J Biomol NMR (2013) 55:379–389
DOI 10.1007/s10858-013-9722-1

ARTICLE

PARAssign—paramagnetic NMR assignments of protein nuclei on the basis of pseudocontact shifts

Simon P. Skinner · Mois Moshev · Mathias A. S. Hass · Marcellus Ubbink

Received: 21 January 2013 / Accepted: 14 March 2013 / Published online: 23 March 2013

Abstract The use of paramagnetic NMR data for the refinement of structures of proteins and protein complexes is widespread. However, the power of paramagnetism for protein assignment has not yet been fully exploited. PARAssign is software that uses pseudocontact shift data derived from several paramagnetic centers attached to the protein to obtain amide and methyl assignments. The ability of PARAssign to perform assignment when the positions of the paramagnetic centers are known and unknown is demonstrated. PARAssign has been tested using synthetic data for methyl assignment of a 47 kDa protein, and using both synthetic and experimental data for amide assignment of a 14 kDa protein. The complex fitting space involved in such an assignment procedure necessitates that good starting conditions are found, both regarding placement and strength of paramagnetic centers. These starting conditions are obtained through automated tensor placement and user-defined tensor parameters. The results presented herein demonstrate that PARAssign is able to successfully perform resonance assignment in large systems with a high degree of reliability. This software provides a method for obtaining the assignments of large systems, which may previously have been unassignable, by using 2D NMR spectral data and a known protein structure.

Keywords Pseudocontact shift · Software · Assignment · Pseudoazurin · Cytochrome P450

Abbreviations

PCS	Pseudocontact shift
RDC	Residual dipolar coupling
PRE	Paramagnetic relaxation enhancement
PPMCC	Pearson Product Moment Correlation Coefficient
HSQC	Heteronuclear single quantum coherence
TROSY	Transverse relaxation optimized spectroscopy
CRINEPT	Cross-correlated relaxation enhanced polarization transfer
CLaNP-5	Caged Lanthanide NMR Probe 5

Simon P. Skinner and Mois Moshev contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s10858-013-9722-1) contains supplementary material, which is available to authorized users.

S. P. Skinner · M. A. S. Hass · M. Ubbink (✉)
Gorlaeus Laboratories, Leiden Institute of Chemistry, Leiden University, P.O. Box 9502, 2300 RA Leiden, The Netherlands
e-mail: m.ubbink@chem.leidenuniv.nl

S. P. Skinner
e-mail: skinnersp@chem.leidenuniv.nl

M. A. S. Hass
e-mail: hassmas@chem.leidenuniv.nl

M. Moshev
Leiden Institute of Advanced Computer Science, Leiden University, Snellius Building, P.O. Box 9512, 2300 RA Leiden, The Netherlands
e-mail: mois@monomon.me

Introduction

NMR spectroscopy is an invaluable technique to obtain structural and dynamic information about proteins and, thereby, enables a greater understanding of the functions that proteins carry out. A prerequisite for any detailed NMR study is that NMR assignments must be obtained for the nuclei of the protein in question. Traditionally,

heteronuclear multidimensional (3D/4D) NMR spectra are used to obtain this information (Fernandez and Wider 2003). For small proteins (<20 kDa), these experiments are sufficient to assign the nuclei. Larger proteins present more of a challenge due to their relaxation properties and spectral crowding. The line broadening problems can be overcome, for a large part, using TROSY and CRINEPT based experiments (Pervushin et al. 1997; Riek et al. 2002; Salzmann et al. 1998; Fiaux et al. 2002). Isotope labeling of methyl groups for larger proteins and protein complexes has also been used to circumvent the difficulties encountered when applying multidimensional NMR experiments on uniformly labeled proteins (Tugarinov et al. 2006). Moreover, the internal mobility of methyl groups causes slow relaxation of methyl protons, and therefore sharp lines are observed in the resulting spectra. Selective methyl labeling does not provide sequential information, but assignment can be obtained, for example by selective mutation of a methyl containing amino acid for another, as was shown for the 20S proteasome (Kay 2011). The method of selective mutation has been developed into a high-throughput technique, which combines automated site-directed mutagenesis, residue-type-specific-isotope labeling and fast NMR experiments (Amero et al. 2011). An additional approach is the so-called 'divide and conquer' approach (Sprangers et al. 2008), which involves assignment of nuclei of parts in a large complex and then transfer of the assignments to the spectra of the entire complex. However, many of these techniques require high concentrations of protein. Paramagnetic NMR can be complementary, or even an alternative to many of these methods. One of the major advantages of using paramagnetic NMR, in particular pseudocontact shifts (PCSs), for large proteins and protein complexes is that concentrations as low as 10 μM can be used to obtain data of sufficient quality to observe these effects (Keizers et al. 2010). Furthermore, PCSs have already proven to be very useful in the assignment and structure solution of proteins using solid-state NMR techniques (Luchinat et al. 2012; Bertini et al. 2009).

Incorporation of a paramagnetic lanthanide ion either into a natural metal binding site or via an attached lanthanide binding tag (Dvoretsky et al. 2002; Ikegami et al. 2004; Leonov et al. 2005; Pintacuda et al. 2004b; Rodriguez-Castañeda et al. 2006; Kamen et al. 2007; Keizers et al. 2007; Su et al. 2008; Jia et al. 2011) gives rise to observable paramagnetic effects such as PCSs, residual dipolar couplings (RDCs) and paramagnetic relaxation enhancements (PREs). PCSs are particularly useful due to the fact that they are easy to measure and provide long-range structural information. A PCS is a change in the observed Larmor frequency of a nuclear spin resulting from the time-averaged dipolar interaction between the observed nucleus and the anisotropy of the static magnetic moment of an unpaired electron spin in the paramagnetic center. PCSs depend on the distance between the center and the nucleus in an anisotropic fashion and are described by a magnetic susceptibility tensor (χ tensor). In order to calculate PCSs theoretically for nuclei in a known 3D structure, it is necessary to know eight parameters, namely, the position of the paramagnetic center (three Cartesian coordinates), the orientation of the χ tensor relative to the molecular frame (three Euler angles) and two anisotropy parameters that represent the axial and rhombic components of the χ tensor (Bertini et al. 2002). With many software packages the χ tensor can be determined from PCS data, given the 3D structure of the protein: Fantasia (Banci et al. 1996), Fantasian (Banci et al. 1997), the PARArestraints module for Xplor-NIH (Schwieters et al. 2003; Banci et al. 2004) and Numbat (Schmitz et al. 2008). These programs either perform a five-parameter χ tensor fit, using a known position of the paramagnetic center, or a complete eight-parameter fit. In all cases, the NMR assignments of the spectra of both the diamagnetic and paramagnetic protein samples are required to perform the fitting. Software packages to enable assignment of protein nuclei through the use of PCS data have been developed: Platypus (Pintacuda et al. 2004a), Echidna (Schmitz et al. 2006), and Possum (John et al. 2007). Platypus provides backbone amide assignment for resonances in the spectra of both the diamagnetic and paramagnetic samples using a known probe position, while simultaneously fitting the five tensor parameters. This has only been demonstrated, however, using residue-selectively ^{15}N-labeled samples and a small data set of 20 resonances. Echidna provides backbone amide assignment of a paramagnetic sample based on the spectral assignment of the equivalent diamagnetic sample, by fitting the magnitudes and Euler angles of the χ tensor: The metal position must be known for this procedure to be carried out. Possum is a method developed to automatically assign methyl groups of a protein using the resonance frequencies of the diamagnetic and paramagnetic samples, where the position of the paramagnetic center, as well as the orientation and magnitudes of the tensor are known beforehand. PARAssign is a tool wherein the assignment of the nuclei of large proteins can be achieved using only 2D spectra and a three-dimensional structure of the protein of interest with only the resonance frequencies observed for the diamagnetic and paramagnetic samples. In contrast to all previously published software, PARAssign uses the data obtained with several paramagnetic centers to determine the assignments. Thus, for each nucleus that needs to be assigned in the diamagnetic sample, a set of PCSs is available, providing sufficient data for a novel and efficient method for simultaneous χ tensor refinement and assignment of the nuclei.

