used. Considering only the GS to residue edges in the bipartite graph, the whole graph may consist of one or more connected, mutually disjoint subgraphs. For each subgraph, the

maximal number of GSs that can be assigned is the lesser of the number of GS vertices and residue vertices. The upper bound on the whole graph is the sum of these estimates for all the subgraphs. The user specifies the minimal fraction of GSs that must be assigned to residues in order to be an acceptable assignment. If the upper bound is less than that fraction, then further exploration from this vertex in the search tree is pointless.

Having chosen a branching procedure and a bound test, the search procedure is very straightforward. (1) Start with the initial bipartite graph at the root. (2) If the current search tree node is an assignment (all edges are required), then test it (step 3) and backtrack (step 4); otherwise apply the bound test (step 5). (3) If the required fraction of GSs have been assigned, and this assignment has not been seen before, add it to the list of solutions. (4) Move up the tree to the first higher node having an unexplored child, and move to it for a repeat of step 2. If no such node can be found, the backtracking has returned to the root, and the search is complete. (5) If the current node fails the bound test, backtrack (step 4); otherwise branch (step 6). (6) Branch by creating two child search nodes by deleting and requiring the first edge with highest score. Then move to the child with the required edge for a repeat of step 2.

The search is illustrated on an extremely simple example in Fig. 2, where the bounding test requires that 100% of the GSs must be used in each assignment, and the branching choices are prioritized by the score described in the greedy algorithm. As shown in the figure, the right-hand branches correspond to requiring an edge, while the left-hand branches eliminate that edge. Note how the graphs become simpler as more edges are required. The greedy algorithm would travel down the right side of the tree always making assignments, whereas the branch-and-bound algorithm searches the tree more thoroughly and discovers a third assignment.

The total size of the search tree grows rapidly with increasing protein size and numbers of GSs, and with decreasing numbers of rungs on the GSs. For high quality data on a small protein, it is possible in a reasonable time to search the whole tree for assignments using nearly all the GSs. Otherwise, branching continues until a preset time limit is reached, but this produces a deceptive set of assignments that all agree on the earlier parts of the tree simply because the branch-and-bound search did not have a chance to backtrack very far from the leaves corresponding to acceptable assignments. In order to get the broader sampling of the greedy algorithm while also enjoying the branch-and-bound algorithm's better local searching for adequately good assignments, we have combined them. The allotted time is divided into, say, tenths for ten different attempts. Each attempt consists of a greedy descent of the search tree with different random choices of required edges,

followed by a branch-and-bound search from that vertex in the tree, enumerating alternative acceptable assignments until the time for that attempt expires. For challenging assignment problems, this gives a fairer assessment of the variety of acceptable assignments where any consistencies are more likely to be genuine features of the data and less likely to be artifacts of the search procedure. So either by the purely greedy algorithm or this combination procedure, the general result is either some set of acceptable assignments or none when the time limit was too low, or the demands on fraction of GSs to be used were too high.

Examining multiple assignments

If multiple assignments are found, some subsequent analysis is helpful to determine which features are consistent throughout all assignments, and can therefore be given high confidence. Also when some alternative features are consistently associated with certain other features, such a set may characterize one of the alternative conformations of the protein. Consider the three assignments shown at the bottom of Fig. 2. GS *c* is always assigned to residue 3, and GS *a* is always assigned to residue 4. GS *b* can be assigned to either residue 2 or 6. In the former case, GS *d* can be assigned to either residue 6 or 7, while in the latter case, GS *d* is consistently assigned to residue 7.

For large proteins and incomplete data, SAGA can sometimes find thousands of different assignments, but they can be summarized in a readable form in the following way. Subsets of residues consistently assigned the same way are given an identifier, starting with 0 for the set of residues that are assigned consistently in all assignments, then 1 for the set of residues consistently assigned in the next largest subset of assignments, and so on. Then for each residue there are zero or more different GSs assigned to it, each alternative being marked with an identifier to show what other residue assignments are consistent with it. Table 1 shows how the three assignments in Fig. 2 can be presented, for example. Note that residue R3 is consistently assigned to GS *c*, and P5 is consistently unassigned of course, so they share the assignment identifier 0. However in 33% of the assignments, residue L6 is assigned to GS *b*, but whenever this happens, residue D2 is unassigned, so to show this correlation, they share the identifier 3.

Software implementation

Saga is written in Python v. 2.6 as a standalone program, and it should be fairly operating system independent. It has been tested on several Macintosh computers with Intel CPUs. For academic use, the program is available from the authors free of charge, with citation obligation. Industrial users should contact the University of Michigan Technology Transfer

Table 1 Summary of multiple assignments from Fig. 2

Residue	GS	Identifier	%
A1		0	100
D2	b	1	67
D2		3	33
R3	c	0	100
E4	a	0	100
P5		0	100
L6	b	3	33
L6	d	4	33
L6		5	33
E7	d	2	67
E7		4	33

Office. Most of the tests in the following section were run on an Apple MacBookPro5 with a 2.4 GHz Intel Core 2 Duo processor and 2 GB of memory.

Results and discussion

Viability

For a simple but realistic test of SAGA, consider the favorite test protein, ubiquitin. There are 78 residues in the chain, of which 74 are assignable. Constructing GSs from a full set of Sparky files (HNCA, HNCB, and HNCO) produces 69 GSs, all of which can occupy one or more residues. The mean number of edges associated with each GS or residue vertex is 11, so there is nontrivial ambiguity in even such a high quality dataset for a small protein. The more traditional approach taken by our earlier CASA algorithm first forms the 69 GSs into five segments of sequentially joined GSs, and these are quickly and unambiguously placed on the sequence, thereby assigning 69 residues. The single assignment agrees exactly with the manual assignment. SAGA, however, does not insist on initially forming segments of GSs. Yet the greedy algorithm finds the same assignment in just 1 min (results not shown). The branch-and-bound search using 10 random greedy starts finds 13 different assignments in a total of 2 h. In each of these 13, all 69 GSs were required to be used, and they differ from the manual assignment mostly by assigning one or two GSs to residues near the C-terminus, which were unassigned in the manual assignment.

Even in this example of a small protein with high quality data, the clique algorithm shows how enormous the task is to examine all assignments. The 69 GSs and the 74 assignable residues correspond to 792 vertices in the assignment graph, which is much less than $69 \times 74 = 5,106$ because the GSs are on average compatible

M 1		L 41		T 81	T81 1	A 121	A121 1	N 161		
A 2		W 42		W 82	W82 1	S 122		G 162	G162 0.5	G124 0.3
D 3		V 43	K120 0.5	L 83	L83 1	N 123	N123 0.4	A 163		
A 4		K 44	K120 0.2	K 84	K84 1	G 124	G124 0.5	V 164	V164 1	
A 5		R 45		D 85	D85 1	N 125		D 165	D165 1	
S 6		P 46		A 86	A86 1	L 126	L126 1	A 166	A166 1	
D 7	N123 0.1	N 47		T 87	T87 1	K 127	K127 1	A 167	A167 1	
L 8		L 48		G 88	G88 1	Q 128	Q128 1	K 168	K168 1	
K 9		F 49		N 89		F 129	F129 1	F 169	F169 1	
S 10		N 50		T 90		T 130	T130 1	T 170	T170 1	
R 11		W 51	W51 1	P 91		I 131	I131 1	F 171		
L 12		H 52	H52 1	F 92	F92 1	N 132	N132 1	T 172		
D 13		M 53	M53 1	M 93	M93 1	V 133	V133 1	P 173		
K 14		T 54	T54 1	L 94	L94 1	G 134	G134 1	P 174		
V 15		Q 55	Q55 1	I 95	I95 1	R 135	R135 1	Q 175	Q175 1	
S 16		P 56		A 96	A96 1	D 136	D136 1	G 176	G176 1	
S 17		D 57	D57 1	R 97	R97 1	G 137	G137 1	V 177	V177 1	
F 18	N123 0.1	E 58	E58 1	N 98	N98 1	T 138	T138 1	T 178	T178 1	
H 19		S 59	S59 1	Q 99	Q99 1	I 139	I139 1	V 179	V179 1	
A 20		I 60	I60 1	S 100		H 140	H140 1	D 180	D180 1	
S 21		L 61	L61 1	S 101		Q 141	Q141 1	D 181	D181 1	
F 22	N123 0.1	V 62	V62 1	D 102	D102 1	F 142	F142 1	Q 182	Q182 1	
T 23		S 63	S63 1	W 103	W103 1	S 143	S143 1	R 183	R183 1	
Q 24		D 64	D64 1	Q 104	Q104 1	A 144	A144 1	K 184	K184 1	
K 25		G 65	G65 1	Q 105	Q105 1	V 145	V145 1	L 185	L185 1	
V 26		K 66	K66 1	Y 106	Y106 1	E 146	E146 1	E 186	E186 1	
T 27		T 67	T67 1	N 107	N107 1	Q 147	Q147 1			
D 28	N123 0.3	L 68	L68 1	I 108	I108 1	D 148	D148 1			
G 29	G162 0.3 G124 0.2	W 69	W69 1	K 109	K109 1	D 149	D149 1			
S 30		F 70	F70 1	Q 110	Q110 1	Q 150	Q150 1			
G 31		Y 71	Y71 1	N 111	N111 1	R 151	R151 1			
A 32		N 72	N72 1	G 112		S 152	S152 1			
A 33		P 73		D 113	D113 1	S 153	S153 1			
V 34		F 74	F74 1	D 114	D114 1	Y 154	Y154 1			
Q 35	K120 0.2	V 75	V75 1	F 115	F115 1	Q 155	Q155 1			
E 36		E 76	E76 1	V 116	V116 1	L 156	L156 1			
G 37		Q 77	Q77 1	L 117	L117 1	K 157	K157 1			
Q 38		A 78	A78 1	T 118	T118 1	S 158	S158 1			
G 39		T 79		P 119		Q 159	Q159 1			
D 40		A 80	A80 1	K 120	K120 0.1	Q 160	Q160 1			

