J Biomol NMR (2014) 58:155–163
DOI 10.1007/s10858-014-9813-7

ARTICLE

# RASP: rapid and robust backbone chemical shift assignments from protein structure

**Christopher A. MacRaild · Raymond S. Norton**

Received: 9 January 2014 / Accepted: 15 January 2014 / Published online: 21 January 2014


**Abstract** Chemical shift prediction has an unappreciated power to guide backbone resonance assignment in cases where protein structure is known. Here we describe Resonance Assignment by chemical Shift Prediction (RASP), a method that exploits this power to derive protein backbone resonance assignments from chemical shift predictions. Robust assignments can be obtained from a minimal set of only the most sensitive triple-resonance experiments, even for spectroscopically challenging proteins. Over a test set of 154 proteins RASP assigns 88 % of residues with an accuracy of 99.7 %, using only information available from HNCO and HNCA spectra. Applied to experimental data from a challenging 34 kDa protein, RASP assigns 90 % of manually assigned residues using only 40 % of the experimental data required for the manual assignment. RASP has the potential to significantly accelerate the backbone assignment process for a wide range of proteins for which structural information is available, including those for which conventional assignment strategies are not feasible.

**Keywords** Chemical shift prediction · Structure-based resonance assignment · Large protein NMR

**Electronic supplementary material** The online version of this article (doi:10.1007/s10858-014-9813-7) contains supplementary material, which is available to authorized users.

C. A. MacRaild (✉) · R. S. Norton
Medicinal Chemistry, Monash Institute of Pharmaceutical Sciences, Monash University, 381 Royal Parade, Parkville 3052, Australia
e-mail: chris.macraild@monash.edu

## Introduction

NMR spectroscopy is frequently the method of choice for the study of protein structure, interactions and dynamics (Mittermaier and Kay 2009), and is a valuable tool for mapping ligand binding in drug development programs (Lepre et al. 2004). The assignment of spectral frequencies (chemical shifts) to specific atoms of the protein is a prerequisite for these analyses. Traditionally, such assignments have required manual analysis of extensive experimental datasets, imposing substantial costs in both labor and instrument time, and have represented a significant bottle-neck in protein NMR studies. Accordingly, much effort has been directed towards the automation of the assignment process. Progress to this end has meant that routine, fully automated spectral assignment of small proteins is now achievable (Schmidt and Güntert 2012). However, current approaches do not scale well to larger or more challenging proteins, where the available data are frequently ambiguous and incomplete.

Increasingly, the goal of protein NMR is not to determine structures, but rather to gain functional insights into systems for which structural information is already available (Barrett et al. 2013). Of the 47 assignment notes describing new protein assignments in a recent issue of *Biomolecular NMR Assignments* (Volume 7, issue 2), a crystal structure was available in the PDB (Berman et al. 2000) for at least 21 targets. Accordingly, several methods have been proposed that aim to exploit available structural information to assist in the resonance assignment problem (Langmead and Donald 2004; Stratmann et al. 2010). Typically, these strategies aim to find the assignment that best matches observed networks of NOEs with those expected from a protein structure, with additional restraints derived from other structural observables such as residual

dipolar couplings (RDCs) or paramagnetic effects. To date, none of these approaches has offered a compelling alternative to conventional triple-resonance assignment strategies, for at least two reasons: they have been limited to smaller proteins for which conventional approaches are straightforward, and they rely on the measurement of structural parameters that are not necessarily of primary interest in cases where protein structure is already well characterized.

Concurrently, significant effort has focused on improving the accuracy of chemical shift prediction from protein structure (Han et al. 2011; Kohlhoff et al. 2009; Shen and Bax 2010). Progress in this area has enabled the use of assigned chemical shift information to drive structure determination (Cavalli et al. 2007; Shen et al. 2008; Wishart et al. 2008). Despite some early efforts (Gronwald et al. 1998), little attention has been directed at the inverse problem: the use of structure-based chemical shift predictions to guide resonance assignment. Here, we show that chemical shift prediction has an unappreciated power to guide backbone resonance assignment in cases where protein structure is known. We exploit this power to develop Resonance Assignment by chemical Shift Prediction (RASP), a structure-based assignment strategy that utilizes conventional triple resonance experiments to aid the assignments of challenging protein targets.

## Methods

### Test set

A test set comprising matched pairs of high-resolution crystal structures and assigned NMR chemical shifts was extracted from the database distributed with TALOS (Shen et al. 2009). In its construction, the TALOS database was filtered to remove assignments to residues with unusually high crystallographic B-factor and those with extreme outlier chemical shifts. This results in numerous stretches missing chemical shift data, mimicking the incompleteness common in experimental datasets of challenging proteins. We excluded from this dataset those proteins for which assignments are missing for more than 40 % of crystallographically resolved residues, those for which experimental amide shifts were absent, and those for which the experimental conditions for the NMR experiments obviously deviated from those in the crystal structure in a significant way. For example, we removed TALOS ID 4568, which paired the assigned chemical shifts of acid-denatured apomyoglobin (BMRB ID 4568) with an X-ray crystal structure of a mutant (but folded) form of the holoprotein (PDB ID 1DTI). The result is a set of 154 proteins comprising on average 139 spin systems predicted from the crystal structure (a range of 52–615) and 114 experimental spin systems (32–495) (Table S1).

