J Biomol NMR (2013) 56:113–123
DOI 10.1007/s10858-013-9727-9

ARTICLE

Prediction of peak overlap in NMR spectra

Frederik Hefke · Roland Schmucki · Peter Güntert

Received: 12 January 2013 / Accepted: 4 April 2013 / Published online: 13 April 2013

Abstract Peak overlap is one of the major factors complicating the analysis of biomolecular NMR spectra. We present a general method for predicting the extent of peak overlap in multidimensional NMR spectra and its validation using both, experimental data sets and Monte Carlo simulation. The method is based on knowledge of the magnetization transfer pathways of the NMR experiments and chemical shift statistics from the Biological Magnetic Resonance Data Bank. Assuming a normal distribution with characteristic mean value and standard deviation for the chemical shift of each observable atom, an analytic expression was derived for the expected overlap probability of the cross peaks. The analytical approach was verified to agree with the average peak overlap in a large number of individual peak lists simulated using the same chemical shift statistics. The method was applied to eight proteins, including an intrinsically disordered one, for which the prediction results could be compared with the actual overlap based on the experimentally measured chemical shifts. The extent of overlap predicted using only statistical chemical shift information was in good agreement with the overlap that was observed when the measured shifts were used in the virtual spectrum, except for the intrinsically disordered protein. Since the spectral complexity of a protein NMR spectrum is a crucial factor for protein structure determination, analytical overlap prediction can be used to identify potentially difficult proteins before conducting NMR experiments. Overlap predictions can be tailored to particular classes of proteins by preparing statistics from corresponding protein databases. The method is also suitable for optimizing recording parameters and labeling schemes for NMR experiments and improving the reliability of automated spectra analysis and protein structure determination.

Keywords Chemical shift distribution · Peak overlap · Peak dispersion · Protein structure determination · CYANA

F. Hefke · R. Schmucki · P. Güntert
Center for Biomolecular Magnetic Resonance, Institute of Biophysical Chemistry, Frankfurt am Main, Germany

F. Hefke · R. Schmucki · P. Güntert (✉)
Frankfurt Institute for Advanced Studies, Goethe University Frankfurt am Main, Max-von-Laue-Str. 9, 60438 Frankfurt am Main, Germany
e-mail: guentert@em.uni-frankfurt.de

Present Address:
R. Schmucki
F. Hoffmann-La Roche AG, Basel, Switzerland

Introduction

The structure determination of membrane proteins, amyloidal fibrils, and large complexes represents one of the biggest challenges in the area of structural biology. NMR spectroscopy is an important tool for such investigations, since it can provide both structural as well as dynamic information. The basis for any detailed investigation of proteins by NMR is the resonance assignment, which can take up considerable time and effort and is often hindered by experimental issues, especially in larger proteins, protein complexes, membrane proteins, or amyloid fibrils. Larger molecules have longer rotational correlation times and consequently shorter transverse relaxation times T_2, leading to line broadening in the NMR spectrum. Especially in 3D and 4D spectra even the natural linewidth given by T_2 relaxation can usually not be reached because the limited acquisition time demands truncation of the free induction decays in the indirect dimensions. Hence there is

a convolution of the signal with a function of broadness inversely proportional to the maximum evolution time, which in general results in broader lines than relaxation (Ernst et al. 1987; Szántay 2007). This results in signal overlap in the spectrum. With membrane proteins the situation is further complicated because the protein has to be surrounded by amphipathic molecules that increase the effective molecular weight, resulting in broader line widths. Combined with the relatively narrow chemical shift dispersion found in α-helices, this increases overlap and leads to assignment ambiguity. Using 3D and 4D spectra can in principle reduce overlap, but, as mentioned above, at least part of the advantage is lost by truncation effects and lower sensitivity. Higher static magnetic fields B_0 can improve the resolution and signal-to-noise ratio but may not be readily available. Another way to reduce the overlap is to sparsely label the sample (Goto and Kay 2000; Kainosho and Güntert 2009; Lian and Middleton 2001), for example by using labeled amino acids (Higman et al. 2009; Kainosho et al. 2006), segmental labeling (Busche et al. 2009; Yamazaki et al. 1998), transmembrane segment enhanced labeling (Reckel et al. 2008), or employing pair-labeling strategies (Hefke et al. 2011).

The extent of overlap is a function of the line width, the number of peaks, and their dispersion in the spectrum. The simplest model for estimating peak overlap assumes that peaks are uniformly distributed in the spectrum (Mumenthaler et al. 1997). To get a first estimate of how overlap becomes more problematic in larger proteins with more shifts and broader resonance lines, we consider N peaks, distributed randomly within a region of size Γ in a n-dimensional spectrum. Each peak is assumed to occupy a "peak region" of size γ, given by the data points of the peak that are significantly above the noise level. A peak is classified as overlapped if the center of at least one other peak falls within its peak region. The expected number of peaks that are not overlapped with other peaks can be approximated by (Kainosho et al. 2006)

$$\tilde{N} = N(1 - \gamma/\Gamma)^{N-1} \approx Ne^{-N\gamma/\Gamma}. \quad (1)$$

The number $\tilde{N}$ of non-overlapped peaks decreases exponentially with the size of the protein and the size of the peak region. The quantity $N\gamma/\Gamma$ in the exponent of Eq. 1 is the fraction of the entire spectral space that would be occupied by peaks in the absence of any overlap. If the n-dimensional peak regions γ and the spectral region Γ are the product of corresponding one-dimensional regions γ_1 and Γ_1, one obtains

$$\tilde{N} \approx Ne^{-N(\gamma_1/\Gamma_1)^n}. \quad (2)$$

With increasing number n of dimensions the space in which the peaks are distributed increases exponentially and overlap is significantly decreased. However, this simplistic model underestimates the amount of overlap that occurs in a real spectrum because it makes the unrealistic assumption that peaks are distributed uniformly and independently in the spectrum. The overlap probability expressed by Eqs. 1 and 2 is therefore overly optimistic. In reality peak positions are determined by the underlying chemical shifts, which in turn are dependent on the type of amino acid they originate from and other factors such as the secondary structure. For instance, α-helical membrane proteins exhibit narrower chemical shift distributions than β-sheet proteins (Oxenoid et al. 2004).

In this article we present a new method to estimate the expected spectral overlap that assigns to peaks in a spectrum a probability of being overlapped by using atom-specific chemical shift distributions from the Biological Magnetic Resonance Data Bank (BMRB) (Ulrich et al. 2008).