Fig. 4 10 min greedy assignment of 186-residue *E. Coli* Chaperone LolA (BMR 10078). All $C\alpha$, $C\beta$ and CO connectivities were given for the white fields in the first columns. No information was given for the black fields. White fields in the second columns show SAGA's

assignments corresponding to the literature assignment. Black fields in the second or third columns show alternative assignments. The numbers in these fields show the fraction of those assignments

H 1		L 41		T 81	T81 1	A 121	A121 1	N 161		
A 2		W 42	W42 1	W 82	W82 1	S 122		G 162	G162 0.3	G124 0.2
D 3		V 43	V43 1	L 83	L83 1	N 123	N123 0.3	A 163		
A 4		K 44	K44 1	K 84	K84 1	G 124	G124 0.3	V 164	V164 1	
A 5	A5 1	R 45	R45 1	D 85	D85 1	N 125		D 165	D165 1	
S 6	S6 1	P 46		A 86	A86 1	L 126	L126 1	A 166	A166 1	
D 7	D7 1	N 47	N47 1	T 87	T87 1	K 127	K127 1	A 167	A167 1	
L 8	L8 1	L 48	L48 1	G 88	G88 1	Q 128	Q128 1	K 168	K168 1	
K 9	K9 1	F 49	F49 1	N 89		F 129	F129 1	F 169	F169 1	
S 10	S10 1	N 50	N50 1	T 90	T90 0.4	T 130	T130 1	T 170	T170 1	
R 11	R11 1	W 51	W51 1	P 91		I 131	I131 1	F 171		
L 12	L12 1	H 52	H52 1	F 92	F92 1	N 132	N132 1	T 172	T254 0.4 T267 0.3 S263 0.2 T67 0.1	
D 13	D13 1	M 53	M53 1	M 93	M93 1	V 133	V133 1	P 173		
K 14	K14 1	T 54	T54 1	L 94	L94 1	G 134	G134 1	P 174		
V 15	V15 1	Q 55	Q55 1	I 95	I95 1	R 135	R135 1	Q 175	Q175 1	
S 16		P 56		A 96	A96 1	D 136	D136 1	G 176	G176 1	
S 17	S17 1	D 57	D57 0.5 D257 0.5	R 97	R97 1	G 137	G137 1	V 177	V177 1	
F 18	F18 1	E 58	E58 0.5 E258 0.5	N 98	N98 1	T 138	T138 1	T 178	T178 1	
H 19	H19 1	S 59	S59 0.5 S259 0.5	Q 99	Q99 1	I 139	I139 1	V 179	V179 1	
A 20	A20 1	I 60	I60 0.5 I260 0.5	S 100		H 140	H140 1	D 180	D180 1	
S 21	S21 1	L 61	L61 0.5 L261 0.5	S 101		Q 141	Q141 1	D 181	D181 1	
F 22	F22 1	V 62	V62 0.5 V262 0.5	D 102	D102 1	F 142	F142 1	Q 182	Q182 1	
T 23	T23 1	S 63	S63 0.5 S263 0.5	W 103	W103 1	S 143	S143 1	R 183	R183 1	
Q 24	Q24 1	D 64	D64 0.5 D264 0.5	Q 104	Q104 1	A 144	A144 1	K 184	K184 1	
K 25	K25 1	G 65	G65 0.5 G265 0.5	Q 105	Q105 1	V 145	V145 1	L 185	L185 1	
V 26	V26 1	K 66	K66 0.5 K266 0.5	Y 106	Y106 1	E 146	E146 1	E 186	E186 1	
T 27	T27 1	T 67	T67 0.5 T267 0.5	N 107	N107 1	Q 147	Q147 1			
D 28	D28 1	L 68	L68 0.5 L268 0.5	I 108	I108 1	D 148	D148 1			
G 29	G29 1	W 69	W69 0.5 W269 0.5	K 109	K109 1	D 149	D149 1			
S 30	S30 1	F 70	F70 0.5 F270 0.5	Q 110	Q110 1	Q 150	Q150 1			
G 31	G31 1	Y 71	Y71 0.5 Y271 0.5	N 111	N111 1	R 151	R151 1			
A 32	A32 1	N 72	N72 0.5 N272 0.2 Y71 0.2 N250 0.1	G 112		S 152	S152 1			
A 33	A33 1	P 73		D 113	D113 1	S 153	S153 1			
V 34	V34 1	F 74	F74 1	D 114	D114 1	Y 154	Y154 1			
Q 35	Q35 1	V 75	V75 1	F 115	F115 1	Q 155	Q155 1			
E 36	E36 1	E 76	E76 1	V 116	V116 1	L 156	L156 1			
G 37	G37 1	Q 77	Q77 1	L 117	L117 1	K 157	K157 1			
Q 38	Q38 1	A 78	A78 1	T 118	T118 1	S 158	S158 1			
G 39	G39 1	T 79		P 119		Q 159	Q159 1			
D 40		A 80	A80 1	K 120	K120 1	Q 160	Q160 1			

Fig. 5 10 min greedy assignment of 186-residue *E. Coli* Chaperone LolA (BMR 10078). All $C\alpha$, $C\beta$ and CO connectivities were given as described in the legend to Fig. 3. Duplicate GSs were given for the

grey area, residues 57–72 (shifted as a group in all dimensions by 0.2 ppm and labeled with numbers 257–272)

intermediate exchange. We simulated this “devilish” situation and show SAGA's performance in Fig. 6. While not as clean as in Fig. 4, the assignment is still completely credible. To show the degeneracy in the assignment, we have represented the sequence with the six NMR-distinguishable residue types (See Fig. S2).

Size limits

Having established that the algorithm performs very well in realistic situations with missing and/or duplicated data, we tested the size limits of the program. As a large example, consider maleate synthase G of *E. Coli*, which contains 723

Fig. 6 30 min greedy assignment of 186-residue *E. Coli* Chaperone LolA (BMR 10078). All $C\alpha$, $C\beta$ and CO connectivities were given as described in the legend to Fig. 3. Duplicate GSs were given for the grey area, residues 57–72 (shifted as a group in all dimensions by 0.2 ppm and labeled with numbers 257–272). The amino acid

sequence is represented as an “NMR” sequence, in which indistinguishable residue types have been collected (A = A; G = G; P = P; S+T = S; D+F+I+L+N+Y = D; E+C+H+K+M+Q+R+V+W = E. See Fig. S2)

residues. The original spectra were assigned using a combination of 4D and 3D NMR methods (Tugarinov et al. 2002), and the assignment list was deposited as BMR 5471. Of the 723 residues, 692 are possibly assignable, and an artificially complete set of three Sparky files based on the BMRB entry produced 654 GSs, 615 of which have 3-rung connectivities, 53 have two-rung connectivities and 5 have 1-rung connectivities.

Using the greedy algorithm, we find in 60 min virtually complete assignments that for the most part agree with the literature assignments (see Fig. 7). Interestingly, the program generated also several stretches of alternative assignments that comprise more than several residues in a row and are hence worth considering. This illustrates the advantage of the program: it will point out areas that are assigned with confidence, but also areas for which additional experiments need to be carried out (e.g. NOESY connectivities or residue-specific labeling), or which should not be used in subsequent work.

Using SAGA to test and extend partial assignments in large proteins

We used SAGA to assess the viability of partial hand assignments of large proteins, using original Sparky peak pick files. As explained in methods, these files were hand-curated to remove noise and sidechain peaks, to add an occasional missed cross peak, and were synchronized between the six triple resonance spectra. That is, HCN cross peaks at the same (H,N) coordinates were given the same root name on authority of the expert spectroscopist.

In the case of a 501-residue example of the Hsp70 chaperone DnaK from *Thermos Thermophilus*, (called TTH-501 hereafter), there are 474 assignable residues and 405 GSs. The hand assignments were based on 351 GSs with $C\alpha$ -matches, 88 GSs with $C\beta$ -matches and 275 GSs with CO-matches. These combined into 82 GSs with 3-rung connectivities, 194 GSs with 2-rung connectivities, and 80 GSs with single-rung connectivities. A 10 h greedy run produced 30 assignments with probabilities between 1 and 0.2. The complete results are shown in Fig. 8, which shows that most of the hand assignments are indeed “found” by the program. However, there are many more valid assignments, and the question arises how one would go about picking the “correct” ones if one does not know the answer. We proceeded as follows. (1) We keep all assignments of probability 0.6 or higher. (2) We discard all assignments with probability less than 0.2. (3) Of the remaining set, we only retain those GSs that are assigned in stretches of 3 or more residues. (4) The results are checked for mutual compatibility. (5) If the same stretch of linked GSs occurs more than once, we retain the one that occurs in a longer overall stretch. The results of this editing are shown in Fig. 9. The retained assignments are very compatible with the hand assignment, but several other credible assignments are found as well.