Chemical shift predictions for the test set were made with Sparta+ (version 2.50F1 Rev 2011.108.15.55) (Shen and Bax 2010), ShiftX2 (version 1.07) (Han et al. 2011) and CamShift (version 1.35) (Kohlhoff et al. 2009). Predictions used the default settings for each predictor, except for ShiftX2, where we excluded results from ShiftY, as most members of the test set are present in the BMRB database and could thus favourably bias ShiftY predictions. No attempt was made to correct for temperature, pH or deuteration in the shift predictions of the test set. By way of comparison, a set of structure-independent shift predictions was derived by taking residue-specific average chemical shifts from the RefDB database (Zhang et al. 2003). Preliminary tests showed the performance of CamShift to be slightly inferior to that of the other predictors in this application, so it was not used further. Sparta+ and ShiftX2 performed almost identically, on average, over the test set, and we report only the results for Sparta+ predictions here. In real applications, however, we advocate the use and close comparison of several chemical shift predictors, as performance varies to some extent on a protein-by-protein basis.

### Comparing predicted and experimental chemical shifts

We use a weighted distance measure to compare the set of chemical shifts comprising an experimental spin system $i$ with the corresponding set of predicted shifts for residue $r$ in the target protein structure:

$$\mathbf{D}_{i,r} = \sqrt{\sum_X \left( \frac{X_{expt} - X_{pred}}{w_X} \right)^2} \quad (1)$$

where the sum is over all shifts X common to spin system $i$ and residue $r$, including $i-1$ $(r-1)$ shifts. The weighting terms $w_X$ are chosen to account for the varying precision with which each shift type can be predicted. Specifically, for shifts predicted by Sparta+ we use the estimated prediction error for each predicted shift. For ShiftX2 and CamShift, we use the prediction RMSDs for each shift type, as reported for the 61 protein test set of Han et al. (2011).

In an attempt to account for the fact that NMR spectra are not uniformly populated, and that spin systems in densely populated spectral regions will tend to be closer to a larger number of predicted shifts, purely by chance, we normalize **D** such that the average shift distance for each spin system is 1, to yield **N**:

$$\mathbf{N}_{i,r} = \frac{n\boldsymbol{D}_{i,r}}{\sum_s^n \boldsymbol{D}_{i,s}} \quad (2)$$

where $n$ is the number of residues. As shown in Fig. 1b, small values of **N** are strongly predictive of assignment accuracy, and a simple empirical function,

$$\mathbf{L}_{i,r} = 0.9\left(1 - \frac{\mathbf{N}_{i,r}^{3.5}}{\mathbf{N}_{i,r}^{3.5} + 0.001}\right) \quad (3)$$

approximates the relationship between the normalised shift distance and the likelihood of a correct assignment. This serves as the first term of the RASP scoring function, Eq. 1.

For the second term, $\mathbf{S}_{i,j}$, we score the agreement between the common shifts $X_j$ and $X_{i-1}$ that support the sequential relationship between spin systems $i$ and $j$. Because this entails the comparison of sets of experimental chemical shifts (rather than comparison of predicated shifts with experimental shifts, as in **L**) we adopt a more direct and stringent scoring function:

$$\mathbf{S}_{i,j} = \prod_X cdf\left(\left|X_{i-1} - X_j\right|\right) \quad (4)$$

where *cdf*() is the normal cumulative distribution function with variance of 0.1 ppm for CA and CO chemical shifts, 0.2 ppm for CB, 0.06 for N and 0.03 for any proton shift. Our choice of these variances was based on the achievable precision for each shift type in conventional triple-resonance spectra of large proteins, taking into account spectral resolution, line shape and potential for overlap. The resulting scoring function does not penalise the absence of any particular shift in either spin system.

## The RASP scoring function

The RASP algorithm casts chemical shift assignment as a combinatorial optimisation problem: it seeks assignments that maximise the sum of two terms, the first of which scores the agreement between the predicted and experimental chemical shifts and the second of which scores agreement between equivalent pairs of shifts from sequentially assigned spin systems:

$$\sum_r \left[\mathbf{L}_{\pi(r),r} + \mathbf{S}_{\pi(r),\pi(r-1)}\right] \quad (5)$$

Here a permutation of residues $r$ is denoted $\pi$, where $\pi$ expresses the assignment such that $\pi(r) = i$ denotes the assignment of spin system $i$ to residue $r$.

## The RASP optimisation heuristic

The RASP optimisation strategy is based on the greedy randomised adaptive search procedures (GRASP) metaheuristic (Resende and Ribeiro 2010). GRASP offers effective optimisation of a wide range of combinatorial optimisation problems with a minimum of adjustable parameters, and showed superior performance when applied to our test set, compared to a range of other optimisation strategies, including simulated annealing and branch-and-bound search strategies (data not shown). Our implementation is described in detail in the Supplementary Methods. Briefly, approximate solutions are constructed by weighted random selection from the best-scoring of an exhaustive list of possible assignments to sequential