PARAssign requires neither prior knowledge of the positions of the paramagnetic centers or the tensor orientations, nor any assignment information, only estimates of the magnitudes of the paramagnetic tensors, and in the case of methyl assignment, specification of the residue type.

Extensive testing of the program was undertaken using synthetic $^{1}H^{N}$ PCS data for pseudoazurin (PAZ) (125 residues) and P450cam (414 residues) to which selective methyl labeling schemes were applied. The robustness of the algorithm was further demonstrated using PAZ experimental data. All data were based on the use of Caged Lanthanide NMR Probe 5 (CLaNP-5) (Keizers et al. 2007; Keizers et al. 2008). However, the use of the software is not restricted to data acquired with this probe and the use of other paramagnetic probes, reviewed elsewhere (Koehler and Meiler 2011), is also possible. It should be noted though that many tags attached via a single arm to the protein induce much weaker paramagnetic effects, probably as a consequence of averaging effects due to considerable mobility of the tag relative to the protein. A large amplitude of tag movements could compromise the reliability of the assignment.

Theory

χ-Tensor fitting

A PCS can be described by a second rank magnetic susceptibility tensor:

$$\delta_{\mathrm{PCS}} = \frac{1}{12\pi r_{\mathrm{i}}^{5}}\left[\Delta\chi_{\mathrm{ax}}\left(3(\vec{r}_{\mathrm{z}}\cdot\vec{r}_{\mathrm{i}})^{2}-r_{\mathrm{i}}^{2}\right)+\frac{3}{2}\Delta\chi_{rh}\left((\vec{r}_{\mathrm{x}}\cdot\vec{r}_{\mathrm{i}})^{2}-(\vec{r}_{\mathrm{y}}\cdot\vec{r}_{\mathrm{i}})^{2}\right)\right] \quad (1)$$

where $\Delta\chi_{\mathrm{ax}}$ and $\Delta\chi_{\mathrm{rh}}$ represent the axial and rhombic components of the second rank magnetic susceptibility tensor, respectively, $r_{\mathrm{i}}=\sqrt{x_{\mathrm{i}}^{2}+y_{\mathrm{i}}^{2}+z_{\mathrm{i}}^{2}}$ (the coordinates of nucleus i, in the tensor frame with the paramagnetic center at the origin) and $\vec{r}_{\mathrm{x}}$, $\vec{r}_{\mathrm{y}}$ and $\vec{r}_{\mathrm{z}}$ represent the unit vectors that determine the orientation of the magnetic susceptibility tensor.

Equation 1 can be fitted using a five parameter fit, such that $\delta_{\mathrm{PCS}}=f(\Delta\chi_{\mathrm{ax}},\Delta\chi_{\mathrm{rh}},\alpha,\beta,\gamma)$ or an eight parameter fit, such that $\delta_{\mathrm{PCS}}=f(\Delta\chi_{\mathrm{ax}},\Delta\chi_{\mathrm{rh}},\alpha,\beta,\gamma,x,y,z)$, where α, β and γ are three Euler angles used to rotate $\vec{r}_{\mathrm{x}}$, $\vec{r}_{\mathrm{y}}$ and $\vec{r}_{\mathrm{z}}$ and x, y and z correspond to the Cartesian coordinates of the paramagnetic center. Using the Z–Y–Z convention, a rotation matrix, $R=R_{\mathrm{z}}(\alpha)R_{\mathrm{y}}(\beta)R_{\mathrm{z}}(\gamma)$ can be constructed to use these angles in a fitting procedure. The fitting procedure employed in PARAssign is a sequential least squares programming procedure (Kraft 1988) as implemented in Scipy 0.8.0 (http://www.scipy.org). The target function used in this fitting procedure is:

$$\frac{1}{N}\sum_{i}^{N}\left(\delta_{\mathrm{PCS,i}}^{\mathrm{pred}}-\delta_{\mathrm{PCS,i}}^{\mathrm{exp}}\right)^{2} \quad (2)$$

where $\delta_{\mathrm{PCS,i}}^{\mathrm{pred}}$ and $\delta_{\mathrm{PCS,i}}^{\mathrm{exp}}$ represent the predicted and experimental PCSs, respectively, and N is the total number of PCSs used for the minimization.

PCS determination procedure

Each paramagnetic center is placed on the protein according to a set of vector equations based on the refined tensor positions of CLaNP-5 on PAZ (Keizers et al. 2008) with some modifications (*a mathematical description is provided in the Supporting Material*). Briefly, the paramagnetic center is placed on a line that intercepts the center of the vector between the Cα atoms of the two residues to which the paramagnetic probe is attached and that is parallel to the z component of the tensor. The distance from the center to both Cα atoms is 8 Å. The $\vec{r}_{\mathrm{x}}$ vector of the paramagnetic tensor is defined as the direction of the Cα–Cα vector; the $\vec{r}_{\mathrm{z}}$ vector is defined perpendicular to $\vec{r}_{\mathrm{x}}$ and to point away from the protein by using the average of the vectors between the Cα and Cγ of the two residues to provide this direction. The $\vec{r}_{\mathrm{y}}$ vector is derived as the cross product of the $\vec{r}_{\mathrm{x}}$and $\vec{r}_{\mathrm{z}}$ vectors (Fig. 1). This provides the initial position and tensor orientation for each paramagnetic center. If a single-armed probe is used, its position can be user-defined or calculated by PARAssign. The $\vec{r}_{\mathrm{x}}$ vector of the paramagnetic tensor is defined as the direction of the Cα–Cβ bond vector; the $\vec{r}_{\mathrm{z}}$ vector is defined perpendicular to $\vec{r}_{\mathrm{x}}$ and to point away from the protein by using the Cα–Cγ bond of the residue to provide this direction. The $\vec{r}_{\mathrm{y}}$ vector is derived as the cross product of the $\vec{r}_{\mathrm{x}}$ and $\vec{r}_{\mathrm{z}}$ vectors. The distance between the Cα atom of the attachment site and the paramagnetic center is user-defined.