Materials and methods

The new method for estimating peak overlap has been implemented in the CYANA software package (Güntert 2009; Güntert et al. 1997). An overview of the algorithm is given in Fig. 1.

Chemical shift database

To account for the different chemical shift distributions of individual atoms, shifts are not treated as uniformly

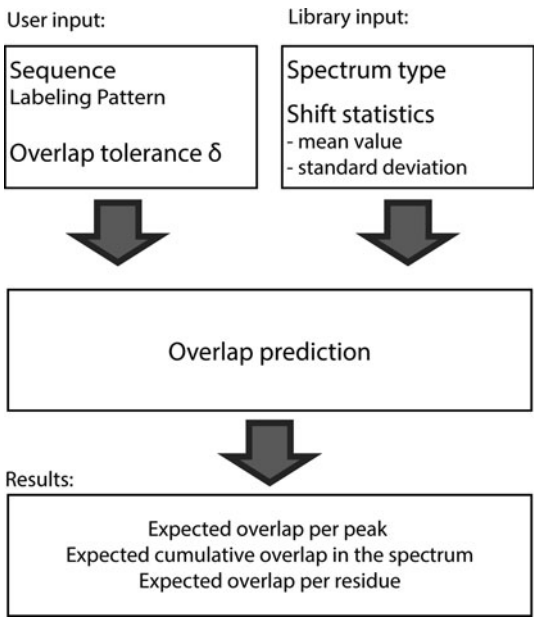

Fig. 1 Flowchart of the overlap prediction procedure

distributed over the entire NMR spectrum. Instead, the chemical shift of atom k is assumed to be distributed normally with mean ω_k and standard deviation σ_k:

$$\mu_N(x;\omega_k,\sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{-\frac{1}{2}\left(\frac{x-\omega_k}{\sigma_k}\right)^2}. \quad (3)$$

The mean value ω_k and standard deviation σ_k are obtained from the shift statistics of the BMRB database that stores for every atom the mean position, standard deviation, and number of occurrences in all protein data sets in the database. The distributions μ_N of Eq. 3 are only reliable if they are based on a sufficient number of chemical shift values. By default, at least 100 measured values were required. Two separate normal distributions are used for the oxidized and reduced forms of cysteine, which the user distinguishes in the protein amino acid sequence by using the residue codes CYSS and CYS for oxidized and reduced cysteine, respectively. Other cases, such as *cis*/*trans* proline, can be handled similarly. The statistics can also be obtained from other sources than the BMRB. For instance, if shifts exist for homologous or otherwise similar proteins, the database can be tailored to a certain class of proteins. Any given set of chemical shift lists and sequences for proteins can be readily processed into a CYANA library with corresponding chemical shift statistics. For the calculations of this paper, we used the general chemical shift statistics from the BMRB database.

Expected peaks

The algorithm estimates overlap probabilities using lists of peaks that are expected based on experiment type-specific magnetization transfer pathways and the covalent structure of the protein (Bartels et al. 1997; Schmidt and Güntert 2012; Schmucki et al. 2009). The magnetization transfer pathways for a spectrum are given as connectivity patterns stored in the CYANA library. For instance, the HNCA spectrum can be described by the magnetization transfer pathways for its intra- and interresidual peaks:

```
SPECTRUM HNCA   HN N C
0.98   HN : H_AMI   N : N_AMI   C : C_ALI   C_BYL
0.80   HN : H_AMI   N : N_AMI   C_BYL   C : C_ALI   N_AMI
```

The first line gives the spectrum name and the atom labels that will be used to identify the respective columns in the peaks lists. The number of atom labels defines the dimensionality n of the spectrum. Each of the following lines specifies a (formal) magnetization transfer pathway, characterized by the probability of the resulting expected peak (not used by the present algorithm) followed by a series of atom types (H_AMI, amide hydrogen; N_AMI, amide nitrogen, C_ALI, aliphatic carbon, C_BYL, carbonyl carbon, etc., as used in the CYANA residue library) that define a molecular pattern of atoms linked by direct covalent bonds. In each pathway the n atoms whose shifts will determine the position of the resulting peak are identified by their corresponding atom labels, followed by ':'. Note that in the case of the HNCA spectrum, the pathways include a "detour" through the carbonyl carbon (C_BYL) to exclude peaks originating from H^{ε}–N^{ε}–C^{δ} in Arg and H^{ζ}–N^{ζ}–C^{ε} in Lys. Through-space type experiments are approximated by the subset of short-range peaks using an extended set of magnetization pathways, which is accurate enough for the present purpose. The magnetization pathway library can be adapted and extended easily.

The peak lists generated by expected peak prediction are "perfect" and contain in general more peaks than can be identified in a real spectrum. Expected peaks are generated only for atoms with, by default, 100 shift values in the BMRB database. Groups of atoms with degenerate chemical shifts, e.g. methyl groups, are represented by a single shift value.

Definition of overlap between two peaks

For the purpose of overlap prediction a peak is considered overlapped if it cannot be resolved from other peaks in n-dimensional space. The most straightforward implementation of this criterion would classify a peak as overlapped if the center of at least one other peak falls within its peak region, and to define the peak region by hard cutoffs for each spectral dimension. However, in order to simplify the theory, we define the ability to distinguish peaks by a Gaussian function of the peak position difference rather than by a fixed "hard" cutoff for this difference because this allows the derivation of analytic expressions for the overlap probabilities (see below). Since the distance between two peaks is not the only factor that decides whether they can be distinguished or not (others including the relative peak intensities, local noise level, peak shape, etc.), the "soft" approach is equally sensible as a hard cutoff, and very similar results are expected for both approaches (see "Results and discussion" below). The probability that two peaks cannot be distinguished from each other in one dimension of a spectrum is defined to be

$$p(\Delta x) = e^{-\frac{1}{2}\left(\frac{\Delta x}{\delta}\right)^2}, \quad (4)$$

where Δx is the difference of the peak positions, and the overlap tolerance δ is a parameter that can be set by the user according to the expected resolution of the spectrum. Equation 4 expresses in a "soft" way the idea that two peaks cannot be distinguished if they are exactly overlapped ($p(0) = 1$), that distinction is difficult for $\Delta x < \delta$, and clear for $\Delta x \gg \delta$.