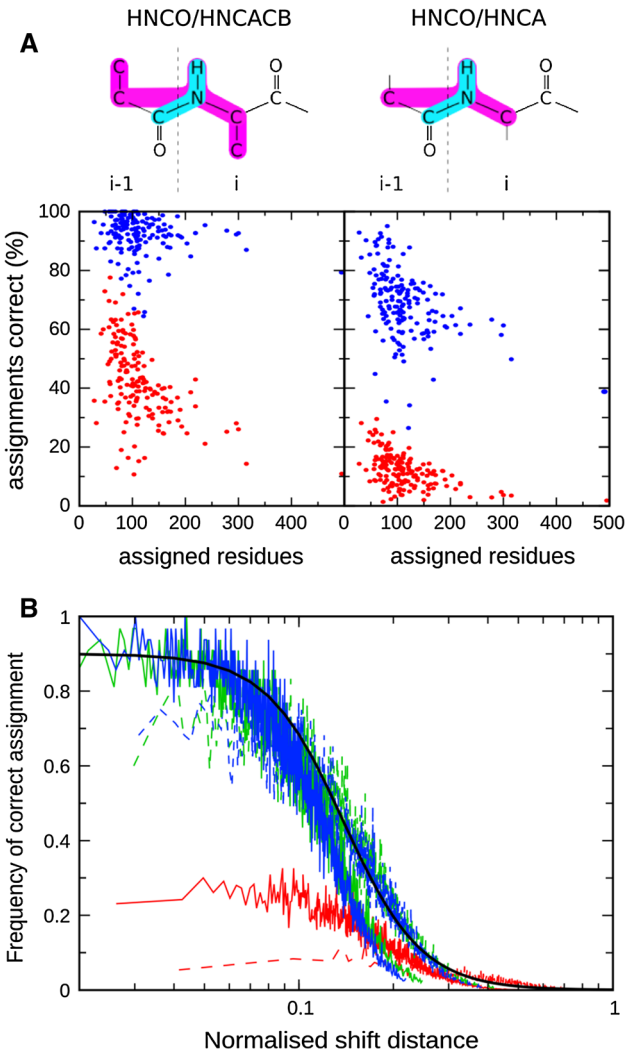



**Fig. 1** **a** Backbone assignments can be inferred simply by identifying the assignment that minimizes the distance between experimental and predicted chemical shifts for spin systems derived from HNCO and HNCACB (*left*) and from HNCO and HNCA (*right*). For each of 154 proteins, the percentage of residues assigned correctly is plotted against the number of residues assigned, using chemical shifts predicted with Sparta+ (*blue*), or with structure-independent average chemical shifts from RefDB (*red*). **b** The normalised shift distance between spin system $i$ and residue $r$, predicts the likelihood of correct assignment of $i$ to $r$. The frequency of correct assignment $i$ to $r$ is plotted against the normalised shift distance $\mathbf{N}_{i,r}$ for chemical shift predictors Sparta+ (*blue*) and Shiftx2 (*green*) or structure-independent average chemical shifts from RefDB (*red*), using spin systems comprising HNCA/HNCO shifts (*dashed lines*) and HNCACB/HNCO shifts (*solid lines*). Equation 3 is plotted as a *heavy black line*

stretches of three residues ('triplets'). The resulting assignment is improved by a systematic search for a local maximum in the scoring function by pairwise swaps of residue assignments, accepting each swap that improves the assignment score.

### Ketopantoate reductase

Chemical shift predictions for ketopantoate reductase (KPR) were made from the 1.7 Å crystal structure of apo KPR (PDB id. 1KS9) (Matak-Vinkovic et al. 2001) using Sparta+ (Shen and Bax 2010) and ShiftX2 (Han et al. 2011) as for the test set, except that, for KPR, shift predictions were run with appropriate corrections for the deuteration, pH and temperature conditions used for acquisition of the NMR data (Headey et al. 2008). In preliminary tests, RASP calculations against the ShiftX2 shift predictions yielded superior assignment coverage, so these predictions were used for all of the results presented here.

MARS (Jung and Zweckstetter 2004) was run against the initial and final sets of KPR spin systems using a fragment size of 5 and chemical shift cutoffs of 0.2 ppm for CA and 0.5 ppm for CB. Secondary structure was input as defined in PDB record 1KS9. PINE (Bahrami et al. 2009) was run against these input spin systems using the PINE web-server (http://pine.nmrfam.wisc.edu/).

## Results and discussion

### Predicted chemical shifts are powerful aids to backbone assignment

The accuracy of chemical shift prediction is currently insufficient to be of use for the direct assignment of individual shifts. However neither the backbone chemical shifts of individual residues, nor their associated prediction errors, are strongly correlated (Fig. S1). This lead us to reason that grouping shift predictions by residue may offer a means to increase the discriminatory power of these predictions, and make them useful for guiding the assignment problem. To test this hypothesis, we assemble experimental shifts into generalized spin systems following the approach of conventional triple-resonance assignment strategies used almost universally for backbone assignment. For example, the HNCO experiment correlates the amide $^{1}$H and $^{15}$N chemical shifts of residue $i$ with the $^{13}$CO shift of residue $i-1$, while the HNCA experiment correlates the amide shifts of residue $i$ with the $^{13}$CA shifts of residues $i$ and $i-1$. Thus, the spin systems assembled from the HNCO and HNCA experiments comprise the shifts [$^{1}H_i$, $^{15}N_i$, $^{13}CO_{i-1}$, $^{13}CA_i$, $^{13}CA_{i-1}$] (Fig. 1a).