SAGA clearly discloses the risk of obtaining partial assignments without confirming them with further scrutiny, such as ^{13}C lineshape comparisons in the different spectra, independent assignment experiments at different temperature or pH, extra information from selective labeling, 4D methods (Tugarinov et al. 2002), NOE connectivities and

E 101 E101 1	E 102 E102 1	E 201 E201 1	E 301 E301 1	D 401 D401 1	E 501 E501 1	E 601 E601 1	E 701 E701 1
S 103 S103 1	D 104 D104 1	D 202 D202 1	D 302 D302 1	A 402 A402 1	G 502 G502 1	G 602 G602 1	E 702 E702 1
D 105 D105 1	D 106 D106 1	G 206 G206 1	G 306 G306 1	D 403 D403 1	E 503 E503 1	D 603 D603 1	E 703 E703 1
S 107 S107 1	D 108 E108 1	E 207 E207 1	E 307 E307 1	D 404 D404 1	D 504 D504 1	D 604 D604 1	D 704 D704 1
D 109 D109 1	S 110 S110 1	S 208 S208 1	E 308 E308 1	D 405 D405 1	G 505 G505 1	G 605 G605 1	D 705 D705 1
E 111 E111 0.5	E 112 E112 0.5	D 210 D210 1	D 310 D310 1	S 411 S411 1	E 506 E506 1	D 606 D606 1	G 706 G706 1
D 112 D112 0.5	A 113 A113 1	E 211 E211 1	E 311 E311 1	S 412 S412 1	E 507 E507 1	E 607 E607 1	S 707 S707 1
A 114 A114 0.5	D 115 D115 1	D 212 D212 1	D 312 D312 1	E 413 E413 1	E 508 E508 1	E 608 E608 1	S 708 S708 1
D 116 D116 1	E 117 E117 0.5	E 213 E213 1	D 313 D313 1	D 414 D414 1	E 509 E509 1	E 609 E609 1	E 709 E709 1
E 118 E118 1	E 119 E119 0.5	E 214 E214 1	D 314 D314 1	E 415 E415 1	E 510 E510 1	E 610 E610 1	E 710 E710 1
E 120 E120 1	A 121 A121 1	D 215 D215 1	D 315 D315 1	A 416 A416 1	E 511 E511 1	E 611 E611 1	E 711 E711 1
D 121 D21 1	E 122 E122 1	E 216 E216 1	D 316 D316 1	E 417 E417 1	D 512 D512 1	E 612 E612 1	D 712 D712 1
E 122 E22 1	E 123 E123 1	E 217 E217 1	E 317 E317 1	E 418 E418 1	E 513 E513 1	E 613 E613 1	E 713 E713 1
E 20 E20 1	A 124 A124 1	E 218 E218 1	E 318 E318 1	E 419 E419 1	E 514 E514 1	E 614 E614 1	A 714 A714 1
D 21 D21 1	A 125 A125 0.5	E 219 E219 1	E 319 E319 1	E 420 E420 1	E 515 E515 1	E 615 E615 1	E 715 E715 1
E 22 E22 1	E 126 E126 0.5	E 220 E220 1	E 320 E320 1	E 421 E421 1	E 516 E516 1	E 616 E616 1	E 716 E716 1
E 23 E23 1	E 127 E127 0.5	E 221 E221 1	E 321 E321 1	E 422 E422 1	E 517 E517 1	E 617 E617 1	E 717 E717 1
E 24 E24 1	E 128 E128 0.5	E 222 E222 1	E 322 E322 1	E 423 E423 1	E 518 E518 1	E 618 E618 1	E 718 E718 1
E 25 E25 1	E 129 E129 0.5	E 223 E223 1	E 323 E323 1	E 424 E424 1	E 519 E519 1	E 619 E619 1	E 719 E719 1
E 26 E26 0.5	E 130 E130 0.5	E 224 E224 1	E 324 E324 1	E 425 E425 0.5	E 520 E520 1	E 620 E620 1	E 720 E720 1
E 27 E27 1	E 131 A131 0.5	E 225 A225 0.5	E 325 E325 1	D 426 D426 0.5	E 521 E521 1	E 621 E621 1	E 721 E721 1
E 28 E28 1	E 132 A132 0.5	E 226 E226 1	E 326 E326 1	E 427 E427 0.5	E 522 E522 1	E 622 E622 1	E 722 E722 1
E 29 E29 1	E 133 A133 1	E 227 E227 1	E 327 E327 1	E 428 E428 0.5	E 523 E523 1	E 623 E623 1	E 723 E723 1
E 30 E30 1	E 134 E134 1	E 228 E228 1	E 328 E328 1	E 429 E429 0.5	E 524 E524 1	E 624 E624 1	E 724 E724 1
E 31 E31 1	E 135 E135 1	E 229 E229 1	E 329 E329 1	E 430 E430 0.5	E 525 E525 1	E 625 E625 1	E 725 E725 1
E 32 E32 1	E 136 E136 1	E 230 E230 1	E 330 E330 1	E 431 E431 1	E 526 E526 1	E 626 E626 1	E 726 E726 1
E 33 E33 1	E 137 E137 0.5	E 231 E231 1	E 331 E331 1	E 432 E432 1	E 527 E527 1	E 627 E627 1	E 727 E727 1
E 34 E34 1	E 138 E138 1	E 232 E232 1	E 332 E332 1	E 433 E433 1	E 528 E528 1	E 628 E628 1	E 728 E728 1
E 35 E35 1	E 139 E139 1	E 233 E233 1	E 333 E333 1	E 434 E434 1	E 529 E529 1	E 629 E629 1	E 729 E729 1
E 36 E36 1	E 140 E140 1	E 234 E234 1	E 334 E334 1	E 435 E435 1	E 530 E530 1	E 630 E630 1	E 730 E730 1
E 37 E37 0.5	E 141 A141 1	E 235 E235 1	E 335 E335 1	E 436 E436 1	E 531 E531 1	E 631 E631 1	E 731 E731 1
E 38 E38 1	E 142 A142 1	E 236 E236 1	E 336 E336 1	E 437 E437 1	E 532 E532 1	E 632 E632 1	E 732 E732 1
E 39 E39 1	E 143 E143 1	E 237 E237 1	E 337 E337 1	E 438 E438 1	E 533 E533 1	E 633 E633 1	E 733 E733 1
E 40 E40 1	E 144 E144 1	E 238 E238 1	E 338 E338 1	E 439 E439 1	E 534 E534 1	E 634 E634 1	E 734 E734 1
E 41 E41 1	E 145 E145 1	E 239 E239 1	E 339 E339 1	E 440 E440 1	E 535 E535 1	E 635 E635 1	E 735 E735 1
E 42 E42 1	E 146 E146 1	E 240 E240 0.5	E 340 E340 1	E 441 E441 1	E 536 E536 1	E 636 E636 1	E 736 E736 1
E 43 E43 1	E 147 E147 0.5	E 241 E241 0.5	E 341 E341 1	E 442 E442 1	E 537 E537 1	E 637 E637 1	E 737 E737 1
E 44 E44 1	E 148 E148 1	E 242 E242 0.5	E 342 E342 1	E 443 E443 1	E 538 E538 1	E 638 E638 1	E 738 E738 1
E 45 E45 1	E 149 E149 1	E 243 E243 0.5	E 343 E343 1	E 444 E444 1	E 539 E539 1	E 639 E639 1	E 739 E739 1
E 46 E46 1	E 150 E150 1	E 244 E244 0.5	E 344 E344 1	E 445 E445 1	E 540 E540 1	E 640 E640 1	E 740 E740 1
E 47 E47 1	E 151 E151 1	E 245 E245 0.5	E 345 E345 1	E 446 E446 1	E 541 E541 1	E 641 E641 1	E 741 E741 1
E 48 E48 1	E 152 E152 1	E 246 E246 1	E 346 E346 1	E 447 E447 1	E 542 E542 1	E 642 E642 1	E 742 E742 1
E 49 E49 1	E 153 E153 1	E 247 E247 1	E 347 E347 1	E 448 E448 1	E 543 E543 1	E 643 E643 1	E 743 E743 1
E 50 E50 1	E 154 E154 1	E 248 E248 1	E 348 E348 1	E 449 E449 1	E 544 E544 1	E 644 E644 1	E 744 E744 1
E 51 E51 1	E 155 E155 1	E 249 E249 1	E 349 E349 1	E 450 E450 1	E 545 E545 1	E 645 E645 1	E 745 E745 1
E 52 E52 1	E 156 E156 1	E 250 E250 1	E 350 E350 1	E 451 E451 1	E 546 E546 1	E 646 E646 1	E 746 E746 1
E 53 E53 1	E 157 E157 1	E 251 E251 1	E 351 E351 1	E 452 E452 1	E 547 E547 1	E 647 E647 1	E 747 E747 1
E 54 E54 1	E 158 E158 1	E 252 E252 1	E 352 E352 1	E 453 E453 1	E 548 E548 1	E 648 E648 1	E 748 E748 1
E 55 E55 1	E 159 E159 1	E 253 E253 1	E 353 E353 1	E 454 E454 1	E 549 E549 1	E 649 E649 1	E 749 E749 1
E 56 E56 1	E 160 E160 1	E 254 E254 1	E 354 E354 1	E 455 E455 1	E 550 E550 1	E 650 E650 1	E 750 E750 1
E 57 E57 1	E 161 E161 1	E 255 E255 1	E 355 E355 1	E 456 E456 1	E 551 E551 1	E 651 E651 1	E 751 E751 1
E 58 E58 1	E 162 E162 1	E 256 E256 1	E 356 E356 1	E 457 E457 1	E 552 E552 1	E 652 E652 1	E 752 E752 1
E 59 E59 1	E 163 E163 1	E 257 E257 1	E 357 E357 1	E 458 E458 1	E 553 E553 1	E 653 E653 1	E 753 E753 1
E 60 E60 1	E 164 E164 1	E 258 E258 1	E 358 E358 1	E 459 E459 1	E 554 E554 1	E 654 E654 1	E 754 E754 1
E 61 E61 1	E 165 E165 1	E 259 E259 1	E 359 E359 1	E 460 E460 1	E 555 E555 1	E 655 E655 1	E 755 E755 1
E 62 E62 1	E 166 E166 1	E 260 E260 1	E 360 E360 1	E 461 E461 1	E 556 E556 1	E 656 E656 1	E 756 E756 1
E 63 E63 1	E 167 E167 1	E 261 E261 1	E 361 E361 1	E 462 E462 1	E 557 E557 1	E 657 E657 1	E 757 E757 1
E 64 E64 1	E 168 E168 1	E 262 E262 1	E 362 E362 1	E 463 E463 1	E 558 E558 1	E 658 E658 1	E 758 E758 1
E 65 E65 1	E 169 E169 1	E 263 E263 1	E 363 E363 1	E 464 E464 1	E 559 E559 1	E 659 E659 1	E 759 E759 1
E 66 E66 1	E 170 E170 1	E 264 E264 1	E 364 E364 1	E 465 E465 1	E 560 E560 1	E 660 E660 1	E 760 E760 1
E 67 E67 1	E 171 E171 1	E 265 E265 1	E 365 E365 1	E 466 E466 1	E 561 E561 1	E 661 E661 1	E 761 E761 1
E 68 E68 1	E 172 E172 1	E 266 E266 1	E 366 E366 1	E 467 E467 1	E 562 E562 1	E 662 E662 1	E 762 E762 1
E 69 E69 1	E 173 E173 1	E 267 E267 1	E 367 E367 1	E 468 E468 1	E 563 E563 1	E 663 E663 1	E 763 E763 1
E 70 E70 1	E 174 E174 1	E 268 E268 1	E 368 E368 1	E 469 E469 1	E 564 E564 1	E 664 E664 1	E 764 E764 1
E 71 E71 1	E 175 E175 1	E 269 E269 1	E 369 E369 1	E 470 E470 1	E 565 E565 1	E 665 E665 1	E 765 E765 1
E 72 E72 1	E 176 E176 1	E 270 E270 1	E 370 E370 1	E 471 E471 1	E 566 E566 1	E 666 E666 1	E 766 E766 1
E 73 E73 1	E 177 E177 1	E 271 E271 1	E 371 E371 1	E 472 E472 1	E 567 E567 1	E 667 E667 1	E 767 E767 1
E 74 E74 1	E 178 E178 1	E 272 E272 1	E 372 E372 1	E 473 E473 1	E 568 E568 1	E 668 E668 1	E 768 E768 1
E 75 E75 1	E 179 E179 1	E 273 E273 1	E 373 E373 1	E 474 E474 1	E 569 E569 1	E 669 E669 1	E 769 E769 1
E 76 E76 1	E 180 D180 1	E 274 E274 1	E 374 E374 1	E 475 E475 1	E 570 E570 1	E 670 E670 1	E 770 E770 1
E 77 E77 1	E 181 E181 1	E 275 E275 1	E 375 E375 1	E 476 E476 1	E 571 E571 1	E 671 E671 1	E 771 E771 1
E 78 E78 1	E 182 E182 1	E 276 E276 1	E 376 E376 1	E 477 E477 1	E 572 E572 1	E 672 E672 1	E 772 E772 1
E 79 E79 1	E 183 E183 1	E 277 E277 1	E 377 E377 1	E 478 E478 1	E 573 E573 1	E 673 E673 1	E 773 E773 1
E 80 E80 1	E 184 E184 1	E 278 E278 1	E 378 E378 1	E 479 E479 1	E 574 E574 1	E 674 E674 1	E 774 E774 1
E 81 E81 1	E 185 D185 1	E 279 E279 1	E 379 E379 1	E 480 E480 1	E 575 E575 1	E 675 E675 1	E 775 E775 1
E 82 E82 1	E 186 E186 1	E 280 E280 1	E 380 E380 1	E 481 E481 1	E 576 E576 1	E 676 E676 1	E 776 E776 1
E 83 E83 1	E 187 E187 1	E 281 E281 1	E 381 E381 1	E 482 E482 1	E 577 E577 1	E 677 E677 1	E 777 E777 1
E 84 E84 1	E 188 E188 1	E 282 E282 1	E 382 E382 1	E 483 E483 1	E 578 E578 1	E 678 E678 1	E 778 E778 1
E 85 E85 1	E 189 E189 1	E 283 E283 1	E 383 E383 1	E 484 E484 1	E 579 E579 1	E 679 E679 1	E 779 E779 1
E 86 E86 1	E 190 E190 1	E 284 E284 1	E 384 E384 1	E 485 E485 1	E 580 E580 1	E 680 E680 1	E 780 E780 1
E 87 E87 1	E 191 E191 1	E 285 E285 1	E 385 E385 1	E 486 E486 1	E 581 E581 1	E 681 E681 1	E 781 E781 1
E 88 E88 1	E 192 E192 1	E 286 E286 1	E 386 E386 1	E 487 E487 1	E 582 E582 1	E 682 E682 1	E 782 E782 1
E 89 E89 1	E 193 E193 1	E 287 E287 1	E 387 E387 1	E 488 E488 1	E 583 E583 1	E 683 E683 1	E 783 E783 1
E 90 E90 1	E 194 E194 1	E 288 E288 1	E 388 E388 1	E 489 E489 1	E 584 E584 1	E 684 E684 1	E 784 E784 1
E 91 E91 1	E 195 E195 1	E 289 E289 1	E 389 E389 1	E 490 E490 1	E 585 E585 1	E 685 E685 1	E 785 E785 1
E 92 E92 1	E 196 E196 1	E 290 E290 1	E 390 E390 1	E 491 E491 1	E 586 E586 1	E 686 E686 1	E 786 E786 1
E 93 E93 1	E 197 E197 1	E 291 E291 1	E 391 E391 1	E 492 E492 1	E 587 E587 1	E 687 E687 1	E 787 E787 1
E 94 E94 1	E 198 E198 1	E 292 E292 1	E 392 E392 1	E 493 E493 1	E 588 E588 1	E 688 E688 1	E 788 E788 1
E 95 E95 1	E 199 E199 1	E 293 E293 1	E 393 E393 1	E 494 E494 1	E 589 E589 1	E 689 E689 1	E 789 E789 1
E 96 E96 1	E 200 D200 1	E 294 E294 1	E 394 E394 1	E 495 E495 1	E 590 E590 1	E 690 E690 1	E 790 E790 1
E 97 E97 1	E 201 E201 1	E 295 E295 1	E 395 E395 1	E 496 E496 1	E 591 E591 1	E 691 E691 1	E 791 E791 1
E 98 E98 1	E 202 E202 1	E 296 E296 1	E 396 E396 1</				