In principle, the overlap tolerance is related to both the collected and processed digital resolution of the spectra and the relaxation times of the involved nuclei. For convenience, because peak positions and chemical shift values are given in ppm, the overlap tolerance δ is specified in ppm, even though, strictly speaking, it should be expressed in Hz, which is the proper unit for both relaxation and signal truncation linewidth. The default values of the overlap tolerance are $\delta_H = 0.03$ ppm for ^{1}H dimensions and $\delta_N = \delta_C = 0.3$ ppm for ^{15}N and ^{13}C dimensions. The overlap tolerance could be set according to the chemical shift error values in the chemical shift data files of the BMRB database (Ulrich et al. 2008). However, since different ways of setting of the chemical shift error values in the BMRB appear to be in use for different proteins, and because we did not have access to the original spectra, we chose to use uniform values of $\delta_H = 0.03$ ppm for ^{1}H dimensions and $\delta_N = \delta_C = 0.3$ ppm for ^{15}N and ^{13}C dimensions for all calculations in this paper. In practice, the overlap tolerances should be set by visually inspecting the spectra and choosing δ based on the smallest separation between neighboring, distinguishable peaks. In addition, the choice of δ may depend on how the spectra will be used: If it is sufficient to detect the presence of a peak, e.g. for resonance assignment, a smaller overlap tolerance is acceptable than for applications that require accurate peak intensities, e.g. NOESY spectra for the collection of conformational restraints, and even larger overlap tolerances are advisable if the peak shape or peak fine structure are to be analyzed, e.g. for determining scalar coupling constants (Szyperski et al. 1992).

The overlap definition of Eq. 4 can be related to a more traditional overlap definition using a hard cutoff δ' to define overlap when $|\Delta x| < \delta'$. The corresponding probability is $p'(x) = \theta(\delta' - |\Delta x|)$, where θ is the Heaviside step function that equals one for positive arguments and zero otherwise. Equating the integrals over the two probabilities, $\int_{-\infty}^{\infty} p(\mathrm{x})dx = \int_{-\infty}^{\infty} p'(\mathrm{x})dx$, yields the relationship $\delta' = \sqrt{\pi/2}\delta \approx 1.25\delta$ between the two overlap tolerance parameters. This means that the expected overlap computed on the basis of Eq. 4 will be approximately equivalent to the expected overlap computed with a 25 % larger hard cutoff.

Overlap probability for two peaks in one dimension

The overlap definition of Eq. 4 suffices to calculate the number of overlapped peaks in a peak list in which all peak positions are fixed. However, given only the sequence of the protein, estimating the overlap for a list of expected peaks whose position is not yet known requires integration over the chemical shift distributions that describe the expected peak positions, as will be described in the following.

We consider two atoms with chemical shifts that are not known precisely. We assume that they follow normal distributions μ_N according to Eq. 3 with mean values ω_1 and ω_2, and standard deviations σ_1 and σ_2, respectively. The probability P_{deg} that two corresponding peaks in a one-dimensional spectrum, or in one dimension of a multidimensional spectrum, cannot be distinguished is

$$\begin{aligned} P_{\mathrm{deg}}(\omega_1,\sigma_1,\omega_2,\sigma_2,\delta) &= \int_{-\infty}^{\infty} dx_1 \mu_N(x_1;\omega_1,\sigma_1) \\ &\times \int_{-\infty}^{\infty} dx_2 \mu_N(x_2;\omega_2,\sigma_2) p(x_1-x_2) \\ &= \frac{\delta}{\sqrt{\sigma_1^2+\sigma_2^2+\delta^2}} \exp\left(-\frac{1}{2}\left(\frac{\omega_1-\omega_2}{\sqrt{\sigma_1^2+\sigma_2^2+\delta^2}}\right)^2\right) \\ &= \sqrt{2\pi}\delta\mu_N\left(\omega_1-\omega_2;0,\sqrt{\sigma_1^2+\sigma_2^2+\delta^2}\right), \end{aligned} \tag{5}$$

where $p(x_1 - x_2)$ is the probability of Eq. 4 that two signals with chemical shift difference $x_1 - x_2$ cannot be distinguished, and it is assumed that the distributions of the two chemical shifts are independent. Substituting $\Delta\omega = \omega_1 - \omega_2$ for the difference between the mean values of the two atom chemical shifts and $\sigma = \sqrt{\sigma_1^2+\sigma_2^2}$ for the geometric mean of their standard deviations, Eq. 5 becomes

$$\begin{aligned} P_{\mathrm{deg}}(\Delta\omega,\sigma,\delta) &= \frac{\delta}{\sqrt{\sigma^2+\delta^2}} \exp\left(-\frac{1}{2}\left(\frac{\Delta\omega}{\sqrt{\sigma^2+\delta^2}}\right)^2\right) \\ &= \sqrt{2\pi}\delta\mu_N(\Delta\omega;0,\sqrt{\sigma^2+\delta^2}), \end{aligned} \tag{6}$$

For $\sigma \gg \delta$, as is usually the case, this further simplifies to

$$\begin{aligned} P_{\mathrm{deg}}(\Delta\omega,\sigma,\delta) &\approx \frac{\delta}{\sigma} \exp\left(-\frac{1}{2}\left(\frac{\Delta\omega}{\sigma}\right)^2\right) \\ &= \sqrt{2\pi}\delta\,\mu_N(\Delta\omega;0,\sigma). \end{aligned} \tag{7}$$

The overlap probability P_{deg} of Eq. 6 can be expressed as a function of only two variables, the dimensionless quantities $\Delta\omega/\delta$ and σ/δ,

$$\begin{aligned} &P_{deg}(\Delta\omega/\delta,\sigma/\delta) \\ &= \frac{1}{\sqrt{1+(\sigma/\delta)^2}} \exp\left(-\frac{1}{2}\left(\frac{\Delta\omega/\delta}{\sqrt{1+(\sigma/\delta)^2}}\right)^2\right) \\ &= \sqrt{2\pi}\,\mu_N(\Delta\omega/\delta;0,\sqrt{1+(\sigma/\delta)^2}). \end{aligned} \tag{8}$$

P_{deg} is an exponentially decaying function of $\Delta\omega$, but decreases only slowly with increasing σ (Fig. 2).