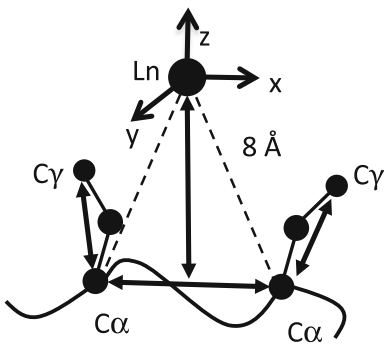

Fig. 1 Placement of the paramagnetic centers relative to the attachment site for a two-armed probe using a modified method to that previously published (see text and supporting information)

Peaks of nuclei close to the paramagnetic center are broadened beyond detection as a result of paramagnetic relaxation enhancement and are disregarded. The cutoff distance can be user-specified or calculated by PARAssign, using the equation for Curie spin relaxation (Schmitz et al. 2006):

$$r_{\text{cutoff}} \approx \sqrt[6]{\frac{1}{5\pi PRE}\left(\frac{\mu_0}{4\pi}\right)^2 B_0^2\gamma_X^2\frac{(g_e\mu_B)^4 S^2(S+1)}{(3k_BT)^2}\left(4\tau_r+\frac{3\tau_r}{1+\omega_X^2\tau_r^2}\right)} \quad (3)$$

where γ_X is the gyromagnetic ratio of the nucleus X, μ_0 is the permeability of a vacuum, T is the temperature, μ_B is the Bohr magneton, g_e is the electron g factor, S is the total spin quantum number of the paramagnetic ion (J in the case of lanthanides), τ_r is the rotational correlation time of the molecule, B_0 is the magnetic field strength, k_B is the Boltzmann constant and ω_X is the Larmor frequency of the nucleus X. As an indication, a PRE of $\approx$200 s^{-1} will usually lead to broadening beyond detection.

Assignment procedure

The Hungarian method for minimal cost assignment (Kuhn 1955) is used to perform the assignment of the experimental PCSs, using a Q factor cost function to populate the assignment matrix:

$$Q = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{\left(\delta_{\text{PCS,i}}^{\text{pred}}-\delta_{\text{PCS,i}}^{\text{exp}}\right)^2}{\left(\delta_{\text{PCS,i}}^{\text{pred}}+\delta_{\text{PCS,i}}^{\text{exp}}\right)^2}} \quad (4)$$

where N is the number of paramagnetic centers. This algorithm ensures that each experimental PCS is assigned exactly once to a predicted PCS and therefore, an assignable atom in the protein structure. The Q value is used as a filter when building the assignment matrix. Pairs of predicted and experimental PCSs whose Q score exceeds a user-defined value are excluded from the assignment matrix. In the case of methyl data, the residue type is also used to avoid incorrect assignments. The chemical shifts of methyl groups of different residues are sufficiently different for this to be done by user-defined residue typing. It should also be noted that predicted and experimental PCS with opposite signs are excluded from Q factor calculation and therefore cannot be assignment possibilities. This avoids any issues that would arise from the summed denominator in equation 4.

The PARAssign algorithm

Data input

PARAssign imports peak lists from spectra of both diamagnetic and paramagnetic forms of the protein from either a [^{13}C,^{1}H] or a [^{15}N,^{1}H] HSQC spectrum. Using the predicted PCSs, generated from the initial paramagnetic tensor orientations and magnitudes, a matching algorithm based on Bayesian statistics is used to match the peaks observed for the diamagnetic form of the protein to those for the paramagnetic form of the protein to produce PCSs (*see Supporting Material*). This procedure is executed for the data for each paramagnetic center and the combined sets of PCSs are then used for initial assignment. In this matching algorithm, the PCSs for both the proton and heteroatom dimensions are used, along with their ratios. These PCSs are then used for paramagnetic tensor refinement and protein assignment. The matching procedure produces on average 85 % correct matches based on the initial tensor orientation. These PCSs can be used as a starting point for spectral analysis to identify additional PCSs, which can then be subsequently used for assignment of the protein nuclei.

Initial assignment

Using the initial placement of the paramagnetic center and orientation of the tensors, the protein structure and the user-defined $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ values, a set of predicted PCSs is calculated for each paramagnetic center. All PCSs that are less than $\pm$0.02 are excluded from the datasets, since these values could be too small to be accurately measured experimentally. The protein structure is imported as a PDB file using the PDBparser module of Biopython (Hamelryck and Manderick 2003; Cock et al. 2009) and protonated. Protein structures can also be downloaded from the RCSB within PARAssign. It is possible that the initial tensor calculated by PARAssign may not have an optimal starting orientation and in order to establish whether it can be improved, two sequential searches of the α and β Euler angles are carried out. The signs of the PCSs are an indicator for the orientation of the tensor. To find an approximate orientation the total number of positive PCSs in the experimental and predicted sets are compared. So, using the predicted PCSs, the total number of positive PCSs of all relevant nuclei in the protein for each paramagnetic center is calculated for the predicted datasets and compared with that of the experimental datasets. The difference between the fractions of positive PCSs in the predicted and experimental datasets is calculated as a cost, c:

$$c = \frac{N_{\text{positive}}^{\text{predicted}}}{N_{\text{all}}^{\text{predicted}}} - \frac{N_{\text{positive}}^{\text{experimental}}}{N_{\text{all}}^{\text{experimental}}} \quad (5)$$

.

This cost is then used a "goodness of fit" statistic for the sequential searches of the α and β Euler angles. This approach is possible, because the value of c is strongly

dependent on the value of the β Euler angle (Fig. 2). By rotating the β Euler angle over the range $\pm 1/2\pi$ from its initial starting value calculated in PARAssign, it is evident that for four proteins of varying size, a similar pattern of sinusoidal variation is observed for all of the tested proteins. This is a good indication that using the number of positive PCSs is a valid manner in which to refine Euler angles prior to initial protein assignment. Due to the fact that rotation operators do not commute, two searches need to be carried out, one beginning with rotation of the α angle followed by rotation of the β angle ('α/β') and another beginning with rotation the β angle followed by rotation of the α angle ('β/α'). The 'α/β' sequential search is performed as follows. Each tensor frame is rotated by a $-1/2\pi$ α angle and then steps of 0.1 radians are performed until a rotation of $1/2\pi$ has been achieved. At each step of the walk, c is used to provide a "goodness of fit" value. The α angle corresponding to the lowest value of c is then selected as the optimal starting orientation for the α angle and denoted α^{opt}. The α angle can take values in the range $-1/2\pi \leq \alpha \leq +1/2\pi$ because the rhombic component of the χ tensor has π radian symmetry and therefore, only half of this angular space needs to be searched.