Fig. 8 10 h greedy assignment of a 501-residue construct of the Hsp70 chaperone DnaK from *Thermos Thermophilus*, with 474 assignable residues and 405 GSs. The hand assignments were based on 82 GSs with 3-rung connectivities, 194 GSs with 2-rung connectivities, and 80 GSs with single-rung connectivities. The connected GSs were given for the *white fields* in the first columns. No information was given for the *black fields*. *White fields* in the second

columns show SAGA’s assignments corresponding to the literature assignment. *Black fields* in the second and following columns show alternative assignments. The numbers in these fields show the rank order of those assignments, with 0 being the best. The amino acid sequence is represented as an “NMR” sequence, in which indistinguishable residue types have been collected

tolerances and larger type tolerances in this additional run. The resulting assignment is shown in Fig. 11. Of the 101 additional GSs, 46 could be placed with a confidence of 0.5 or better. Clearly, some of these could not ever have been found by “hand”: Consider GS X954, which is placed with 100% confidence on V18. This GS has no C α (i) or C β (i) rungs, while its CO(i) rung is unmatched, because G19 does not show any CO(i-1) rungs. SAGA, however, placed GS X954 on the basis of its (unmatched) i–1 rungs that put it next to a residue of that type.

Using the assignment to assess experimental assignment strategies

Having established the reliability of the program, we assess here the inverse question: how complete do NMR data sets

need to be in order to be uniquely assignable? As is shown in Figures S3 and S4, there is no real need for CO connectivities, even for assigning the resonances of a 381-residue protein. This substantiates the assertion in the introduction that, theoretically, two sets of complete rung connectivities are sufficient to uniquely sequentially link the GSs of proteins, maybe even larger than 1,000 residues. The loss of the CO-rungs does not affect the placement of the GSs on the sequence, as the CO chemical shifts have almost no predictive value for residue type (see Fig. S2).