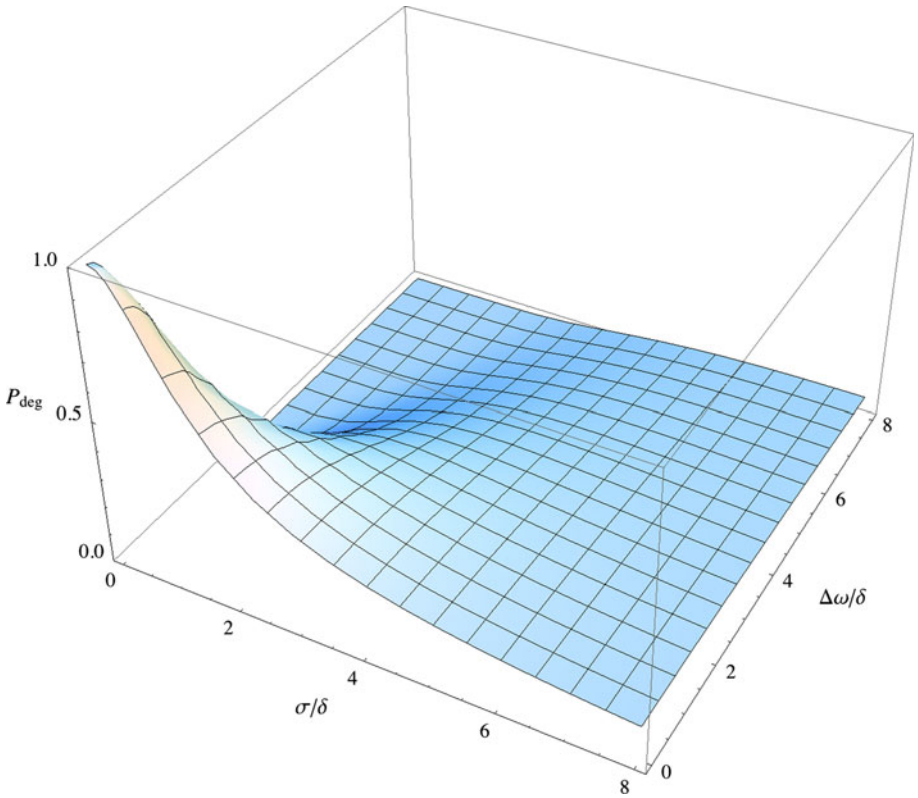

Fig. 2 Overlap probability $P_{\mathrm{deg}}(\Delta\omega/\delta, \sigma/\delta)$ of Eq. 8 for two peaks in one dimension. The chemical shifts of the two atoms are assumed to be normally distributed with mean values ω_1 and ω_2, and standard deviations σ_1 and σ_2, respectively; $\Delta\omega = \omega_1 - \omega_2$ is the difference between the mean values of the two atom chemical shifts, $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$ is the geometric mean of their standard deviations, and δ is the overlap tolerance parameter introduced in Eq. 4

The approach can also be used if the position of one of the two peaks is already known by setting the corresponding standard deviation to zero, e.g. $\sigma_1 = 0$. Equation 8 remains valid with $\sigma = \sigma_2$. If the positions of both peaks are fixed, $\sigma = \sigma_1 = \sigma_2 = 0$, and Eq. 8 reduces to Eq. 4.

Overlap probability for two peaks in n dimensions

Two peaks cannot be distinguished in an n-dimensional spectrum if they overlap in each of the n dimensions. Their overlap probability therefore becomes:

$$P_{\mathrm{deg}}^{(n)}\left(\frac{\Delta\omega^{(1)}}{\delta^{(1)}}, \ldots, \frac{\Delta\omega^{(n)}}{\delta^{(n)}}; \frac{\sigma^{(1)}}{\delta^{(1)}}, \ldots, \frac{\sigma^{(n)}}{\delta^{(n)}}\right) = \prod_{i=1}^{\prime n} P_{\mathrm{deg}}\left(\frac{\Delta\omega^{(i)}}{\delta^{(i)}}, \frac{\sigma^{(i)}}{\delta^{(i)}}\right)$$

To account for the fact that peaks assigned to the same atom must be aligned in the corresponding dimension, these peaks are considered as fully overlapped in the respective dimension and the corresponding P_{deg} term is omitted from the product.

Overlap probability for N peaks in n dimensions

In a data set of N peaks in an n-dimensional spectrum, the probability $P_{\mathrm{deg}}^{(\mathrm{tot})}$ that a given peak j overlaps with one or more other peaks is the complement of the probability that it does not overlap with any other peak:

$$P_{\mathrm{deg}}^{(\mathrm{tot})}(j) = 1 - \prod_{k=1}^{\prime N}\left[1 - P_{\mathrm{deg}}^{(n)}\left(\frac{\Delta\omega_{jk}^{(1)}}{\delta^{(1)}}, \ldots, \frac{\Delta\omega_{jk}^{(n)}}{\delta^{(n)}}; \frac{\sigma_{jk}^{(1)}}{\delta^{(1)}}, \ldots, \frac{\sigma_{jk}^{(n)}}{\delta^{(n)}}\right)\right]. \quad (9)$$

The product in Eq. 9 runs over all peaks other than peak j; $\Delta\omega_{jk}^{(i)} = \omega_j^{(i)} - \omega_k^{(i)}$ and $\sigma_{jk}^{(i)} = \sqrt{\sigma_j^{(i)2} + \sigma_k^{(i)2}}$. The expected total number of overlapping peaks thus becomes

$$N_{\mathrm{deg}} = \sum_{j=1}^{N} P_{\mathrm{deg}}^{(\mathrm{tot})}(j). \quad (10)$$

In the special case of equal overlap probabilities $P_{\mathrm{deg}}^{(n)}$ for all peak pairs, Eqs. 9 and 10 reduce to Eq. 1 with $P_{\mathrm{deg}}^{(n)} = \gamma/\Gamma$. Thus, the present theory is a generalization of the simple earlier approaches (Kainosho et al. 2006; Mumenthaler et al. 1997).

Verification by Monte-Carlo simulation

Monte-Carlo simulation was used to verify the correctness of the theory of Eqs. 3–10 by simulating for a given sequence a large number of peak lists according to the model of Eq. 3. Shift positions of atoms were sampled from normal

distributions with mean values and standard deviations corresponding to the shift statistics, and peak lists were generated according to the magnetization transfer pathways of the NMR experiment. A peak pair was considered to be overlapped with the probability of Eq. 4. The procedure was repeated 50,000 times and the average overlap probability was compared with the analytical result of Eq. 10.