We initially use a simple weighted distance measure, **D** (Eq. 1), to compare the set of chemical shifts comprising the experimental spin system $i$ with the corresponding set of predicted shifts for residue $r$ in the target protein structure. A simplistic assignment strategy, then, is to find the assignment of spin systems to residues that minimizes the total shift distance by application of the Hungarian algorithm (Kuhn 1955) to the matrix of distances **D**. We tested this strategy against a data set drawn from the TALOS database of assigned chemical shifts matched with high-quality X-ray crystal structures (Shen et al. 2009). The results of this naïve approach are summarized in Fig. 1a. Remarkably, it yields essentially complete and correct assignments for a number of smaller proteins in our test set, and correct assignments for more than 75 % of assignable residues in even the largest protein when CB shifts (from an HNCACB, for example) are considered. Although this remains well short of the accuracy required for a practically useful assignment strategy, it suggests that the potential of chemical shift predictions to guide the assignment process has perhaps been underestimated. It is of particular note that, whereas CB chemical shifts are strictly required for triple-resonance assignment by conventional means, Fig. 1a shows that, even when shifts from only the HNCO and HNCA experiments are considered, shift predictions retain significant power to guide assignment, with an average of 67 % of possible assignments made correctly by this approach.

To highlight the specific role that structure-based information plays in these results, we assemble a set of structure-independent chemical shift predictions, in the form of residue-type specific average chemical shifts from the RefDB database (Zhang et al. 2003). Equivalent statistical chemical shift data are used in some form by essentially all conventional manual or automatic assignment strategies. When the simple assignment approach described above is applied using these shift predictions, we achieve an average accuracy of 12 % in the absence of CB shifts, or 43 % when they are included (Fig. 1a). This result confirms that significant additional information is available in structure-based chemical shift predictions. Thus, we seek to exploit this information to improve the accuracy and efficiency of chemical shift assignment.

To further explore the potential of predicted chemical shifts in backbone assignment, we asked to what extent good agreement between experimental and predicted shifts for any given residue could predict the likelihood that the corresponding assignment was correct. We first normalise the shift distances **D** as described in Methods, above, to account for the fact that chemical shifts are not uniformly distributed, and that spin systems in densely populated regions of NMR spectra will tend to be closer to a larger number of predicted shifts purely by chance. Figure 1b

shows that small values of the normalised shift distance **N** are strongly predictive of an accurate assignment. Moreover, the relationship between **N** and assignment accuracy is relatively insensitive to the composition of the experimental spin systems, or to the chemical shift prediction algorithm employed. For this reason, we choose to use this relationship as a scoring function for the assignment algorithm developed below. Once again, conventional structure-independent chemical shift statistics have only moderate predictive power in this context and, as expected, this limited power depends critically on knowledge of CB chemical shifts (Fig. 1b).

## The RASP assignment algorithm

The initial assignment strategy considered above treats the assignment of each spin system independently, ignoring the sequential information encoded in the $X_i/X_{i-1}$ pairs of chemical shifts. This sequential information is the basis of conventional triple-resonance assignment strategies, and can serve as a further constraint on an assignment strategy based on chemical shift predictions. Therefore, we seek to simultaneously optimize our chemical shift based scoring function with the agreement between the common shifts $X_j$ and $X_{i-1}$ for sequentially assigned spin systems $i$ and $j$ as described in Methods (Eqs. 3–5). Solving this optimization problem uniquely is not computationally feasible, even for small proteins. Moreover, a single unique solution is not necessarily desirable, as it gives no insight into the extent to which any specific residue assignment is constrained by the input data. Instead, we seek to sample diverse near-optimal solutions, yielding an ensemble of possible assignments. In such an ensemble, assignments that are robustly constrained by the data are consistent across the ensemble, while the diversity of assignments at other spin systems is indicative of the assignments that are consistent with the available data. In this sense, our approach is similar to that introduced recently in SAGA (Crippen et al. 2010), where various optimization heuristics yield a collection of plausible assignments for further analysis. An advantage of assignment ensembles of this type lies in their power to suggest further experiments that might resolve any remaining ambiguities in the assignment (e.g. further triple resonance experiments to yield additional sequential information, or residue specific labeling or unlabelling schemes (Jaipuria et al. 2012)). Moreover, as described below, assignment ensembles and chemical shift predictions can assist the further analysis of existing spectra, guiding the search for spin systems that may have been misidentified due to overlap, weak signal, or other sources of ambiguity.

To determine an assignment ensemble based on Eq. 5, we have developed a sampling strategy based on the GRASP optimization metaheuristic (Resende and Ribeiro 2010), which efficiently samples near-optimal solutions, even for very large proteins. In developing this strategy, we sought an approach that will assist the semi-automated assignment of large and spectroscopically challenging proteins, noting that robust fully automatic assignment strategies already exist for small proteins (Bahrami et al. 2009; Schmidt and Güntert 2012). For this reason we do not address the problem of spin-system assembly: this process has been effectively automated for relatively simple systems, but fails for larger or spectroscopically challenging proteins. In such cases, manual spin-system assembly by an experienced spectroscopist is necessary, but relatively straightforward. As we demonstrate below, our algorithm is robust to errors and incompleteness in spin-system assembly. This robustness permits an iterative assignment procedure in which results in initial rounds provide highly reliable information on which to base subsequent refinement of the input spin systems. Iterative assignment of this type is essential to manual assignment approaches, and in practice is common when applying automated assignment algorithms to complex problems.

## The performance of RASP

The resulting algorithm, which we call RASP, has been applied to our test set, yielding an ensemble of approximately 100 possible residue assignments for each of the 17,530 experimental spin systems defined in the test set. When the spin systems comprise only those shifts available from HNCA and HNCO spectra, there are on average seven unique assignment possibilities for each spin system, and the correct assignment is present in the ensemble in 17,491 cases (99.8 %). A single unique residue assignment is found for 3,157 spin systems and 99.9 % of these assignments are correct, while the most frequent residue assignment in the ensemble is correct for 97.8 % of all spin systems. Further improvements in assignment coverage can be achieved by the inclusion of CB chemical shifts (Table 1, Fig S2), although the acquisition of these data (in the form of HNCACB spectra, for example), is significantly more demanding owing to the reduced sensitivity of the corresponding experiments. The capacity of RASP to generate extensive and accurate assignments even in the absence of CB chemical shifts is to our knowledge unique, and potentially enables detailed NMR studies of proteins for which CB chemical shifts are not experimentally accessible.