However, as Fig. S5 shows, one cannot rely on just complete C α and CO for an assignment of a 20 kDa protein or larger. The reason, of course, is that the C α resonance statistics are not sufficiently restrictive to allow placement of the GSs on the sequence (see Fig. S2). Figure S5 also shows how well SAGA performs: it finds many feasible

E 1		E 101	E101	1	3	D 201	D201	1	2	E 301	E301	1	2	E 401	E401	0.6	2		
F 2	A2	0.8	2	F 102	F102	1	3	E 202	E202	1	3	D 302	D302	1	2	E 402	E402	0.6	2
F 3	E3	0.8	3	E 103	E103	1	3	E 203	D203	1	3	D 303	S303	1	2	S 403	S403	0.6	
F 4	A4	1	3	G 104	G96	0.6	1	G 204	G204	1	1	E 304	E304	1	2	E 404	E404		
F 5	E5	1	2	E 105	E105	0.8	3	E 205	E205	0.8	1	P 305				D 405	D405		
G 6	G6	1	1	D 106	D106	1	3	G 206	G206	0.8		D 306				D 406	D406		
D 7	D7	1	1	D 107	D107	1	3	E 207	E207			D 307				P 407	P407		
D 8	D497	0.6		S 108	S108	1		D 208	A175	0.6	2	E 308	E308			E 408	E408	0.4	3
D 9				P 109				E 209	E115	1	2	E 309	E309			D 409	D409	0.4	3
G 10				E 110	E110	1	1	E 210	D116	1		D 310				S 410	S410	0.4	3
S 11				E 111	E111	1	2	E 211	E211			E 311	E311			E 411	S411	0.4	2
S 12				D 112	D112	1	2	G 212	A212	0.8	3	G 312	G312			D 412	D412	0.4	2
D 13				S 113	S113	1	2	S 213	S213	1	2	P 313				P 413	P413		
S 14				A 114	A114	0.8	2	S 214	S214	0.8		E 314	E314	0.8	2	S 414	S414		
E 15	E15	1	2	E 115	D176	0.6		G 215	G215			E 315	E315	0.8	2	E 415	E415		
D 16	D16	1	2	D 116				D 216	D216			E 316	E316	0.8	2	E 416	E260	0.4	2
A 17	A17	1	2	D 117				S 217	S217			A 317	A317	0.8	3	E 417	D263	0.4	2
E 18	E18	1	3	E 118				E 218	E218			D 318	D318	0.8	3	E 418	E264	0.4	2
D 19	D19	1	2	E 119				D 219	D219			E 319	E319	0.8	1	D 419	D265	0.4	2
E 20	E20	1	2	D 120	D120	0.4	2	G 220	G220	0.6		D 320	D320	1	2	D 420	D265	0.4	2
G 21	G221	0.6		E 121	E121	0.4	1	G 221	A221	1	2	A 321	A321	1	2	S 421	S421		
G 22				E 122	E122	0.4	2	S 222	S222			G 322	G322	1	1	S 422	S422		
E 23				D 123	D123	0.4	2	D 223	D275	0.6		D 323	D323	1	3	A 423	A423		
P 24				A 124	A124	0.8	3	E 224	E224			S 324	S324	1		E 424	E424		
E 25	E25	0.8	2	S 125	S125	1	1	E 225	E225			P 325				E 425	E425		
E 26	E26	1	2	S 126	E126	1	2	A 226	A227	0.6	2	A 326				D 426	D426		
D 27	D27	1	3	E 127	E127	1	3	D 228	D228	0.6	2	E 327	E327			E 427	E427		
E 28	E28	1	2	D 128	D128	1	2	E 229	E229			D 328	D328			S 428	E384	1	1
D 29	D29	1	2	G 129	G129	1	1	E 230	E230			D 329	D329	0.6	2	A 429	A429	1	3
A 30	A30	1	2	E 130	E130	0.6	2	E 231	E231			E 330	E330	0.6	3	E 430	E430	1	2
A 31	E31	1	2	E 131	E131	0.6	2	E 232	E232			E 331	E331	0.6	2	E 431	E431	1	3
G 32	G32	1	1	S 132	D132	0.6		D 232	D232			D 332	D332	0.6		D 432	D432	1	3
E 33	E130	0.4	2	S 134				A 233	A233	0.6		D 333	D333			E 433	E433	1	3
E 34	E363	0.4	2	E 135				E 234	E234			E 334	E334			E 434	E434	1	3
E 35	E364	0.4	2	A 136				E 235	E235			D 335	D335			D 435	D435	1	3
S 36	E365	0.4		E 137				D 236	D236			G 336	G336			E 436	E436	1	2
P 37				D 137				E 237	E237			A 337	A337			G 437	G437	1	1
S 38				S 138				E 238	E238			S 338	S338			E 438	E438	1	2
E 39	E39	0.6	2	E 139				E 239	E239	0.8	2	E 339	E339			E 439	E439	1	1
E 40	E40	0.6	2	P 140				E 240	E240	0.8	2	E 340	E340			P 440	P440		
A 41	A41	0.6	2	A 141				G 241	G241	0.6	1	P 341				E 441	E441	1	3
E 42	D42	0.6	3	D 142				E 242	E242	0.6		A 342				A 442	A442	1	2
E 43	E43	0.6		D 143				D 244	D244	0.6	2	E 343	E343			E 443	E443	1	2
E 44				D 144				E 245	E245	0.6	2	A 344				A 444	A444	1	2
G 45	G45	0.8	1	D 145				E 246	E246	0.6	2	E 345	E75	0.4	3	D 445	D445	1	3
E 46	E46	0.8	2	A 146				A 246	A246	0.6	3	E 346	D76	0.4	2	E 446	E446	1	2
S 47	S47	1	1	E 147	E147	0.4	2	D 247	D247	0.6		E 347	E77	0.4	2	S 447	S447	1	2
D 48	D48	1	3	E 148	E148	0.4	2	E 249	E249			E 348	E78	0.4	3	D 448	D448	1	2
E 49	E49	1	2	E 149	E149	0.4	2	E 250	E250			E 349	E79	0.4	2	G 449	G449	1	1
G 50	G50	0.8	1	A 150	A150	0.4	2	D 251	D251			D 350	E80	0.4		E 450	E450	1	3
E 51				S 151	S151	0.6	1	E 252	E252	0.8	2	D 351				D 451	D451	1	2
E 52				A 152	A152	1	2	E 253	E253	0.6	3	G 352				E 452	E452	1	2
A 53	A53	0.4	2	D 153	D153	1	2	D 254	D254	0.6	2	E 353	E353	0.8	2	E 453	E453	1	2
E 54	E54	0.4	2	A 154	A154	1	2	D 255	D255	0.6	2	E 354	E354	0.6		E 454	E454	1	2
E 55	E55	0.4	2	G 155	G155	1	1	E 256	E256	0.6	2	P 355				G 455	G455	1	1
E 56	E56	0.4	2	E 156	E156	1	1	A 257	A257	0.6	2	D 356				D 456	D456	1	1
A 57	A57	0.4	2	D 157	D157	1	2	E 259	E259			E 357				P 457	P457		
E 58	E58	0.4	2	A 158	A158	1	2	E 260	E260			S 358	D333	0.4	2	P 458	P458		
D 59	D59	0.8	2	G 159	G159	1	1	E 261	E261			E 359	E334	0.4	1	E 459	E459		
D 60	D60	0.8		D 160	D160	1	3	E 262	E262			D 360	D232	0.4		P 460	P460		
E 61				E 161	E161	1	3	E 263	D263	0.6	2	E 361				A 461	A461		
E 62	E62	0.8		E 162	E162	1	3	E 264	E264	0.6	2	D 362				E 462	E462		
G 63				D 163	D163	1	2	E 265	E265	0.6	2	E 363				E 463	E463		
S 64				E 164	E164	1	2	S 266	S266	0.6		E 364				P 464	P464		
D 65	D254	0.4	2	D 165	D165	1	2	E 269	E269			E 365				E 465	E465		
D 66	D255	0.4	2	D 166	D166	1	2	E 270	E270	0.8	2	A 366				D 466	D462	0.6	1
E 67	E256	0.4	2	D 167	D167	1	2	E 271	S271	0.8	2	E 367				E 467	E467	1	3
D 68	D236	0.4	2	E 168	E168	0.8		S 272	S272	1	3	G 368	G368	0.6	1	E 468	E468	1	2
E 69	E237	0.4	2	P 169				S 273	S273	1	3	A 369	A369	0.8	1	E 469	E469	1	3
E 70	E238	0.4	2	S 170	S170	0.6	1	E 274	E274	1	2	A 370				D 470	D470	1	3
D 71				A 171	A171	0.6		S 275	S275	1	3	E 371				D 471	D471	1	3
D 72	D72	0.6	2	D 172	D172	1	2	D 276	D276	1	2	E 372				D 472	D472	1	3
G 73	G73	1	1	A 173	A173	1	2	D 277	D277	1	2	A 373				D 473	D473	1	3
E 74	E74	0.6	3	D 174	D174	1	2	S 278	S278	1	2	G 374				A 474	A474	1	3
E 75	E75	0.6	3	A 175	A175	1	2	D 279	D279	1	2	E 375	E375	0.6		D 475	D475	1	3
D 76	D76	0.6		D 176	D176	1	2	P 280				D 376				G 476	G476	1	1
E 77				G 177	G177	1	1	E 281	E281			E 377				D 477	D477	1	3
E 78				D 178	D178	1	3	D 282	D282			E 379				D 478	D478	1	3
E 79				D 179	D179	1	3	A 279	A279	0.6	2	G 380				E 479	E479	1	3
E 80	E344	0.6	3	E 180	E180	1	3	D 280	D280	0.6	3	E 380				E 480	E480	1	2
E 81	E81	0.8	2	E 181	E181	1	2	D 281	D281	0.8	3	E 381	E218	0.6	2	S 481	S481	1	3
E 82	E82	0.8	2	G 182	G182	1	2	P 282				D 382	D219	0.6	2	A 482	A482	1	3
A 83	A83	0.8	3	D 183	D183	1	3	A 283	A283	1	1	E 383				E 483	E483	1	3
E 84	E84	0.8	3	E 184	E184	1	2	S 284	S284	1	2	E 384				E 484	E484	1	2
E 85	E85	0.8		S 185	S185	1	1	E 285	E285	1	2	D 385				E 485	E485	1	2
E 86				E 186	E186	1	2	S 286	S286	1		D 386				S 486	S486	1	1
P 87				D 187	D187	1	3	P 287				D 387	D387	1	1	S 487	S487	1	1
D 88				E 188	E188	1	2	D 288	D288	0.4	3	E 388	E388	1					