Normally distributed random numbers were generated by the transformation method (Press et al. 1986): Two random numbers x_1, x_2, distributed uniformly in the interval $[-1, 1]$, are generated. If $r^2 = x_1^2 + x_2^2 > 1$, they are rejected, and a new pair of random numbers is generated. Otherwise, two normally distributed random numbers u_1, u_2 are obtained as $u_{1,2} = x_{1,2}\sqrt{-2\log(r^2)/r^2}$.

Test data sets

The algorithm was evaluated for eight different proteins to which we refer in this paper by four-letter acronyms (Table 1): CPRP, the chicken prion protein fragment 128–242 (Calzolai et al. 2005); ENTH, the ENTH-VHS domain At3g16270 from *Arabidopsis thaliana* (López-Méndez and Güntert 2006; López-Méndez et al. 2004); FSH2, the Src homology 2 domain from the human feline sarcoma oncogene Fes (Scott et al. 2004, 2005); FSPO, the F-spondin TSR domain 4 (Pääkkönen et al. 2006); PBPA, the *Bombyx mori* pheromone binding protein (Horst et al. 2001); RHOD, the rhodanese homology domain At4g01050 from *Arabidopsis thaliana* (Pantoja-Uceda et al. 2004, 2005); SCAM, stereo-array isotope labeled (SAIL) calmodulin (Kainosho et al. 2006); DSRP, the delta subunit of RNA polymerase from *Bacillus subtilis* (Motáčková et al. 2010). The proteins CPRP, ENTH, PBPA, and SCAM are predominantly α-helical; FSH2, and RHOD have mixed α/β secondary structure. The protein FSPO has an unusual fold with little regular secondary structure (Pääkkönen et al. 2006). The protein SCAM has two domains connected by a flexible linker; DSRP is an intrinsically disordered protein that contains a disordered C-terminal region of 81 amino acids with a highly repetitive sequence; all others have a well-defined single-domain structure.

In addition, overlap prediction was also carried out for the [^{1}H,^{15}N]-HSQC spectra of 2,174 proteins for which chemical shift assignments are available from the BMRB that are sufficiently complete to assign more than 70 % of the expected peaks.

Results and discussion

Our goal was to provide a flexible and user-friendly algorithm that is capable of predicting spectral overlap in NMR spectra and that can estimate the usefulness of labeling schemes, given a specific sequence, prior to producing samples and measuring NMR spectra. Overlap prediction for a spectrum with several hundred peaks takes about 2 s on a standard desktop computer with 2.4 GHz Intel processor. The maximal runtime of 28 s was measured for a TOCSY spectrum with several thousand peaks for the largest protein in the BMRB.

Table 1 Overview of protein data sets used for overlap prediction

Acronym	PDB code	BMRB code	Amino acids	Assignment completeness (%)
FSPO	1VEX	10002	56	98.6
FSH2	1WQU	6331	114	97.2
CPRP	1U3M	6269	117	97.8
RHOD	1VEE	5929	134	98.4
ENTH	1VDY	5928	140	96.0
PBPA	1GM0	4849	142	99.3
SCAM	1X02	6541	148	100.0
DSRP	2KRC	16912	172	98.4

The assignment completeness gives the percentage of the aliphatic ^{1}H and backbone $^1H^N$ resonances that are assigned

Measured and predicted overlap in a [^{1}H,^{15}N]-HSQC spectrum

As a first test application the predicted overlap was compared to the overlap observed in the experimental [^{1}H,^{15}N]-HSQC of the protein RHOD, for which the chemical shift assignments and the experimental peak list are available (Fig. 3a). Expected peaks were generated using the magnetization transfer rules in the CYANA library (Schmidt and Güntert 2012; Schmucki et al. 2009), and the overlap probability was calculated for the peaks at the positions given by the experimental chemical shift ("measured overlap", Fig. 3b) and by Eq. 9 without knowledge of the peak positions ("predicted overlap", Fig. 3c). In both cases most overlap occurs in the same regions of the spectrum. As expected, the predicted overlap is distributed over many peaks in the crowded regions, whereas the measured overlap affects specific peaks. In principle, the experimental spectrum is an instance taken from the general distribution over which overlap prediction by Eq. 9 is averaging. Overlap prediction is able to distinguish peaks in crowded regions from those in better resolved regions and could thus be used to optimize a labeling pattern that reduces the peak overlap without undue loss of signals.

The effect of additional dimensions on the overlap

Higher-dimensional NMR spectroscopy reduces overlap significantly. To show that the algorithm correctly predicts