In assessing the assignment ensembles that result from RASP, it is convenient to define a frequency threshold, $C_f$, such that any residue assignment occurring more frequently in the ensemble than the threshold is regarded as uniquely determined. The choice of $C_f$ reflects a tradeoff

**Table 1** RASP assignments of the test set

| Input spin systems | Assignment ensemble diversity (accuracy)[a] | Accuracy of most frequent assignment (%) | Accuracy (coverage) at $C_f = 0.7$[b] | Accuracy (coverage) at $C_f = 0.9$[b] |
|---|---|---|---|---|
| HNCA/HNCO | 7.3 (99.8 %) | 97.8 | 99.7 % (88 %) | 99.9 % (61 %) |
| HNCACB/HNCO | 2.1 (99.7 %) | 98.6 | 99.6 % (98 %) | 99.9 % (95 %) |

[a] The average number of unique residue assignments to each spin system, and the proportion of spin systems that include the correct residue assignment in the assignment ensemble

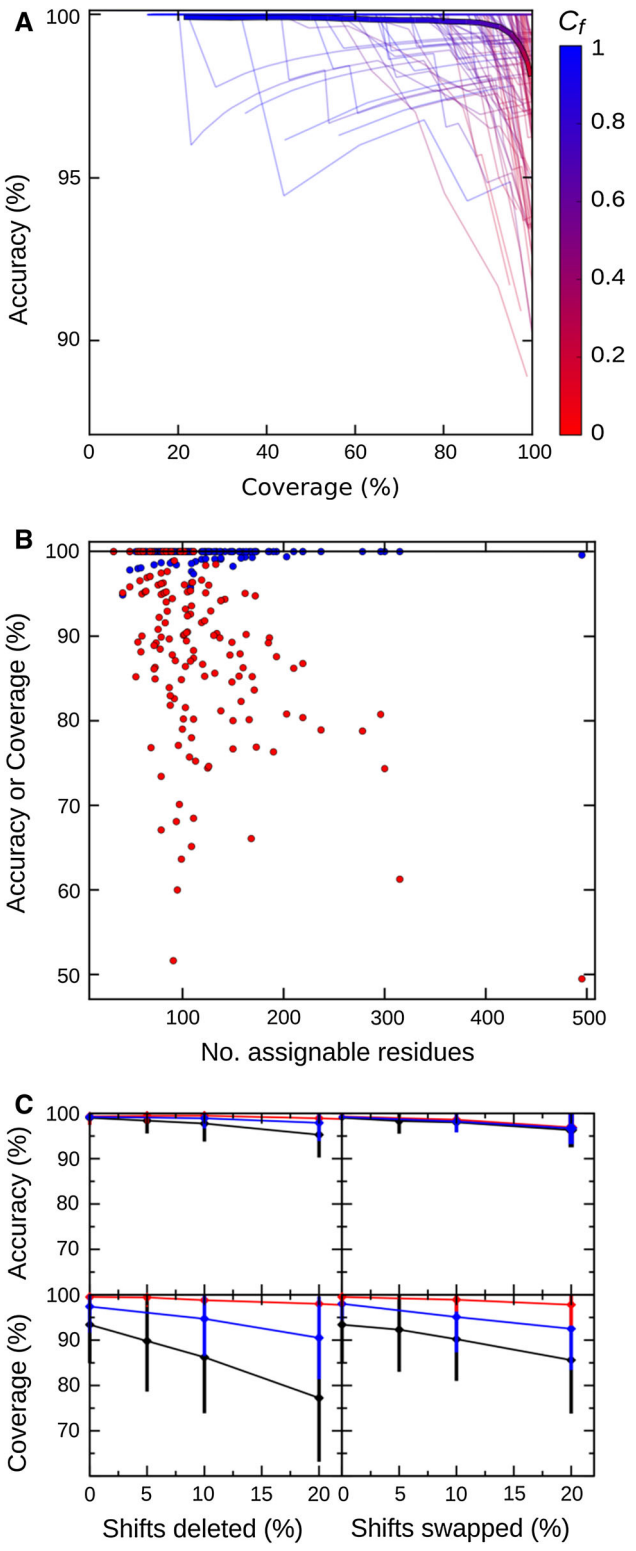
[b] The proportion of assignments more frequent than $C_f$ that are correct, and the proportion of spin systems that are thus assigned

between the number of spin systems uniquely assigned on the one hand, and the accuracy of those assignments on the other. We quantify the former as assignment coverage: the proportion of spin systems for which assignments are made. It is clear from Fig. 2a that the relationship between coverage and accuracy varies somewhat over the proteins of the test set, reflecting variation in the protein size, dataset completeness and other parameters that contribute to the difficulty of the assignment problem. Importantly, however, over a wide range of values of $C_f$, this variation manifests as variation in coverage, rather than accuracy; for a given threshold, assignment accuracy is essentially invariant, while coverage decreases as the assignment difficulty increases (Fig. 2b). Thus, $C_f$ can be chosen to yield a desired accuracy independent of the difficulty of the assignment problem at hand. We find $C_f = 0.7$ to represent an appropriate tradeoff, and use that value in Fig. 2b and in the following. This yields an accuracy of 99.7 % and an average coverage over the test set of 88 % (or 98 %, if CB shifts are considered; Table 1). It is again instructive to compare this performance of RASP using structure-based chemical shift predictions to its performance when relying on structure-independent chemical shift statistics (Fig S2). Using RefDB average chemical shifts and spin systems that include CB shifts, RASP performs reasonably well, with average assignment coverage of 85 % at an accuracy of 99.0 %. When CB shifts are excluded from analysis, however, only 12 % of residues are assigned, with an accuracy of 79 %. These results further highlight the necessity of CB shifts for deriving assignments from shift statistics alone in the absence of structural information.