E 1		A 81	A243	0.3	1	A 161	A161	1	2	D 241		D 321	
S 2		E 82	E244	0.3	3	G 162	G162	1	2	D 242		E 322	
E 3	E330	E 83	D245	0.4	2	S 163	S163	1	2	A 243		D 323	D323 0.3
G 4		E 84	E246	0.3		D 164	D164	1	3	E 244		A 324	
P 5		S 85				A 165	A165	1		D 245		E 325	
A 6	A6	D 86				G 166				E 246		D 326	
E 7	E7	E 87				D 167				E 247		D 327	
G 8	G8	E 88	E88	0.6	3	D 168				E 248	E248	E 328	
D 9		E 89	E89	0.6	3	E 169				E 249	E249	S 329	S329 0.8
D 10		E 90	E90	0.6		D 170				E 250	E250	E 330	
D 11		P 91				E 171				E 251	E251	D 331	D331 0.8
G 12		D 92	D92	0.2	3	D 172	D172	0.2	3	D 252	D252	E 332	
S 13		E 93	E93	0.2	3	D 173	D173	0.2	3	D 253		D 333	
S 14		E 94	E94	0.2	3	D 174	D174	0.2		S 254	S254	D 334	
D 15		E 95	E95	0.2	3	E 175				E 255	E255	E 335	
S 16		D 96	D96	0.2	3	P 176				D 256	D256	D 336	
E 17		D 97	D97	0.2		S 177				E 257		E 337	
E 18		A 98				A 178				E 258	E258	G 338	G312 0.5
G 19	G19	G 99				A 179				A 259	A259	G 339	
E 20	E20	E 100				A 180	A180	0.9		E 260	E260	S 340	S313 0.5
D 21	D21	P 101				D 181	D181	1	1	E 261		S 341	
E 22	E22	E 102	E102	0.4	3	A 182	A182	1	3	E 262	E262	E 342	E342 1
E 23		E 103	E103	0.4	3	D 183	D183	1	3	D 263	D263	D 343	
G 24		E 104	E104	0.4	3	G 184	G184	1	3	E 264	E264	P 344	
E 25		E 105	E105	0.4		D 185	D185	1	3	S 265	S265	E 345	
E 26	E26	E 106				D 186	D186	1	3	A 266	A266	D 346	
E 27	E27	D 107	D107	1	3	E 187	E187	1	3	E 267	E267	E 347	
D 28	D28	E 108	E108	1	3	E 188	E188	1		E 268		E 348	E348 1 3
D 29	D29	G 109	G109	1	2	E 189				E 269		D 349	D349 1 3
A 30	A30	E 110	E110	1	3	G 190	G190	1	1	A 270		D 350	D350 1 3
D 31		S 111	S111	1	3	A 191	A191	1	3	E 271		E 351	E351 1 3
D 32		E 112	E112	1	3	E 192	E192	1	3	E 272		D 352	D352 1 3
E 33		S 113	S113	1	2	E 193	E193	1	3	S 273		D 353	D353 1
G 34		D 114	D114	1	3	D 194	D194	1		D 274	D198 0.2	D 354	
D 35		D 115	D115	1		E 195				S 275		D 355	
E 36		P 116				D 196				S 276		G 356	
S 37		E 117				D 197				S 277		E 357	
S 38		E 118	E118	0.7		D 198				S 278		E 358	
P 39		E 119				D 199				E 279	E279	D 359	D359 1 3
S 40	S40	S 120				D 200				A 280	A280	D 360	D360 0.2 2
D 41	D41	S 121	S121	0.8	2	G 201				S 281	S281	E 361	E361 1
E 42	E42	E 122	E122	0.4	2	G 202				D 282	D282	S 362	
A 43	A43	E 123	E123	0.8		G 203				E 283	E283	D 363	D150 0.5 2
D 44	D44	D 124				S 204				D 284		D 364	D151 0.3
S 45	S45	S 125				D 205	D205	1		D 285		P 365	
D 46	D46	E 126				D 206				S 286		D 366	
S 47	S47	E 127				E 207	E207	1		D 287	D198 0.2	E 367	D310 1 2
E 48	E48	E 128	D343	0.2		S 208				D 288		A 368	E311 0.4 1
E 49	E49	E 129				D 209	D209	1	3	E 289	E289	E 369	E132 1 2
D 50	D50	D 130	D323	0.7	3	D 210	D210	1	3	G 290	G290	A 370	A133 0.4 3
D 51	D51	A 131	A324	0.3	3	S 211	S211	1	3	D 291	D291	D 371	D134 1
G 52	G52	E 132	E325	0.7		D 212	D212	1	3	D 292	D292	G 372	
D 53		A 133				E 213	E213	1	2	D 293		A 373	
A 54		D 134	D199	1	1	D 214	D214	1	1	D 294		A 374	
A 55	A55	D 135	D135	1	1	G 215	G215	1	1	S 295		E 375	
E 56	E56	G 136	G136	1	1	D 216	D216	1	3	S 296		E 376	E376 1 1
D 57	D57	E 137	E137	1		D 217	D217	1	3	D 297		A 377	A377 1 1
E 58	E58	S 138				E 218	E218	1	3	S 298	S298	A 378	A378 1 3
E 59	E59	E 139	E139	1	3	E 219	E219	1	3	E 299	E299	D 379	D379 1 3
A 60	A60	S 140	S140	1	2	E 220	E220	1	3	A 300	A300	D 380	D380 1 3
E 61	E61	D 141	D141	1	3	S 221	S221	1	2	E 301	E301	S 381	S381 1 3
D 62	D62	A 142	A142	1		S 222	S222	1		D 302	D302	G 382	G382 1
P 63		E 143				A 223				E 303		D 383	
S 64	S64	E 144				G 224				E 304	E304	E 384	
D 65	D65	S 145	S145	1	1	D 225				D 305	D305	S 385	
S 66	S66	E 146	E146	0.3		S 226				D 306	D306	E 386	
E 67	E67	P 147				E 227				A 307	A307		
D 68	D68	A 148				D 228	D228	1	3	D 308	D308		
D 69		D 149				G 229	G229	0.2	1	D 309			
A 70		D 150	D198	0.2		G 230	G230	1		D 310			
E 71		D 151				E 231				E 311			
E 72	E335	D 152				D 232				G 312			
D 73	D336	S 153				D 233				S 313			
D 74	E337	E 154				D 234	D234	0.3	3	D 314			
G 75	G338	E 155				D 235	D235	0.3	3	D 315			
E 76	E76	E 156				E 236	E236	0.3	3	P 316			
E 77	E77	A 157				E 237	E237	0.3		E 317	D366	0.5	3
D 78		S 158				E 238				E 318	E367	0.8	
D 79	D206	E 159	E159	1	3	D 239				E 319			
D 80		D 160	D160	1	3	E 240				A 320			

Fig. 10 120 min greedy check of an unpublished, partial hand assignment of the 386-residue nucleotide binding domain of human Hsc70 (206 GSs/374 assignable residues). The hand assignments are based on 110 GSs with 3-rung connectivities, 30 GSs with 2-rung connectivities, and 21 GSs with single-rung connectivities (they are given in the *white fields* in the left columns). Tolerances used: $C\alpha$, $C\beta$ and CO range, 2.5, 3 and 3 sigma, respectively; $C\alpha$, $C\beta$ and CO match 0.1, 0.2 and 0.1 ppm respectively. *White fields* in the second columns

show SAGA's assignments corresponding to the literature assignment. *Black fields* in the second or third columns show alternative assignments. The middle numbers in these fields show the fraction of those assignments. The last number gives the number of rung connectivities between the GSs. The amino acid sequence is represented as an "NMR" sequence. Duplicate assignments have been removed on the basis of frequency and/or number of connecting rungs, except for D198 (*italic*)