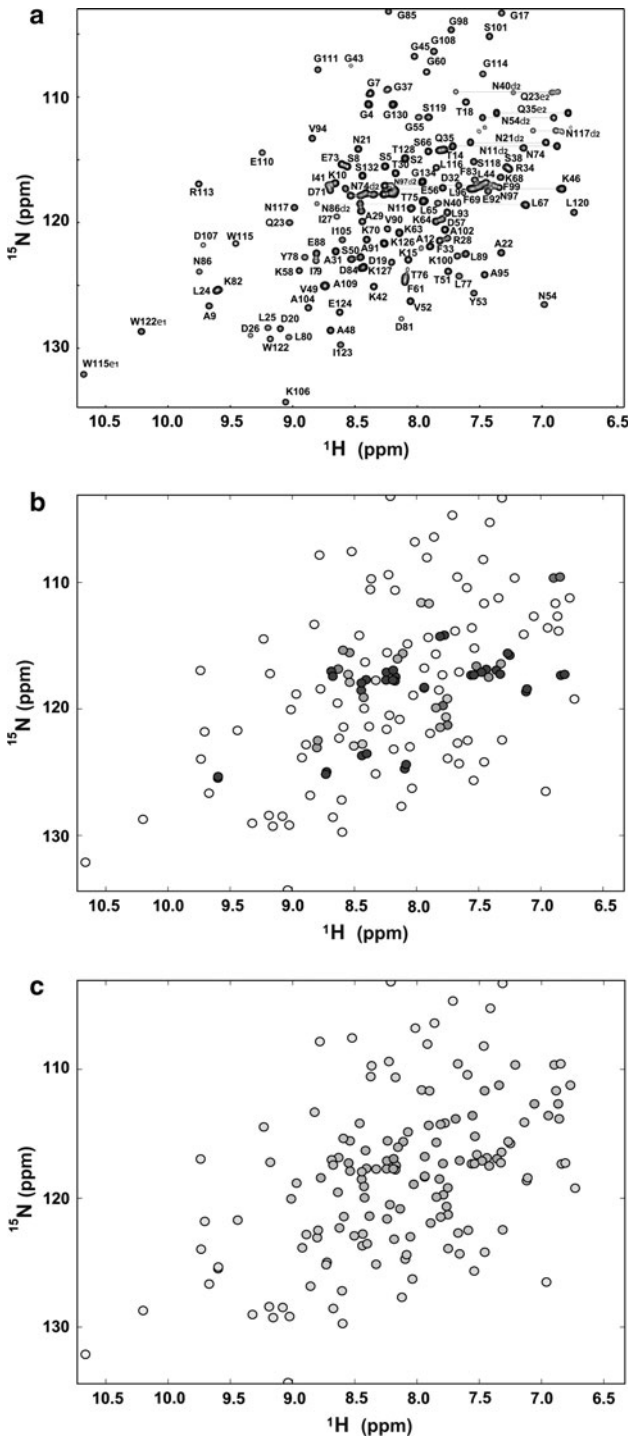

Fig. 3 Overlap in the [^{1}H,^{15}N]-HSQC spectrum of RHOD. **a** Experimental spectrum (Pantoja-Uceda et al. 2004). **b** Spectrum simulated using the experimental chemical shifts. Signals are colored from *white* to *black* with increasing overlap calculated for the fixed peak positions using Eq. 4. **c** Same spectrum as in **b**, colored according to the overlap probability predicted by Eq. 9 using only the sequence and spectrum type information

this behavior we compared overlap predictions for the protein RHOD using two pairs of corresponding two- and three-dimensional spectra, i.e. 2D NOESY versus 3D ^{13}C-resolved NOESY, and [^{1}H,^{15}N]-HSQC versus HNCA (Fig. 4). The overlap predicted using Eq. 9 is strongly reduced by the presence of the extra dimension, especially in case of 2D NOESY (Fig. 4a) versus 3D NOESY (Fig. 4b). [^{1}H,^{15}N]-HSQC (Fig. 4c) and HNCA (Fig. 4d) show less overlap overall, but again the introduction of the third dimension in the HNCA removes most of the signal overlap present in the [^{1}H,^{15}N]-HSQC spectrum.

Overlap prediction for spectra of a test set of eight proteins

To show the overlap prediction with a variety of different types of spectra, eight proteins were analyzed for which chemical shift assignments are available (Table 1). The amount of overlap was predicted by Eq. 9 based only on the sequence and the general chemical shift statistics of the BMRB (blue bars in Fig. 5) and compared to the overlap measured on the basis of the known chemical shift assignments using the chemical shift list of the given protein from the BMRB (green crosses in Fig. 5). For comparison, the percentage of overlap and its standard deviation were also predicted using the Monte Carlo method (blue dots and error bars in Fig. 5).

The overlap measured on the basis of the known chemical shift assignments (green crosses in Fig. 5) and the overlap predicted from the sequence (red dots in Fig. 5) are highly correlated for the spectra of a given protein with Pearson correlation coefficients of 0.84–0.98 (significance <0.00011 in all cases). As expected, overlap increases with protein size. Among the spectra analyzed for any given protein, the overlap is in general largest for the homonuclear 2D spectra, and smallest for triple resonance backbone assignment spectra. Generally, the overlap probability increases with the number of peaks in a spectrum, although this is not universal. The overlap prediction depicts faithfully differences in the measured overlap between different spectra. For longer proteins the predictions appear to become more accurate, which may be an effect of the law of large numbers and the central limit theorem from which it follows that the more peaks are analyzed, the better the assumptions of the theory are fulfilled. As expected, the intrinsically disordered protein DSRP is an exception in that the measured overlap exceeds significantly the one predicted on the basis of the general chemical shift statistics, which is derived almost exclusively from folded globular proteins. It will thus be necessary to derive separate chemical shift statistics for intrinsically disordered proteins in order to obtain more realistic results for this class of proteins. At present, the scarcity of chemical shift assignments available for intrinsically disordered proteins does not yet provide reliable statistics.