In its construction, the TALOS database was filtered to remove assignments to residues with unusually high crystallographic B-factor and those with extreme outlier chemical shifts (Shen et al. 2009). This results in numerous stretches lacking chemical shift data, such that on average 18 % of residues in our test set lack any chemical shift data. Of the remaining 82 %, a further 1, 4, and 16 % of residues are missing CA, CB or C shifts, respectively. This level of missing data mimics to a reasonable extent the incompleteness common in experimental datasets of challenging proteins. To further examine the robustness of RASP to incomplete and noisy data, we deliberately degraded the test set by randomly deleting a further fraction of the backbone carbon chemical shifts, or by swapping carbon shifts of the same type between pairs of spin systems (thus mimicking errors in the assembly of spin systems, as may occur, for example, where amide shifts are degenerate). The results of these degradations are shown in Fig. 2c; the dominant effect is a reduction in coverage, with assignment accuracy robust to even this extent of missing and noisy data.

To test the performance of RASP on an experimental dataset, we assembled spin systems from a subset of the spectra recently used to assign KPR, a 34 kDa enzyme of the bacterial pantothenate biosynthetic pathway. KPR exhibits extensive conformational heterogeneity which manifests as peak broadening and duplication; more than 360 backbone amide peaks are identifiable in the $^{1}H,^{15}N$-TROSY experiment, where only 288 are expected on the basis of the KPR sequence (Headey et al. 2008).

Considering the challenging nature of the KPR assignment problem, we chose to use both CA and CB shifts in the assignment process. Thus, spin systems were assembled on the basis of HNCO, HNCA and HNCACB spectra only, with $i$ and $i - 1$ peaks distinguished on the basis of peak intensity. In terms of instrument time, these spectra represent $\sim$40 % of the total dataset required for the manual assignment of KPR. Spin systems were assembled manually from these spectra without reference to the existing assignments or to any other spectra, with the exception of a high-resolution $^{1}H$-$^{15}N$ TROSY which was used to identify potentially overlapped spin systems. From these three spectra 265 spin systems were defined, including 12 for which no CB shifts could be identified, and 61 and 22 for which the $CB_{i-1}$ and $CA_{i-1}$ chemical shifts, respectively, could not be defined unambiguously. Spin systems for which $CA_i$ could not be assigned unambiguously, or those that appeared to be duplications arising from slow conformational exchange, were not included in the calculation. On the basis of this initial set of spin systems, RASP generated 183 residue assignments at $C_f = 0.7$ with three errors. This represents 66 % of the manually assigned residues of KPR at an accuracy of 98.4 %.




◀**Fig. 2** RASP performance on a 154-protein test set. **a** Assignment accuracy and coverage achieved by RASP using spin systems derived from HNCO and HNCA spectra for each protein in the test set varies as a function of the frequency threshold, $C_f$ (*color scale*), used to filter the assignment ensemble. For 63 of 154 proteins accuracy is 100 % for all values of $C_f$. The average over all proteins is plotted in *bold*. **b** RASP assignment accuracy (*blue*) and coverage (*red*) achieved at $C_f = 0.7$ as a function of protein size, using spin systems derived from HNCO and HNCA spectra. **c** Robustness of RASP assignments to incomplete and erroneous spin systems definitions. RASP assignment accuracy (*top panels*) and coverage (*bottom panels*) in the face of random carbon shift deletion (*left panels*) or exchange between spin systems (*right panels*), for spin systems derived from HNCO and HNCA spectra (*black*), HNcaCO and HNCA spectra (*blue*) or HNCO and HNCACB spectra (*red*)

Reasoning that the relatively poor coverage achieved here may reflect errors in the input spin systems, we re-examined them on the basis of this preliminary set of assignments. Assuming the initial set of RASP assignments to be accurate, we sought spin systems for which ambiguity that existed in the initial analysis could now be resolved by virtue of those spin systems being unique matches for gaps in the initial RASP assignments. In this way we found two spin systems that had been excluded from the initial input that were no longer ambiguous, and we found five erroneous spin systems for which the assignment of $i/i-1$ shift pairs had been inverted. We were also able to resolve ambiguities in the assignment of CA or CB shifts in a further 13 existing spin systems. Using this refined set of spin systems as input, we reran RASP, yielding 222 residue assignments (80 % coverage). Subsequent rounds of input refinement resulted in a further nine spin system modifications and six new spin systems, and achieved a final assignment of 237 residues (85 % coverage). This assignment was in agreement with the published assignment (Headey et al. 2008), with three exceptions: residue 34, a residue flanked by prolines and therefore isolated with respect to sequential connectivities, which appears to be misassigned by RASP, and residues 32 and 135, which are assigned by RASP but were not assigned in the published assignment (Fig. 3).