M 1		A 81	X913 1	A 161	A161 1 2	F 241		L 321	
S 2		V 82	X911 1 1	G 162	G162 1 2	I 242	X935 0.41 1	R 322	
K 3		V 83	X947 0.76	T 163	T163 1 2	A 243	A243 1 1	D 323	D323 1 3
G 4	X978 0.62	Q 84		I 164	I164 1 3	E 244	E244 1 3	A 324	A324 1 3
P 5		S 85		A 165	A165 1	F 245	F245 1 2	K 325	K325 1 1
A 6	A6 1 3	D 86	X937 0.61	G 166		K 246	K246 1	L 326	X963 1
V 7	V7 1 3	M 87		L 167	X928 0.27	R 247		D 327	
G 8	G8 1	K 88	K88 1 3	N 168		K 248	K248 1 1	K 328	
I 9		H 89	H89 1 3	V 169		H 249	H249 1 2	S 329	S329 1 3
D 10		W 90	W90 1	L 170		K 250	K250 1 3	Q 330	Q330 1 3
L 11		P 91		R 171	R171 1 3	K 251	K251 1 1	I 331	I331 1
G 12		F 92	F92 1 3	I 172	I172 1 3	D 252	D252 1	H 332	
T 13		M 93	M93 1 3	I 173	I173 1 3	I 253		D 333	
T 14		V 94	V94 1 3	N 174	N174 1	S 254	S254 1 1	I 334	
Y 15		V 95	V95 1 3	E 175		E 255	E255 1 3	V 335	V335 1 1
S 16		N 96	N96 1 3	P 176		N 256	N256 1	L 336	L336 1 3
C 17		D 97	D97 1	T 177		K 257		V 337	V337 1 3
V 18	X954 1	A 98	X976 1	A 178	X974 1	R 258	R258 1 3	G 338	G338 1
G 19	G19 1 2	G 99		A 179		A 259	A259 1 3	G 339	
V 20	V20 1 3	R 100	X948 0.53	A 180	A180 1	V 260	V260 1	S 340	
F 21	F21 1 3	P 101		I 181	I181 1 1	R 261	X985 0.51 1	T 341	
Q 22	Q22 1	K 102	K102 1 3	A 182	A182 1 3	R 262	R262 1 3	R 342	R342 1 3
H 23	X972 0.75	V 103	V103 1 3	Y 183	Y183 1 3	L 263	L263 1 3	I 343	I343 1
G 24		Q 104	Q104 1 3	G 184	G184 1 1	R 264	R264 1 3	P 344	
K 25		V 105	V105 1	L 185	L185 1 3	T 265	T265 1 3	K 345	
V 26	V26 1 3	E 106		D 186	D186 1 3	A 266	A266 1 2	I 346	
E 27	E27 1 3	Y 107	Y107 1 3	K 187	K187 1 3	C 267	C267 1	Q 347	X984 0.51 1
I 28	I28 1 3	K 108	K108 1 3	K 188	K188 1	E 268		K 348	K348 1 3
I 29	I29 1 3	G 109	G109 1 2	V 189		R 269		L 349	L349 1 3
A 30	A30 1	E 110	E110 1 3	G 190	G190 1 1	A 270		L 350	L350 1 3
N 31		T 111	T111 1 3	A 191	A191 1 3	K 271		Q 351	Q351 1 3
D 32		K 112	K112 1 3	E 192	E192 1 3	R 272		D 352	D352 1 3
Q 33		S 113	S113 1 2	R 193	R193 1 3	T 273		F 353	F353 1
G 34	X981 0.55 1	F 114	F114 1 3	N 194	N194 1 2	L 274		F 354	
N 35	X931 0.98 2	Y 115	Y115 1	V 195	X960 1	S 275		N 355	
R 36	X914 1	P 116		L 196		S 276		G 356	
T 37		E 117	E117 1 3	I 197		S 277		K 357	
T 38	X991 0.53	E 118	E118	F 198	F198 1 2	T 278		E 358	
P 39		V 119		D 199	D199 1	Q 279	Q279 1 3	L 359	L359 1 3
S 40	S40 1 3	S 120		L 200		A 280	A280 1 3	N 360	N360 1 2
Y 41	Y41 1 3	S 121	S121 1 2	G 201	X920 0.58	S 281	S281 1 2	K 361	K361 1
V 42	V42 1 3	M 122	M122 1 2	G 202		I 282	I282 1 3	S 362	
A 43	A43 1 3	V 123	V123 1	G 203		E 283	E283 1 3	I 363	
F 44	F44 1 3	L 124		T 204		I 284	X925 1 2	N 364	
T 45	T45 1 3	T 125	X955 1 3	F 205	F205 1 3	D 285	X910 1 3	P 365	
D 46	D46 1 2	K 126	X957 0.82	D 206	D206 1 2	S 286	X952 1	D 366	D366 1 3
T 47	T47 1 1	M 127	X945 0.5	V 207	V207 1	L 287		E 367	E367 1
E 48	E48 1 2	K 128		S 208		Y 288		A 368	
R 49	R49 1 3	E 129		I 209	I209 1 3	E 289	E289 1 2	V 369	
L 50	L50 1 3	I 130		L 210	L210 1 3	G 290	G290 1 1	A 370	
I 51	I51 1 3	A 131	X971 1 2	T 211	T211 1 3	I 291	I291 1 3	Y 371	
G 52	G52 1	E 132	E132 1 2	I 212	I212 1 3	D 292	D292 1 3	G 372	
D 53		A 133	A133 1 3	E 213	E213 1 2	F 293	X918 1 3	A 373	
A 54		Y 134	Y134 1 1	D 214	D214 1 1	Y 294	X1002 1	A 374	X973 0.5 2
A 55	A55 1 3	L 135	L135 1 1	G 215	G215 1 1	T 295		V 375	X904 0.5
K 56	K56 1 3	G 136	G136 1 1	I 216	I216 1 3	S 296		Q 376	Q376 1 1
N 57	N57 1 3	K 137	K137 1	F 217	F217 1 3	I 297		A 377	A377 1 1
Q 58	Q58 1 3	T 138		E 218	E218 1 3	T 298	T298 1 2	A 378	A378 1 3
V 59	V59 1 2	V 139	V139 1 3	V 219	V219 1 3	R 299	R299 1 3	I 379	I379 1 3
A 60	A60 1 3	T 140	T140 1 2	K 220	K220 1 3	A 300	A300 1 2	L 380	L380 1 3
M 61	M61 1 2	N 141	N141 1 3	S 221	S221 1 2	R 301	R301 1 3	S 381	S381 1 3
N 62	N62 1	A 142	A142 1	T 222	T222 1 2	F 302	F302 1	G 382	G382 1
P 63		V 143		A 223	X903 1	E 303		D 383	
T 64	T64 1 2	V 144		G 224		E 304	E304 1 2	E 384	
N 65	N65 1 3	T 145	T145 1 1	D 225		L 305	L305 1 3	S 385	
T 66	T66 1 3	V 146	V146 1	T 226		N 306	N306 1 3	E 386	X980 0.52
V 67	V67 1 3	P 147		H 227		A 307	A307 1 3		
F 68	F68 1	A 148		L 228	L228 1 3	D 308	D308 1		
D 69		Y 149	X992 1	G 229	G229 1 1	L 309			
A 70		F 150	F150 1 2	G 230	G230 1	F 310	F310 1 2		
K 71		N 151	N151 1	E 231		R 311	R311 1 1		
R 72	X930 0.5 3	D 152		D 232		G 312	G312 1 1		
L 73	X909 0.5	S 153	X986 0.23 1	F 233	X917 1 3	T 313	T313 1		
I 74		Q 154	X969 0.21	D 234	D234 1 3	L 314			
G 75	X993 1	R 155		N 235	N235 1 3	D 315			
R 76	R76 1 3	O 156	X946 1 3	R 236	R236 1 3	P 316			
R 77	R77 1	A 157	X902 1 2	M 237	M237 1	V 317			
F 78		T 158	X999 1 1	V 238		E 318			
D 79	X924 1 3	K 159	K159 1 3	N 239		K 319			
D 80	X943 1 3	D 160	D160 1 3	H 240		A 320			

Fig. 11 Improvement of the assignment of Hsc70 NBD (386 residues). The 206 hand assigned GSs were constrained to their positions, and an additional 101 GSs present in the spectra (mostly with incomplete rungs) were allowed to fill in the open stretches using default tolerances (see [Methods](#)). 46 of these were placed with high confidence. They are shown in the *grey fields*. The middle numbers in the fields show the fraction of those assignments. The last number

gives the number of rung connectivities between the GSs. Duplicate assignments have been removed on the basis of frequency and/or number of connecting rungs. The final result is the assignment of 249 out of the 374 assignable residues, based on 120 GSs with 3-rung connectivities, 37 GSs with 2-rung connectivities, and 28 GSs with single-rung connectivities

pick tables. Currently, the Sparky input mode of SAGA already accepts intensity data. In future enhancements of the program, we envision using this information to weigh against assignments that place GSs of very different intensities next to each other. This is to encapsulate the common knowledge that intensity differences in (a single class of) triple resonance spectra are due to dynamical processes, which, as argued before, often occur in stretches rather than at isolated residues. ^{13}C -line width differences are also mostly determined by dynamics, and can potentially distinguish between correct and wrong GS linkages even when the chemical shifts are identical. Elegant computer-assisted hand-assignment programs such as XEASY have long incorporated an easy visualization of this parameter (Bartels et al. 1995). Coding it into SAGA will be achieved in future versions, and will likely improve the reliability of GS linkages where only few rungs are available.