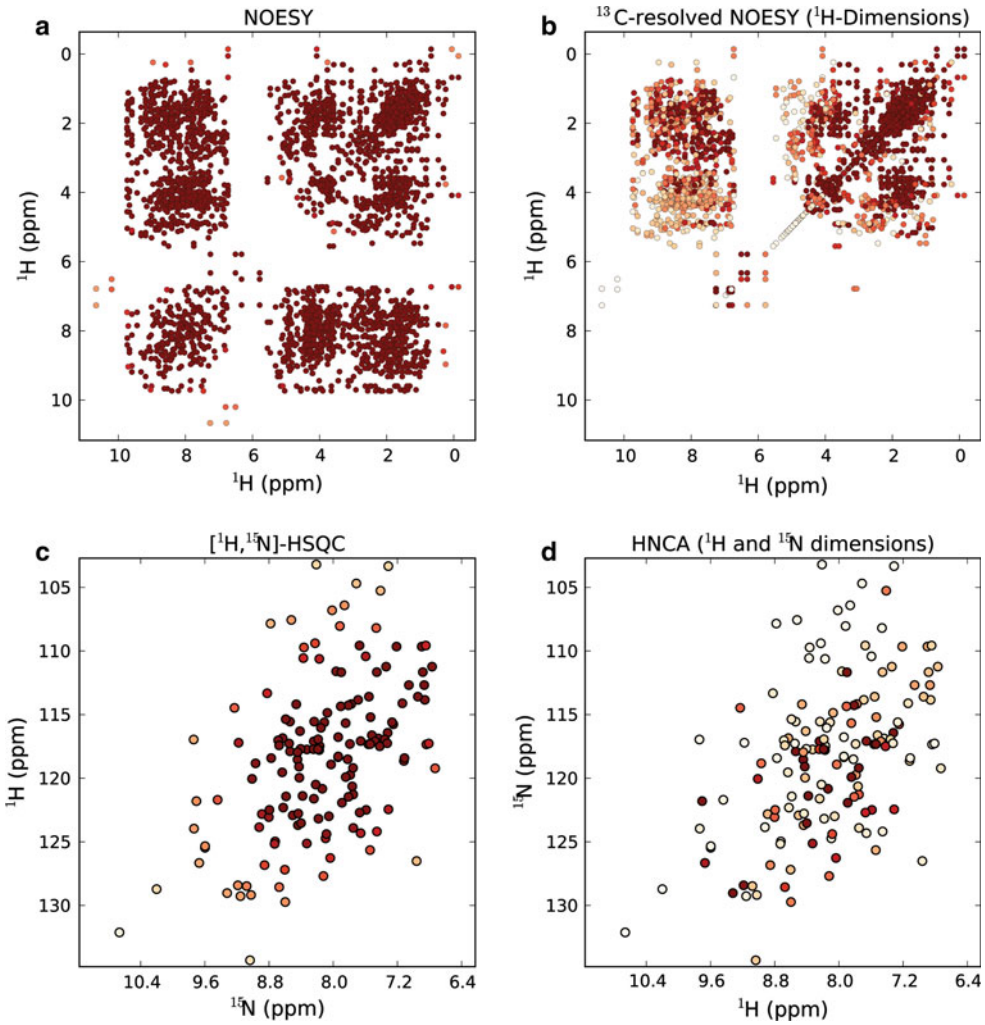

Fig. 4 Overlap comparison for spectra with different numbers of dimensions for the protein RHOD. **a** 2D homonuclear NOESY spectrum. **b** 3D ^{13}C-resolved NOESY spectrum. **c** 2D [^{1}H,^{15}N]-HSQC spectrum. **d** 3D HNCA spectrum, projected onto the [^{1}H,^{15}N]-plane. Signals are colored in *red* from *white* to *black* with increasing overlap predicted by Eq. 9 using only the sequence and general chemical shift statistics. The peak positions correspond to the known chemical shift assignments for RHOD

The correctness of the overlap prediction by the analytic formulas of Eqs. 8–10 (red in Fig. 5) was verified by Monte Carlo simulation (blue in Fig. 5). The average overlap values obtained by Monte Carlo simulation are always in close agreement with the analytical result. The standard deviation is often considerable, indicating that the amount of overlap observed for a single given protein can deviate significantly from the analytical average result even if the chemical shift values of the atoms follow the assumed normal distributions.

Analyzing proteins in the BMRB

As a large-scale application, we calculated for all 2,174 proteins with sufficiently complete chemical shift entries in the BMRB the extent of overlap in [^{1}H,^{15}N]-HSQC spectra by prediction based on the sequence alone and, for comparison, by measurement based on the chemical shift assignments from the BMRB (Fig. 6). This provided a means to investigate the prediction power of our method for a large variety of proteins and to rationalize the use of the soft overlap criterion of Eq. 4. Figure 6a shows that for all proteins the use of a hard cutoff or the "soft" criterion of Eq. 4 yielded very similar results. Figure 6b shows a comparison of the predicted and measured numbers of overlapped peaks. Overall, they are correlated with a correlation coefficient of 0.79 (significance $<10^{-10}$). The spread between for individual proteins is comparable to the standard deviation seen in the Monte Carlo simulation results depicted in Fig. 5. In addition, there are some proteins for which the measured amount of overlap exceeds the predicted overlap considerably. The manual inspection of individual cases showed that these correspond either to intrinsically unfolded proteins, similar to the example of DSRP in Fig. 5, or to symmetric multimers. In principle, more realistic prediction results could be obtained for the former by using a separate chemical shift statistics restricted to intrinsically unfolded proteins, and for the latter by explicitly taking into account the symmetry in a modified theory.