The β angle sequential search is then performed analogously, starting from the frame oriented by the angle α^{opt} and producing the angle β^{opt}. The 'β/α' search is performed in the same way, except the β angle is rotated first. The 'α/β' and 'β/α' searches each provide an "optimal" starting orientation and the search that provides the lowest value of c is used to rotate the tensor to the corresponding initial orientation. Only two angles need to be rotated because the orientation of one of the axes is always fixed in this searching procedure. Moreover, if the γ Euler angle were included in the 'β/α' search, this would be tantamount to performing the last search twice, since both α and γ determine the orientations of the x and y axes of the tensor frame. Using the orientations calculated from the individual sequential searches, the predicted PCSs are re-calculated and the initial assignment of the protein is determined. Each paramagnetic tensor is then fitted based on this assignment.

Fitting the tensors and assigning the protein

The paramagnetic tensors can be fitted using a constrained or unconstrained five or eight parameter fit. Each paramagnetic tensor is fitted to its assigned dataset individually and the PCSs associated with that tensor are re-calculated after application of the resulting fitting parameters. These predictions are then used to populate the assignment matrix required by the Hungarian method and an assignment based on these predictions is carried out. This procedure of fitting and assignment is carried out iteratively until a convergence limit is reached or a user-defined number of minimization steps has been performed. PARAssign then outputs the predicted and experimental values per paramagnetic center along with the predicted assignment, the deviation between predicted and experimental PCSs for each paramagnetic center as well as the average Q score per residue. A PDB file containing the final paramagnetic center positions and tensor orientations is also written (using Scientific Python 2.8) and, in addition, PARAssign produces a scatter plot showing the fit of predicted to experimental PCSs per paramagnetic center, along with the

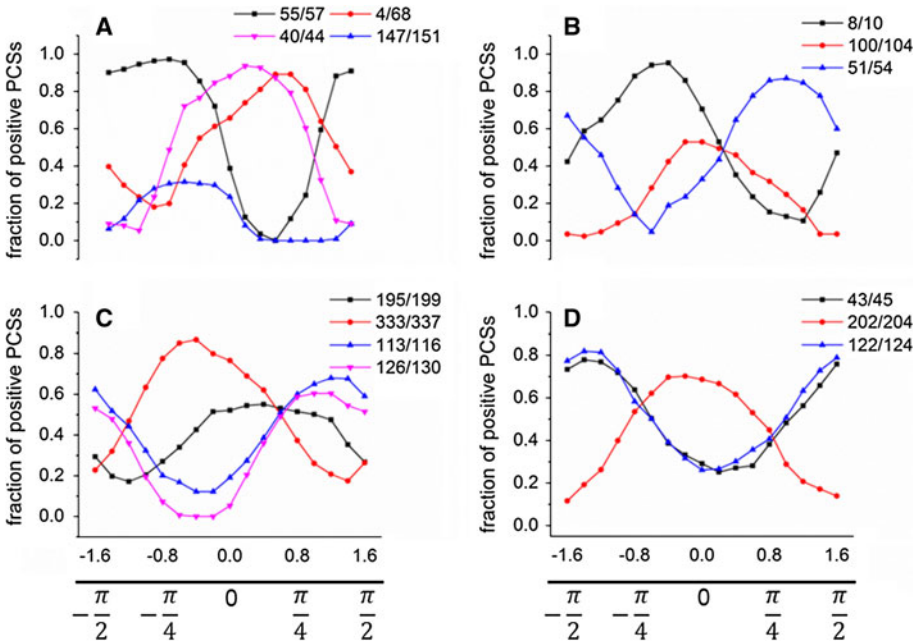

Fig. 2 Positive PCS. The fraction of predicted positive PCSs is plotted as a function of the β angle (radians) for four proteins: T4 lysozyme (A), PAZ (B), cytochrome P450cam (C) and Green Fluorescent Protein (D). The lines represent different attachment sites of the paramagnetic center

Pearson Product Moment Correlation Coefficient (Pearson's r), the linear least squares regression line fitted to the data and a line of y = x for comparison. After the final fitting and assignment step, all predicted PCSs are included in a subsequent assignment, such that all PCSs $< \pm 0.02$ are re-introduced into the dataset. In addition, a user-defined number of alternative assignments are produced by removing current assignments as possibilities from the assignment matrix and performing the Hungarian Algorithm again. This ensures that any alternative assignments will be different from previous assignments and also that the assignment corresponding to the next closest Q score to the previous assignment will be selected as an alternative assignment. This procedure can be carried out as many times as the user determines. All assignments have a reliability indicator associated with them. The trimmed average Q score is calculated using a 5 % upper limit for exclusion and this is used to determine assignment reliability; if the moduli of all PCSs used in an assignment of a nucleus are more than 0.02 and the average Q score is less than the trimmed average Q score, this assignment is deemed a highly reliable '**' (two-star) assignment. If only one of the conditions is met, it is deemed a reliable '*' (one-star) assignment and if neither conditions are met, the assignment is deemed unreliable. The complete algorithm is depicted in Fig. 3.

Testing of the algorithm

Two proteins were selected to test the robustness of the PARAssign algorithm. Pseudoazurin (PAZ), a blue copper protein of 13.5 kDa (PDB code 1PY0) (Prudencio et al. 2004) with Yb-CLaNP-5 attached at three positions was used to test the backbone amide assignment capabilities and cytochrome P450cam, a 46.6 kDa heme protein (PDB code 1DZ4) (Schlichting et al. 2000) with Tm-CLaNP-5 attached at four positions was used to test the methyl assignment capabilities using selective labeling schemes (Ile-Cδ1, 17 residues, Leu-Cδ1, 26 residues, Val-Cγ1, 22 residues, Ala-Cβ, 23 residues, Leu-Cδ1/Val-Cγ1, 46 residues, Ile-Cδ1/Leu-Cδ1/Val-Cγ1, 63 residues). Five synthetic datasets and one experimental dataset were used for testing with PAZ and five synthetic datasets were used for testing with P450cam. A dataset represents all sets of PCSs generated from all paramagnetic centers in a certain position, with a certain tensor size and orientation. Two separate tests were carried out, namely five parameter fitting and eight parameter fitting. The synthetic datasets for five-parameter fitting were generated using a random number generator to produce five sets of α and β Euler angles for each protein with the ranges $-\pi \leq \alpha \leq +\pi$ and $-1/2\pi \leq \beta \leq +1/2\pi$, as well as random $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ values within a range based on published values (Keizers et al. 2008) for Yb-CLaNP-5 (7×10^{-32} m$^3 \leq \Delta\chi_{ax} \leq 10 \times 10^{-32}$ m^3 and 1×10^{-32} m$^3 \leq \Delta\chi_{rh} \leq 4 \times 10^{-32}$ m^3) and Tm-CLaNP-5 (45×10^{-32} m$^3 \leq \Delta\chi_{ax} \leq 55 \times 10^{-32}$ m^3 and 9×10^{-32} m$^3 \leq \Delta\chi_{rh} \leq 11 \times 10^{-32}$ m^3).