We compared these results to those produced by the widely-used backbone assignment algorithms MARS (Jung and Zweckstetter 2004) and PINE (Bahrami et al. 2009). Against the initial set of manually defined input spin systems, MARS achieves 53, 39 and 42 assignments of high, medium and low confidence, respectively, with 0, 7 (17 %) and 11 (26 %) errors, for a total 48 % coverage at 87 % accuracy, while PINE reports greater than 95 % probability for 224 assignments, with 27 errors. Thus, RASP significantly outperforms MARS in terms of both coverage and accuracy on this initial dataset, while PINE offers better coverage than RASP, although at the expense of a substantial decrease in accuracy. When tested against the final set of spin systems refined against the RASP assignments, MARS assigned 87, 69 and 44 residues with high, medium, and low reliability respectively (72 % total coverage), with 5 medium and 8 low reliability assignments disagreeing with the published assignments (Fig. 3). PINE reports 249

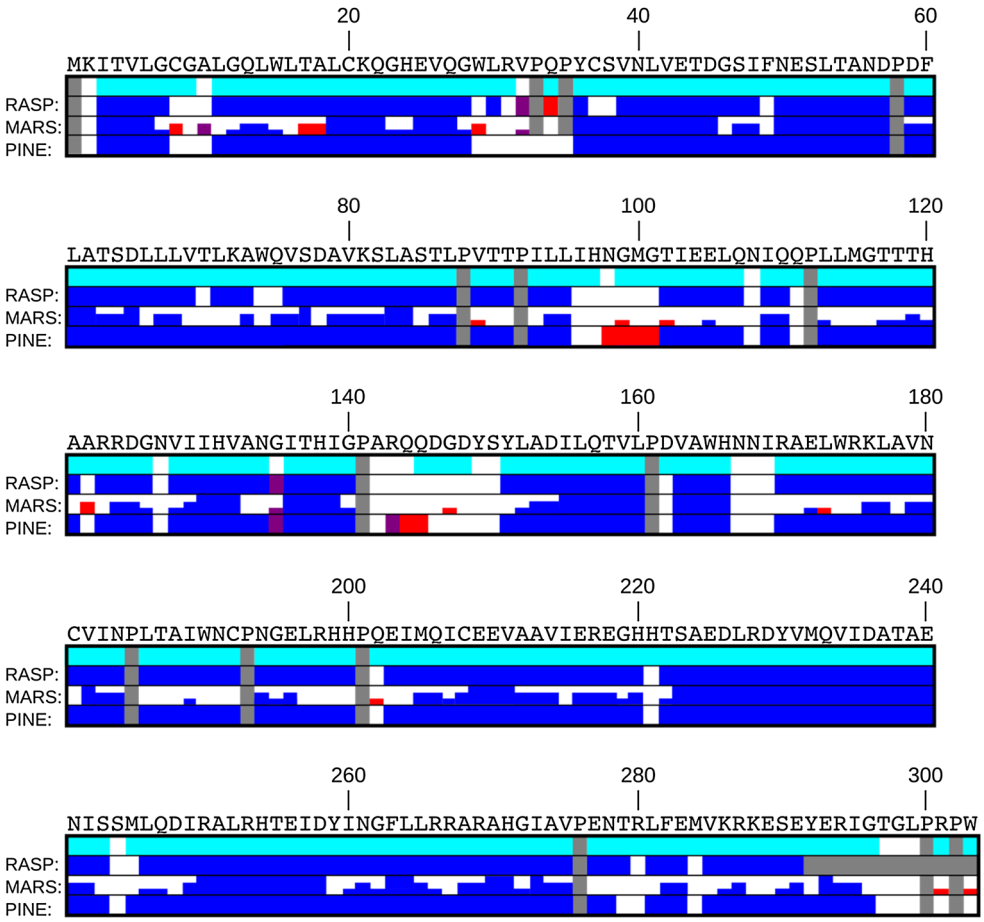



**Fig. 3** The sequential assignment of KPR. Conventional (Headey et al. 2008) (*top*), RASP, MARS (Jung and Zweckstetter 2004) and PINE (Bahrami et al. 2009) assignments are depicted below the KPR sequence, color-coded as unassignable (*grey*), not assigned (*white*), assigned conventionally (*cyan*), assigned in agreement with the conventional assignment (*blue*), assigned by RASP, MARS or PINE but not assigned conventionally (*purple*), misassigned by RASP, MARS or PINE (*red*). Bar height for MARS assignments encodes reported assignment reliability (high/medium/low)

assignments (87 % coverage) with 6 errors. Even this level of performance could not be achieved for this data set using MARS or PINE alone, as it is based on the refined set of spin systems assembled with the assistance of RASP. When the initial set of manually assembled spin systems was used as input for either of these established assignment strategies, they did not achieve the accuracy or coverage required for spin system refinement.

The process of iterative refinement of the input spin systems employed for the KPR assignments is analogous to conventional manual assignment strategies, in which peak-picking and spin system assembly is continuously refined in light of the current partial assignment. RASP assists this process significantly, thanks to the ensemble of assignment possibilities that RASP provides for each spin system (or each residue). Because the correct assignment is almost always found within this ensemble (Table 1), the search for likely candidates for an unassigned spin system is significantly facilitated.