Conclusions

SAGA is a versatile program for automatic sequential assignment that can handle not only small proteins with high quality data, but even the largest feasible proteins with realistic, flawed data. No single search algorithm is optimal for all datasets, so the branch-and-bound search gives a very thorough search on rather unambiguous data, while the greedy search produces very useful results on large proteins with lower quality data. In real applications, many different assignments satisfy reasonable acceptance criteria, so SAGA summarizes them all, highlighting consistently assigned residues, seldom assigned residues, and different alternative assignments for parts of the polypeptide chain.

Acknowledgments E.R.P.Z. acknowledges support from NIH grants GM063027 and GM063027-08S1 (E.R.P.Z., PI) and NS059690 (J.E. Gestwicki, PI). We thank Drs. A.V. Kurochkin and D.S. Weaver for the preparation of the Hsc70 NMR samples.

References

- Atreya HS, Szyperski T (2004) G-matrix fourier transform NMR spectroscopy for complete protein resonance assignment. *Proc Natl Acad Sci USA* 101:9642–9647
- Baldus M (2007) ICMRBS founder's medal 2006: biological solid-state NMR, methods and applications. *J Biomol NMR* 39:73–86
- Bartels C, Xia T, Billeter M, Güntert P, Wüthrich K (1995) The program XEASY for computer-supported NMR spectral-analysis of biological macromolecules. *J Biomol NMR* 6:1–10
- Benod C, Delsuc M-A, Pons J-L (2006) CRAACK: consensus program for NMR amino acid type assignment. *J Chem Inf Model* 46:1517–1522
- Billeter M, Wagner G, Wüthrich K (2008) Solution NMR structure determination of proteins revisited. *J Biomol NMR* 42(3):155–158
- Bron C, Kerbosch J (1973) Finding all cliques of an undirected graph. *Comm ACM* 16:575–577
- Cavanagh J, Fairbrother WJ, Palmer AG, Rance M, Skelton NJ (2007) *Protein NMR spectroscopy*, 2nd edn. Academic Press, Amsterdam
- Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A (1995) NMRPipe: a multidimensional spectral processing system based on UNIX pipes. *J Biomol NMR* 6:277–293
- Eghbalian HR, Bahrami A, Wang L, Assadi A, Markley JL (2005) Probabilistic identification of spin systems and their assignments including coil-helix inference as output (PISTACHIO). *J Biomol NMR* 32:219–233
- Friedrichs MS, Mueller L, Wittekind M (1994) An automated procedure for the assignment of protein ^1HN , ^{15}N , $^{13}\text{C}^\alpha$, $^1\text{H}^\alpha$, $^{13}\text{C}^\beta$ and $^1\text{H}^\beta$ resonances. *J Biomol NMR* 4:703–726
- Frueh DP, Arthanari H, Koglin A, Walsh CT, Wagner G (2009) A double TROSY hNCAnH experiment for efficient assignment of large and challenging proteins. *J Am Chem Soc* 131:12880–12881
- Goddard TD, Kneller DG (2003) *Sparky-NMR assignment and integration software*. University of California, San Francisco
- Güntert P (2009) Automated structure determination from NMR spectra. *Eur Biophys J* 38(2):129–143
- Hare BJ, Prestegard JH (1994) Application of neural networks to automated assignment of NMR spectra of proteins. *J Biomol NMR* 4:35–46
- Hopcroft JE, Karp RM (1973) An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM J Comput* 2:225–231
- Hyberts SG, Wagner G (2003) IBIS—A tool for automated sequential assignment of protein spectra from triple resonance experiments. *J Biomol NMR* 26:335–344
- Kao M-Y, Lam T-W, Sung W-K, Ting H-F (2001) A decomposition theorem for maximum weight bipartite matchings. *SIAM J Comput* 31:18–26
- Kuhl FS, Crippen GM, Friesen DK (1984) A combinatorial algorithm for calculating ligand binding. *J Comput Chem* 5:24–34
- Kuhn HW (1955) Variants of the hungarian method for assignment problems. *Naval Research Logistics Quarterly* 2:83–97
- Lemak A, Steren CA, Arrowsmith CH, Llinás M (2008) Sequence specific resonance assignment via multicanonical Monte Carlo search using an ABACUS approach. *J Biomol NMR* 41:29–41
- Leutner M, Gschwind RM, Liermann J, Schwarz C, Gemmecker G, Kessler H (1998) Automated backbone assignment of labeled proteins using the threshold accepting algorithm. *J Biomol NMR* 11:31–43
- Lin G, Wan X, Tegos T, Li Y (2006) Statistical evaluation of NMR backbone resonance assignment. *Int J Bioinf Res App* 2:147–160
- Lukin JA, Gove AP, Talukdar SN, Ho C (1997) Automated probabilistic method for assigning backbone resonances of (^{13}C , ^{15}N)-labeled proteins. *J Biomol NMR* 9:151–166
- Meadows RP, Olejniczak ET, Fesik SW (1994) A computer-based protocol for semiautomated assignments and 3D structure determination of proteins. *J Biomol NMR* 4:79–96
- Moseley HNB, Monleon D, Montelione GT (2001) Automatic determination of protein backbone resonance assignments from triple resonance nuclear magnetic resonance data. *Meth in Enzymol* 339:91–108
- Nakada S, Sakakura M, Takahashi H, Tokuda H, Shimada I (2007) Backbone resonance assignment for the outer membrane lipoprotein receptor LolB from *Escherichia coli*. *Biomol NMR Assign* 1:121–123
- Olson JB Jr, Markley JL (1994) Evaluation of an algorithm for the automated sequential assignment of protein backbone resonances:

- A demonstration of the connectivity tracing assignment tools (CONTRAST) software package. *J Biomol NMR* 4:385–410
- Oschkinat H, Croft D (1994) Automated assignment of multidimensional nuclear magnetic resonance spectra. *Meth Enzymol* 239:308–318
- Parker RG, Rardin RL (1988) *Discrete Optimization*. Academic Press, New York
- Revington M, Zuiderweg ERP (2004) TROSY-driven NMR backbone assignments of the 381-residue nucleotide-binding domain of the *Thermus Thermophilus* DnaK molecular chaperone. *J Biomol NMR* 30:113–114
- Revington M, Zhang Y, Yip GN, Kurochkin AV, Zuiderweg ERP (2005) NMR investigations of allosteric processes in a two-domain *Thermus thermophilus* Hsp70 molecular chaperone. *J Mol Biol* 349:163–183
- Seavey BR, Farr EA, Westler WM, Markley J (1991) A relational database for sequence-specific protein NMR data. *J Biomol NMR* 1:217–236
- Stratmann D, Guittet E, van Heijenoort C (2010) Robust structure-based resonance assignment for functional protein studies by NMR. *J Biomol NMR* 46:157–173
- Tugarinov V, Muhandiram R, Ayed A, Kay LE (2002) Four-dimensional NMR spectroscopy of a 723-residue protein: chemical shift assignments and secondary structure of malate synthase G. *J Am Chem Soc* 124:10025–10035
- Vitek O, Bailey-Kellogg C, Craig B, Kuliniewicz P, Vitek J (2005) Reconsidering complete search algorithms for protein backbone NMR assignment. *Bioinformatics* 21:230–236
- Wan X, Lin G (2007) GASA: A graph-based automated NMR backbone resonance sequential assignment program. *J Bioinf Comput Biol* 5:313–333
- Wang J, Wang T, Zuiderweg ERP, Crippen GM (2005) CASA: An efficient automated assignment of protein mainchain NMR data using an ordered tree search algorithm. *J Biomol NMR* 33:261–279
- Williamson M, Craven C (2009) Automated protein structure calculation from NMR data. *J Biomol NMR* 43:131–143
- Xiong F, Pandurangan G, Bailey-Kellogg C (2008) Contact replacement for NMR resonance assignment. *Bioinformatics* 24:205–213
- Zimmerman DE, Montelione GT (1995) Automated analysis of nuclear magnetic resonance assignments for proteins. *Curr Opin Struct Biol* 5:664–673
- Zimmerman DE, Kulikowski CA, Montelione GT (1993) A constraint reasoning system for automating sequence-specific resonance assignments from multidimensional protein NMR spectra. *Proc Int Conf Intell Syst Mol Biol* 1:447–455
- Zimmerman D, Kulikowski C, Wang L, Lyons B, Montelione GT (1994) Automated sequencing of amino acid spin systems in proteins using multidimensional HCC(CO)NH-TOCSY spectroscopy and constraint propagation methods from artificial intelligence. *J Biomol NMR* 4:241–256
- Zimmerman DE, Kulikowski CA, Huang Y, Feng W, Tashiro M, Shimotakahara S, Chien CY, Powers R, Montelione GT (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. *J Mol Biol* 269:592–610