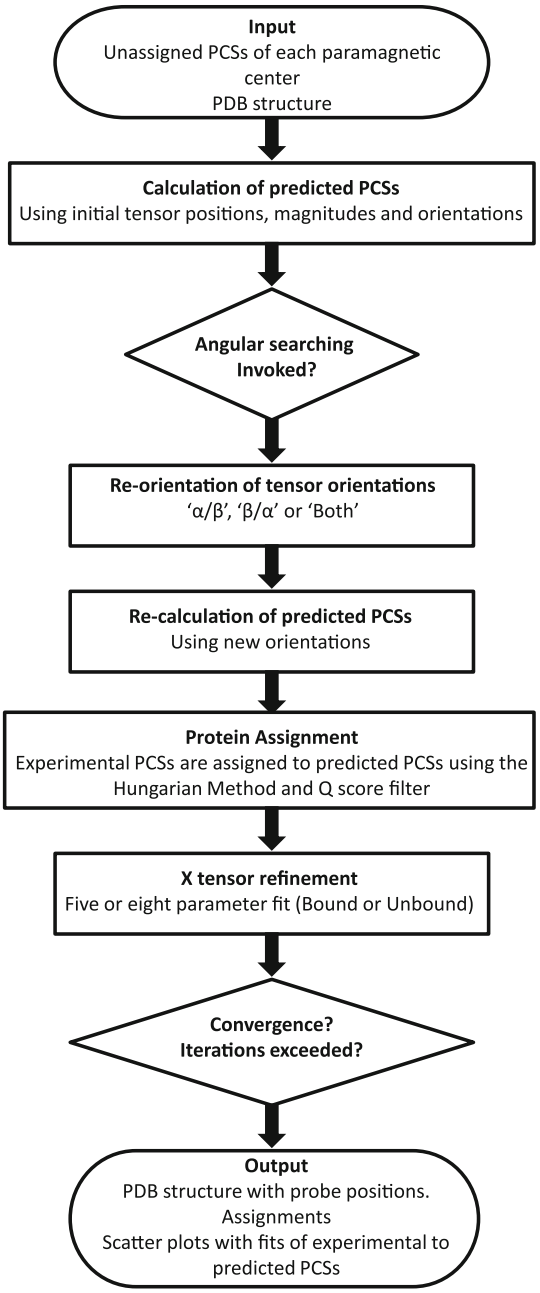

Fig. 3 Flowchart of the PARAssign Algorithm

The synthetic datasets for the eight parameter fits were produced using the datasets from the five parameter fitting and random x, y and z positions generated from a grid of ±3 Å x ± 3 Å x ± 3 Å around the initial position for each paramagnetic center calculated by PARAssign. These random values were used to re-orient the tensor and, where necessary, re-position the paramagnetic center and sets of

PCSs were calculated based on the new orientation and position. Noise was also added to each PCS as a random value in the range $-0.01 \leq x \leq 0.01$ ppm to represent experimental error. This error value was considered appropriate since all data sets used were ^{1}H PCS data. For heteronuclear data larger error margins may be required.

These datasets were then used as target ("quasi-experimental") data for PARAssign and the ability PARAssign to refine the tensor and assign each dataset was tested. In all tests, the starting values for $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ were 9.0×10^{-32} m^3 and 2.0×10^{-32} m^3, respectively for Yb-CLaNP and 50.0×10^{-32} m^3 and 10.0×10^{-32} m^3, respectively for Tm-CLaNP. The cutoff distance for excluding residues whose signals would be broadened beyond detection was set to 10 Å for Yb-labeled PAZ and 16 Å for Tm-labeled P450cam and all PCSs that would theoretically be unobservable were excluded from all synthetic datasets. The starting position and orientation were always those obtained using the procedure described above (PCS determination procedure).

Results and discussion

Synthetic data—PAZ backbone amides

The PARAssign algorithm performs simultaneous χ tensor fitting and assignment of protein nuclei. There are two types of fitting available in PARAssign; namely five parameter fitting, with fixed positions for the paramagnetic centers, and eight parameter fitting. The user can set boundaries for $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ and the metal position if a bound fit is selected. Initially, five parameter fitting and assignment was tested using five synthetic datasets generated from published $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ values and random Euler angles. All initial assignments were generated without applying any angular searching ('None'), such that the initial position derived from placing the tensor was used to perform the assignment. The Q score filter (Equation 4) was set to 0.25 for all tests and incrementally increased by 0.25 to a maximum of 0.75, if the number of amides assigned in the final output was below the maximal assignable for the protein, taking into account the line broadening caused by the probe (*see above*). The Q score filter for assignment is set by the user and is not incremented automatically. In cases where the final assignment was not optimal using the initial tensor placement as the starting position, two sequential Euler angle searches were included separately, namely an 'α/β' and a 'β/α' search (see Initial Assignment section) to derive a better initial orientation for assignment and the Q score filter values were incremented as previously stated. If application of neither of the searches to the initial starting orientation resulted in an optimal assignment, both the 'α/β' and 'β/α' searches were applied ('Both') in the same assignment run and the search that produced the lowest cost, c (equation 5) was selected to move the tensor to its starting orientation and the Q score filter was incremented; this being the final test for each datasets for which an optimal assignment had not been achieved. The five synthetic datasets yielded diverse results (Table 1). For all datasets, the initial orientation of the probe was not sufficient to obtain an optimal assignment, shown by the fact that a search was required in all cases. The starting point for assignment and tensor fitting is a strong determining factor of the final assignment results, as shown in Table 1, along with the Q score filter value used for assignment.