The few errors in the RASP assignment (both for KPR and across the test set) occur almost exclusively at the extremities of contiguous assigned stretches, and in short, isolated stretches of assignable residues, as observed for example where one or two residues are flanked by prolines or residues that are otherwise unassignable. Careful examination of such regions in a proposed assignment is therefore warranted as a means of identifying possible errors. Conversely, such examination may also identify assignments that are missed by the algorithm, for example because of erroneous shift predictions for a particular residue, but that are strongly supported by sequential connectivities or other data.

## Conclusions

For proteins where structural information is available, RASP promises to significantly accelerate the backbone assignment process by reducing the requirements for both data acquisition and manual analysis. Several avenues for further development are immediately apparent. Although our current focus has been the use of data available in a conventional triple-resonance assignment campaign, other structural parameters such as NOEs or RDCs could be incorporated into the scoring function, and are likely to further improve assignment accuracy and coverage for challenging systems (Langmead and Donald 2004; Stratmann et al. 2010). Chemical shifts measured for proteins in

the solid state are generally in good agreement with those observed in solution, suggesting that chemical shift predictions are likely to offer similar advantages to the resonance assignment problem in solid-state NMR. RASP is available from http://sourceforge.net/p/raspnmr. We are currently applying RASP to proteins of interest that have hitherto resisted assignment by existing methods (MacRaild et al. 2011; Richard et al. 2010).

**Acknowledgments** We thank Stephen Headey and Martin Scanlon for sharing the KPR data, and David Chalmers for helpful discussions on optimization strategies. This work was supported in part by an Australian National Health and Medical Research Council project grant (1025150). RSN acknowledges fellowship support from the NHMRC.

## References

Bahrami A, Assadi AH, Markley JL, Eghbalnia HR (2009) Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. PLoS Comput Biol 5:e1000307

Barrett PJ, Chen J, Cho MK, Kim JH, Lu Z, Mathew S, Peng D, Song Y, Van Horn WD, Zhuang T, Sonnichsen FD, Sanders CR (2013) The quiet renaissance of protein nuclear magnetic resonance. Biochemistry 52:1303–1320

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242

Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci USA 104:9615–9620

Crippen GM, Rousaki A, Revington M, Zhang Y, Zuiderweg ERP (2010) SAGA: rapid automatic mainchain NMR assignment for large proteins. J Biomol NMR 46:281–298

Gronwald W, Willard L, Jellard T, Boyko RF, Rajarathnam K, Wishart DS, Sonnichsen FD, Sykes BD (1998) CAMRA: chemical shift based computer aided protein NMR assignments. J Biomol NMR 12:395–405

Han B, Liu Y, Ginzinger SW, Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 50:43–57

Headey SJ, Vom A, Simpson JS, Scanlon MJ (2008) Backbone assignments of the 34 kDa ketopantoate reductase from *E. coli*. Biomol NMR Assign 2:93–96

Jaipuria G, Krishnarjuna B, Mondal S, Dubey A, Atreya HS (2012) Amino acid selective labeling and unlabeling for protein resonance assignments. Adv Exp Med Biol 992:95–118

Jung YS, Zweckstetter M (2004) Mars—robust automatic backbone assignment of proteins. J Biomol NMR 30:11–23

Kohlhoff KJ, Robustelli P, Cavalli A, Salvatella X, Vendruscolo M (2009) Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. J Am Chem Soc 131:13894–13895

Kuhn HW (1955) The Hungarian method for the assignment problem. Naval Res Logist 2:83–87

Langmead CJ, Donald BR (2004) An expectation/maximization nuclear vector replacement algorithm for automated NMR resonance assignments. J Biomol NMR 29:111–138

Lepre CA, Moore JM, Peng JW (2004) Theory and applications of NMR-based screening in pharmaceutical research. Chem Rev 104:3641–3676

MacRaild CA, Anders RF, Foley M, Norton RS (2011) Apical membrane antigen 1 as an anti-malarial drug target. Curr Top Med Chem 11:2039–2047

Matak-Vinkovic D, Vinkovic M, Saldanha SA, Ashurst JL, von Delft F, Inoue T, Miguel RN, Smith AG, Blundell TL, Abell C (2001) Crystal structure of *Escherichia coli* ketopantoate reductase at 1.7 A resolution and insight into the enzyme mechanism. Biochemistry 40:14493–14500

Mittermaier AK, Kay LE (2009) Observing biological dynamics at atomic resolution using NMR. Trends Biochem Sci 34:601–611

Resende MGC, Ribeiro CC (2010) Greedy randomised adaptive search procedures: advances and applications. In: Gendreau M, Potvin J-Y (eds) Handbook of metaheuristics. Springer, New York, pp 283–319

Richard D, MacRaild CA, Riglar DT, Chan J-A, Foley M, Baum J, Ralph SA, Norton RS, Cowman AF (2010) Interaction between *Plasmodium falciparum* apical membrane antigen 1 and the rhoptry neck protein complex defines a key step in the erythrocyte invasion process of malaria parasites. J Biol Chem 285:14815–14822

Schmidt E, Güntert P (2012) A new algorithm for reliable and general NMR resonance assignment. J Am Chem Soc 134:12817–12829

Shen Y, Bax A (2010) SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. J Biomol NMR 48:13–22

Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44:213–223

Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci USA 105:4685–4690

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

Wishart DS, Arndt D, Berjanskii M, Tang P, Zhou J, Lin G (2008) CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. Nucleic Acids Res 36:W496–W502

Zhang H, Neal S, Wishart DS (2003) RefDB: a database of uniformly referenced protein chemical shifts. J Biomol NMR 25:173–195