The diversity in "Best Search" used to obtain the optimal assignment for each of the five datasets indicates that all search possibilities should be used and a suitable Q filter should accompany each search. In some situations, a low Q score value of 0.25 was sufficient to obtain a nearly complete and accurate assignment of PAZ amides, however, for other datasets, a higher Q score value was required to enable potentially erroneous assignments to be used to guide the minimization and consequently, result in those erroneous assignments being corrected. An assignment result was deemed optimal based on whether the maximum number of assignable residues had been reached, the value of the target function per paramagnetic center and the overall fitting statistics for each mutant, i.e. the closeness of the least-squares regression line to the line of $y = x$ and the corresponding Pearson Product Moment Correlation Coefficient (PPMCC). Moreover, the reliability of an individual assignment was determined based on whether the average Q score was less than the trimmed average Q score for that assignment and whether the moduli of all PCSs for that amide were greater than 0.02. An example of the final output from PARAssign, based on experimental PAZ data (*described below*), including the reliability indicators is shown in Table S1. The five parameter fitting tests showed that three quarters of the assignments achieved in all cases were in the most reliable category (2*). The 1* assignments were often put in this class due to the fact that one or more of the PCS for that amide were below the threshold for a reliable assignment (0.02). All starred (2* and 1*) assignments were correct for this test and all others reported below. Within the 0* assignments produced by PARAssign at most 20 % false positives (wrong assignments) were found in these tests, indicating that 0* assignments are useful suggestions, but need to be checked with other methods. This shows that PARAssign not only produces many reliable assignments of amides on the basis of PCS data, but also that those that are unreliable can be easily identified in the final output.

Table 1 Summary of results from testing of five parameter fitting using PAZ backbone amide synthetic data

Dataset	Best search	Total assignable amides	Q filter	Total assigned amides	Correct/assigned		
					2*	1*	0*
1	Both	85	0.75	82/83	51/51	18/18	13/14
2	β/α	85	0.25	78/80	47/47	25/25	6/8
3	Both	85	0.75	83/83	45/45	27/27	11/11
4	α/β	85	0.25	76/78	43/43	26/26	7/9
5	β/α	85	0.50	84/84	43/43	33/33	8/8

The eight parameter fitting abilities of PARAssign using PAZ synthetic datasets were carried out in the same manner as the five parameter fitting tests. These results are shown in Table 2. These results show that PARAssign can perform assignment of proteins to an almost identical degree of accuracy whether the positions of the paramagnetic centers are known or unknown.

Experimental data—PAZ backbone amides

A set of previously published experimental PCS data for PAZ (Keizers et al. 2008) was used to test the PARAssign algorithm with a real data set derived from [^{15}N,^{1}H]-HSQC spectra. The PCSs used for the assignment were produced both via the peak matching algorithm and from manual identification of the PCSs from the spectra. The PCSs identified by the peak matching algorithm were used as a starting point for building the PCS lists used in assignment and additional PCSs were identified by spectral analysis, providing a comprehensive set of experimental data for the assignment of PAZ amides. The assignment was carried out using a restrained eight-parameter fit. This was performed on a ±3 Å x ± 3 Å x ± 3 Å grid, with the tensor magnitudes restricted to 7×10^{-32} m^3 $< \Delta\chi_{ax} <$ 10×10^{-32} m^3 and 1×10^{-32} m^3 $< \Delta\chi_{rh} <$ 4×10^{-32} m^3 and the Euler angles restrained using the following limits: $-\pi < \alpha/\gamma < \pi$ and $-1/2\pi < \beta < 1/2\pi$. The starting values for $\Delta\chi_{ax}$ and $\Delta\chi_{rh}$ were 9.0×10^{-32} m^3 and 2.0×10^{-32} m^3, respectively.

Using this real data, only three amides were misassigned out of the 57 assignable amides. The reason that only 57 of the amides were assignable and not 85 as stated for the synthetic data was not related to the PARAssign software, but due to the limited quality of one experimental data sets, such that only a subset of all resonances could be observed. Of the correct assignments, 32 were 2* assignments, 17 were 1* assignments and five had no star. All three incorrect assignments were not starred and therefore deemed unreliable (Table S1). It is concluded that the PARAssign algorithm can handle and successfully assign amides using experimental data, as evidenced by the reliability of the assignments obtained and the fitting statistics, Fig. 4, which show that the fits of experimental to predicted data were excellent.

Synthetic data—P450cam

The assignment of methyl groups was also tested using PARAssign. Five and eight parameter fits were tested in the same way as for amide assignment. The results of five parameter fitting using P450cam methyl data showed diverse performance of PARAssign (Table 3). All datasets for single residue selective labeling did not converge to a minimum that provided an assignment. In all cases, the number of residues assigned below the Q score filter was below 80 % of the total set (*data not shown*).

This shows that a simple manner to assign a protein for which the number of residues of a single type is too small, is to combine datasets of several residue types. The fraction of 2* assignments was approximately two-thirds for all except dataset 1 for both the Ile-Cδ1/Leu-Cδ1/Val-Cγ1 and Leu-Cδ1/Val-Cγ1 synthetic data. Furthermore, 1* assignments accounted for most of the remaining methyls assigned by PARAssign and these were deemed 1* due to the fact that the absolute values of some PCSs in the dataset were less than 0.02. The testing of eight parameter fitting

Table 2 Summary of results from testing of eight parameter fitting using PAZ backbone amide synthetic data

Dataset	Best search	Total assignable amides	Q filter	Total assigned amides	Correct/assigned		
					2*	1*	0*
1	Both	85	0.75	82/83	51/51	18/18	13/14
2	α/β	85	0.5	78/80	44/44	25/25	9/11
3	Both	85	0.75	83/83	45/45	27/27	11/11
4	α/β	85	0.25	76/78	43/43	26/26	7/9
5	β/α	85	0.50	84/84	43/43	33/33	8/8

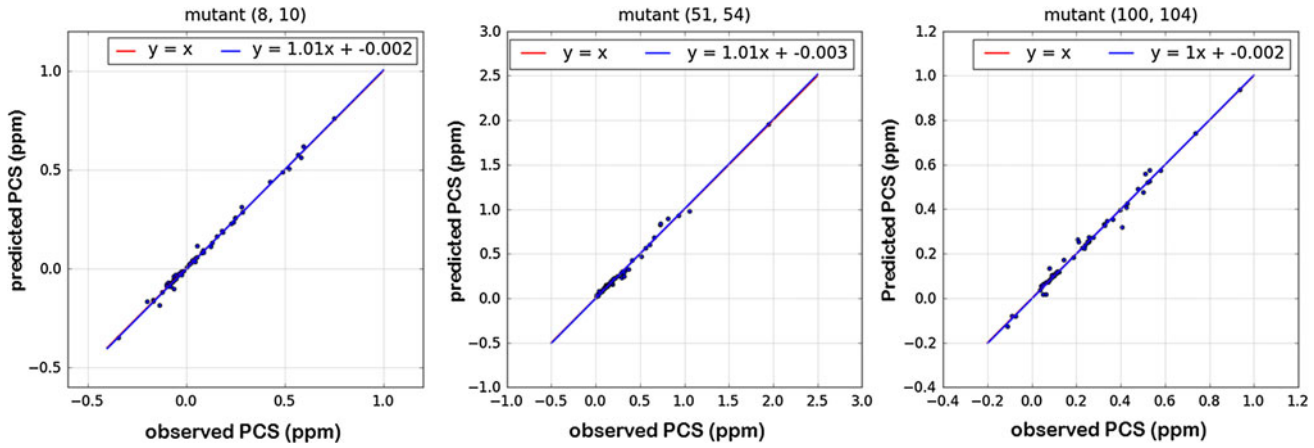

Fig. 4 Scatter plots of predicted against observed ^{1}H PCSs for PAZ (Keizers et al. 2008) after fitting and assignment by PARAssign using the structure of PAZ (PDB entry 1PY0) (Prudencio et al. 2004)

Table 3 Summary of results from testing of five parameter fitting using P450 synthetic data

Dataset	Total assignable methyl groups	Q filter	Best search	Total assigned methyl groups	Correct/assigned		
					2*	1*	0*
Ile-Cδ1/Leu-Cδ1/Val-Cγ1							
1	63	0.50	β/α	63/63	33/33	29/29	1/1
2	63	0.75	β/α	63/63	39/39	18/18	6/6
3	63	0.75	None	60/63	38/38	17/17	5/8
4	63	0.75	β/α	61/63	37/37	24/24	0/2
5	63	0.50	α/β	59/63	36/36	23/23	0/4
Leu-Cδ1/Val-Cγ1							
1	46	0.50	β/α	43/46	19/19	24/24	0/3
2	46	0.75	β/α	46/46	31/31	15/15	0/0
3	46	0.75	None	45/46	31/31	13/13	1/2
4	46	0.75	β/α	44/46	28/28	16/16	0/2
5	46	0.50	α/β	43/46	26/26	15/15	2/5

Table 4 Summary of results from testing of eight parameter fitting using P450 synthetic data

Dataset	Total assignable methyl groups	Q filter	Best search	Total assigned methyl groups	Correct/assigned		
					2*	1*	0*
Ile-Cδ1/Leu-Cδ1/Val-Cγ1							
1	63	0.50	β/α	63/63	39/39	23/23	1/1
2	63	0.75	β/α	63/63	38/38	20/20	5/5
3	63	0.75	None	59/63	35/35	22/22	2/6
4	63	0.75	β/α	61/63	35/35	26/26	0/2
5	63	0.50	α/β	63/63	40/40	23/23	0/0
Leu-Cδ1/Val-Cγ1							
1	46	0.50	β/α	46/46	31/31	12/12	3/3
2	46	0.75	β/α	45/46	30/30	13/13	2/3
3	46	0.75	None	45/46	26/26	17/17	2/3
4	46	0.75	β/α	44/46	24/24	18/18	2/4
5	46	0.50	α/β	46/46	29/29	15/15	2/2

was carried out in an identical manner to the five parameter fitting and a summary of the results is shown in Table 4.

Methyl assignment using an eight-parameter fit showed the same issue as for single residue type datasets. However, for both the combined Ile-Cδ1/Leu-Cδ1/Val-Cγ1 and the combined Leu-Cδ1/Val-Cγ1 datasets, a similar degree of success in assignment was achieved as with the five parameter fitting. The reliability of these assignments varied when compared with the reliability of those produced from five parameter fitting. In some cases, the number of correct and reliable assignments increased when eight parameter fitting was employed as opposed to five parameter fitting. This is particularly true for the Leu-Cδ1/Val-Cγ1 datasets. In two cases, datasets one and five, more correct and more 2* assignments were obtained. In addition, more correct and 2* assignments were obtained for dataset five of the Ile-Cδ1/Leu-Cδ1/Val-Cγ1 series. It is possible that the increase in assignment accuracy is due to the fact that a larger parameter space is searched in eight parameter fitting. These data demonstrate that PARAssign is able to successfully and reliably assign methyls in a large protein with selective labeling, at least using synthetic data. Validation on experimental data sets in progress.

Conclusions

PARAssign enables the user to perform amide and methyl assignments of a protein in the absence of any prior assignment information. PARAssign performs well both in situations where the positions of the paramagnetic centers are known and when they are unknown. When handling such an assignment problem, a small number of peaks represents a serious impediment to obtaining a reliable and accurate assignment, due to the multiparameter fitting problem that needs to be solved. The only information known before commencing this type of assignment are the 3D structure of the protein and an estimate of the magnitudes of the axial and rhombic components of the magnetic susceptibility tensor. However, the problem of insufficient data can be overcome by combining datasets as was demonstrated by the use of Leu-Cδ1/Val-Cγ1 and Ile-Cδ1/Leu-Cδ1/Val-Cγ1 methyl data. This reduces the relative complexity, because these data sets share the same tensor parameters. PARAssign represents a manner in which PCS-based assignments can be carried out on larger proteins and future versions should contain the ability to work with large multimeric systems, since nuclei in such systems are at present extremely challenging to assign by traditional means. RDCs and residual anisotropic chemical shift corrections should also be included in future versions for working with TROSY spectra.

PARAssign is available for download at: http://protchem.lic.leidenuniv.nl/software/parassign/registration.

Acknowledgments We thank Dr. Peter Keizers for providing the experimental PAZ data used in validation of the software. This work was supported financially by the Netherlands Organisation for Scientific Research (NWO), VICI Grant 700-58-441.

Conflict of interest The authors declare no competing financial interest.

References

Amero C, Asuncion Dura M, Noirclerc-Savoye M, Perollier A, Gallet B, Plevin M, Vernet T, Franzetti B, Boisbouvier J (2011) A systematic mutagenesis-driven strategy for site-resolved NMR studies of supramolecular assemblies. J Biomol NMR 50(3):229–236

Banci L, Bertini I, Bren KL, Cremonini MA, Gray HB, Luchinat C, Turano P (1996) The use of pseudocontact shifts to refine solution structures of paramagnetic metalloproteins: Met80Ala cyano-cytochrome c as an example. J Biol Inorg Chem 1(2): 117–126

Banci L, Bertini I, Savellini GG, Romagnoli A, Turano P, Cremonini MA, Luchinat C, Gray HB (1997) Pseudocontact shifts as constraints for energy minimization and molecular dynamics calculations on solution structures of paramagnetic metalloproteins. Proteins 29(1):68–76

Banci L, Bertini I, Cavallaro G, Giachetti A, Luchinat C, Parigi G (2004) Paramagnetism-based restraints for Xplor-NIH. J Biomol NMR 28(3):249–261

Bertini I, Luchinat C, Parigi G (2002) Magnetic susceptibility in paramagnetic NMR. Prog Nucl Mag Res Sp 40(3):249–273

Bertini I, Bhaumik A, De Paëpe G, Griffin RG, Lelli M, Lewandowski JzR, Luchinat C (2009) High-resolution solid-state NMR structure of a 17.6 kDa protein. J Am Chem Soc 132(3): 1032–1040

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJ (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25(11):1422–1423

Dvoretsky A, Gaponenko V, Rosevear PR (2002) Derivation of structural restraints using a thiol-reactive chelator. FEBS Lett 528(1–3):189–192

Fernandez C, Wider G (2003) TROSY in NMR studies of the structure and function of large biological macromolecules. Curr Opin Struct Biol 13(5):570–580

Fiaux J, Bertelsen EB, Horwich AL, Wuthrich K (2002) NMR analysis of a 900K GroEL GroES complex. Nature 418(6894): 207–211

Hamelryck T, Manderick B (2003) PDB file parser and structure class implemented in Python. Bioinformatics 19(17):2308–2310

Ikegami T, Verdier L, Sakhaii P, Grimme S, Pescatore B, Saxena K, Fiebig KM, Griesinger C (2004) Novel techniques for weak alignment of proteins in solution using chemical tags coordinating lanthanide ions. J Biomol NMR 29(3):339–349

Jia X, Yagi H, Su XC, Stanton-Cook M, Huber T, Otting G (2011) Engineering [Ln(DPA)3] 3- binding sites in proteins: a widely applicable method for tagging proteins with lanthanide ions. J Biomol NMR 50(4):411–420

John M, Schmitz C, Park AY, Dixon NE, Huber T, Otting G (2007) Sequence-specific and stereospecific assignment of methyl

groups using paramagnetic lanthanides. J Am Chem Soc 129(44):13749–13757
Kamen DE, Cahill SM, Girvin ME (2007) Multiple alignment of membrane proteins for measuring residual dipolar couplings using lanthanide ions bound to a small metal chelator. J Am Chem Soc 129(7):1846–1847
Kay LE (2011) Solution NMR spectroscopy of supra-molecular systems, why bother? A methyl-TROSY view. J Magn Reson 210(2):159–170
Keizers PH, Desreux JF, Overhand M, Ubbink M (2007) Increased paramagnetic effect of a lanthanide protein probe by two-point attachment. J Am Chem Soc 129(30):9292–9293
Keizers PH, Saragliadis A, Hiruma Y, Overhand M, Ubbink M (2008) Design, synthesis, and evaluation of a lanthanide chelating protein probe: CLaNP-5 yields predictable paramagnetic effects independent of environment. J Am Chem Soc 130(44):14802–14812
Keizers PH, Mersinli B, Reinle W, Donauer J, Hiruma Y, Hannemann F, Overhand M, Bernhardt R, Ubbink M (2010) A solution model of the complex formed by adrenodoxin and adrenodoxin reductase determined by paramagnetic NMR spectroscopy. Biochemistry 49(32):6846–6855
Koehler J, Meiler J (2011) Expanding the utility of NMR restraints with paramagnetic compounds: background and practical aspects. Prog Nucl Magn Reson Spectrosc 59(4):360–389
Kraft D (1988) A software package for sequential quadratic programming (trans: Center DGA). Tech Rep. DFVLR-FB 88–28. DFVLR-FB 88–28; Institute for Flight Mechanics: Koln, Koln
Kuhn HW (1955) The Hungarian method for the assignment problem. Nav Res Logist Q 2(1–2):83–97
Leonov A, Voigt B, Rodriguez-Castaneda F, Sakhaii P, Griesinger C (2005) Convenient synthesis of multifunctional EDTA-based chiral metal chelates substituted with an S-mesylcysteine. Chemistry 11(11):3342–3348
Luchinat C, Parigi G, Ravera E, Rinaldelli M (2012) Solid-state NMR crystallography through paramagnetic restraints. J Am Chem Soc 134(11):5006–5009
Pervushin K, Riek R, Wider G, Wuthrich K (1997) Attenuated T2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. Proc Natl Acad Sci USA 94(23):12366–12371
Pintacuda G, Keniry MA, Huber T, Park AY, Dixon NE, Otting G (2004a) Fast structure-based assignment of 15N HSQC spectra of selectively 15N-labeled paramagnetic proteins. J Am Chem Soc 126(9):2963–2970
Pintacuda G, Moshref A, Leonchiks A, Sharipo A, Otting G (2004b) Site-specific labelling with a metal chelator for protein-structure refinement. J Biomol NMR 29(3):351–361
Prudencio M, Rohovec J, Peters JA, Tocheva E, Boulanger MJ, Murphy ME, Hupkes HJ, Kosters W, Impagliazzo A, Ubbink M (2004) A caged lanthanide complex as a paramagnetic shift agent for protein NMR. Chemistry 10(13):3252–3260
Riek R, Fiaux J, Bertelsen EB, Horwich AL, Wuthrich K (2002) Solution NMR techniques for large molecular and supramolecular structures. J Am Chem Soc 124(41):12144–12153
Rodriguez-Castañeda F, Haberz P, Leonov A, Griesinger C (2006) Paramagnetic tagging of diamagnetic proteins for solution NMR. Mag Reson Chem 44:S10–S16
Salzmann M, Pervushin K, Wider G, Senn H, Wuthrich K (1998) TROSY in triple-resonance experiments: new perspectives for sequential NMR assignment of large proteins. Proc Natl Acad Sci USA 95(23):13585–13590
Schlichting I, Berendzen J, Chu K, Stock AM, Maves SA, Benson DE, Sweet RM, Ringe D, Petsko GA, Sligar SG (2000) The catalytic pathway of cytochrome p450cam at atomic resolution. Science 287(5458):1615–1622
Schmitz C, John M, Park AY, Dixon NE, Otting G, Pintacuda G, Huber T (2006) Efficient chi-tensor determination and NH assignment of paramagnetic proteins. J Biomol NMR 35(2): 79–87
Schmitz C, Stanton-Cook MJ, Su XC, Otting G, Huber T (2008) Numbat: an interactive software tool for fitting deltachi-tensors to molecular coordinates using pseudocontact shifts. J Biomol NMR 41(3):179–189
Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160(1):65–73
Sprangers R, Li X, Mao X, Rubinstein JL, Schimmer AD, Kay LE (2008) TROSY-based NMR evidence for a novel class of 20S proteasome inhibitors. Biochemistry 47(26):6727–6734
Su XC, Man B, Beeren S, Liang H, Simonsen S, Schmitz C, Huber T, Messerle BA, Otting G (2008) A dipicolinic acid tag for rigid lanthanide tagging of proteins and paramagnetic NMR spectroscopy. J Am Chem Soc 130(32):10486–10487
Tugarinov V, Kanelis V, Kay LE (2006) Isotope labeling strategies for the study of high-molecular-weight proteins by solution NMR spectroscopy. Nat Protoc 1(2):749–754