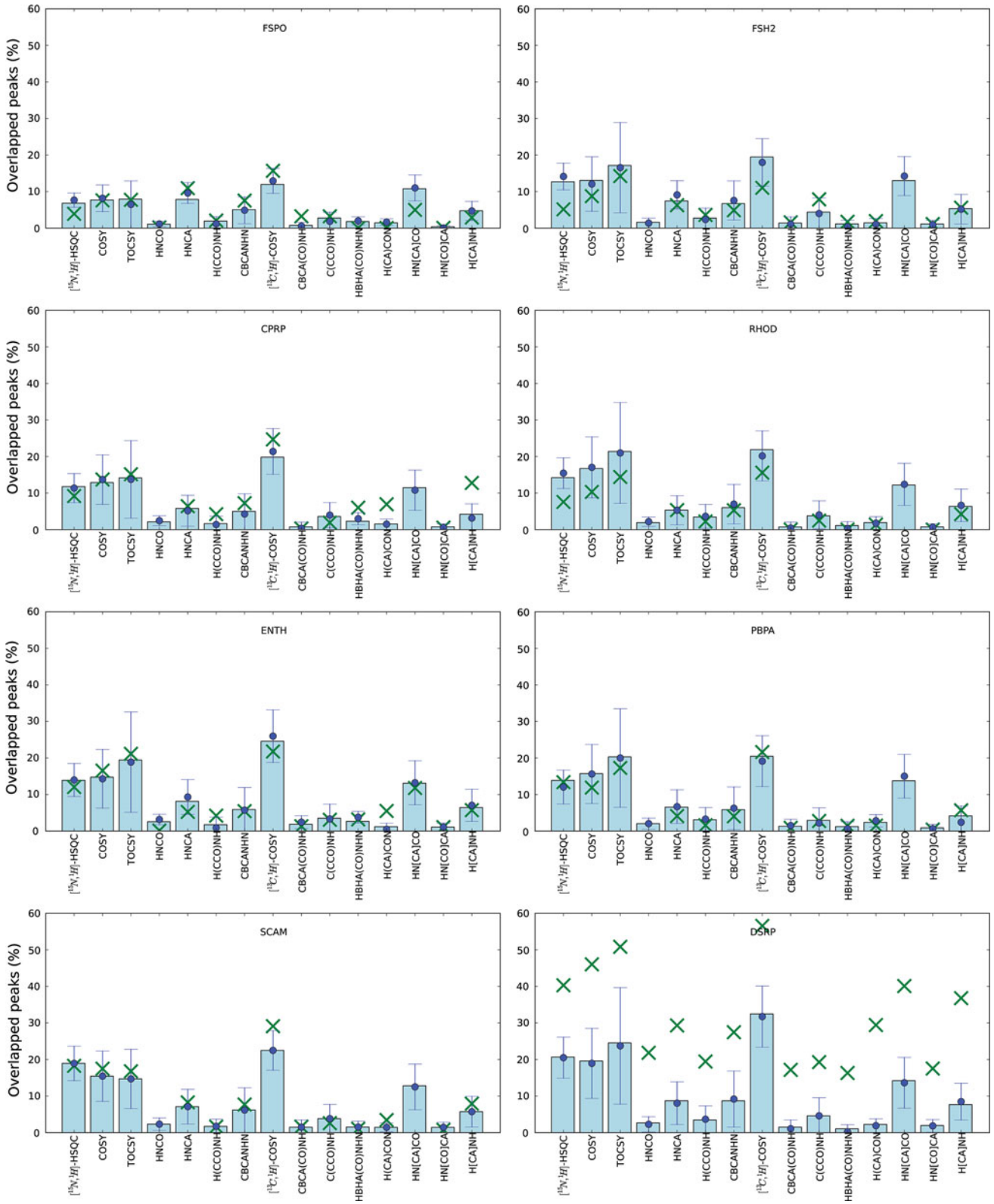

Fig. 5 Overlap prediction and measurement for eight proteins, one of which (DSRP) is intrinsically unstructured. The percentage of overlapped peaks predicted from the sequence alone using Eqs. 9–10 is shown as *blue bars*. The average value and the standard deviation of the predicted overlap obtained by Monte Carlo simulation are shown in *blue*. The measured overlap percentage obtained by applying Eq. 4 to the expected peaks at the positions given by the known experimental chemical shift assignments is indicated by *green crosses*. Where experimental assignments for a certain class of nuclei, e.g. carbonyls, were not available, only the predicted overlap for all atoms is reported

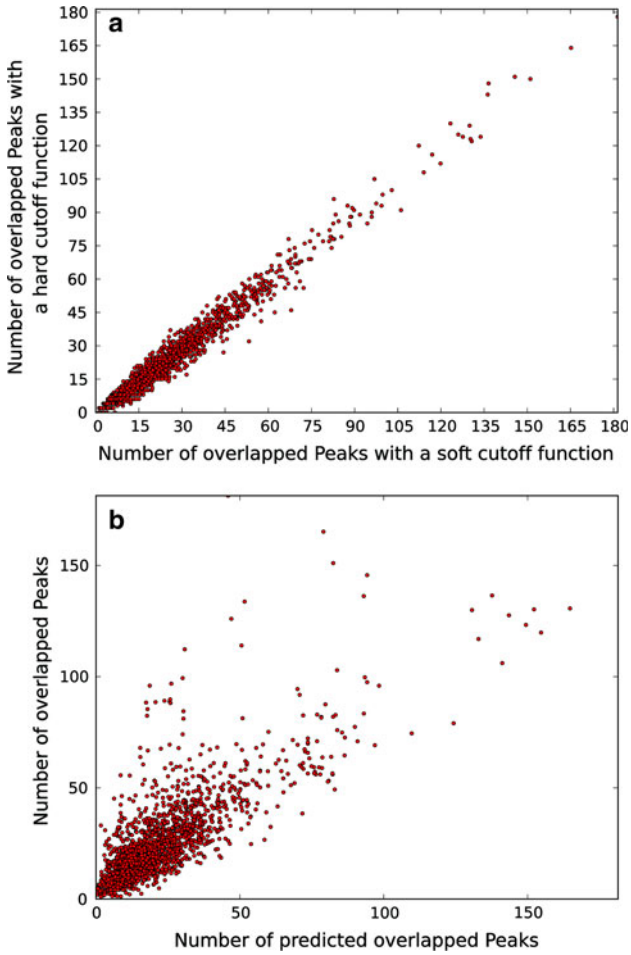

Fig. 6 Number of overlapped peaks in [^{1}H,^{15}N]-HSQC spectra for 2,174 proteins with chemical shift assignments from the BMRB that are sufficiently complete such that more than 70 % of the [^{1}H,^{15}N]-HSQC peaks are assigned. **a** Overlap measured by a hard cutoff $\delta' = 1.25\delta$ (see "Materials and methods"), where δ is the overlap tolerance, for the expected peaks at the positions given by the experimental chemical shift assignments plotted versus the overlap measured for the same peaks using the "soft" Gaussian overlap probability function of Eq. 4. **b** Overlap measured using the soft criterion for the same peaks as in panel a plotted versus the overlap predicted from the sequence alone using Eqs. 9–10

Conclusions

In this paper we have introduced a new general method for estimating the overlap of peaks in NMR spectra that can be applied already if only the sequence of the protein is known, e.g. before starting sample preparation and NMR measurements. Results for the average overlap are in agreement with the amount of overlap measured in experimental spectra although the method can obviously not predict with certainty whether an individual peak will be overlapped or not. The overlap estimation can be used to distinguish proteins with potentially heavily overlapped spectra from those with better chemical shift dispersion based on primary structure information alone.

Overlap estimation can be used, for instance, to support setting proper signal sampling parameters for NMR experiments, e.g. the number of dimensions, maximum evolution times, use of linear prediction, non-uniform sampling and other resolution-improving techniques. Overlap prediction can support the design of overlap-optimized labeling schemes. For a given sequence and a given number of amino acid types that are to be labeled, these can be chosen so that the predicted amount of overlap is minimal, while preserving the maximal information possible. Consequently, there will not be a unique optimal solution, but rather a set of efficient solutions that are characterized by the fact that their overlap cannot be improved further without losing information. It can also be envisaged to use overlap prediction in automated assignment algorithms, e.g. to define a priori probabilities for the observation of peaks, for locally steering peak picking algorithms, and for weighing peak assignments and penalties for peak degeneracy in scoring functions for assignments (Schmidt and Güntert 2012). Similar applications are conceivable for the identification of conformational restraints for structure calculations. Conformational restraints derived from peaks in less overlapped regions are potentially safer to introduce into the structure calculation. A priori overlap prediction based on the present theory can therefore play a role in improving the reliability of automated spectra analysis and protein structure determination.

Acknowledgments We gratefully acknowledge financial support by the Lichtenberg program of the Volkswagen Foundation.

References

Bartels C, Güntert P, Billeter M, Wüthrich K (1997) GARANT—a general algorithm for resonance assignment of multidimensional nuclear magnetic resonance spectra. J Comput Chem 18: 139–149

Busche AE, Aranko AS, Talebzadeh-Farooji M, Bernhard F, Dötsch V, Iwai H (2009) Segmental isotopic labeling of a central domain in a multidomain protein by protein trans-splicing using only one robust DnaE intein. Angew Chem 48:6128–6131

Calzolai L, Lysek DA, Perez DR, Güntert P, Wüthrich K (2005) Prion protein NMR structures of chickens, turtles, and frogs. Proc Natl Acad Sci USA 102:651–655

Ernst RR, Bodenhausen G, Wokaun A (1987) The principles of nuclear magnetic resonance in one and two dimensions. Clarendon Press, Oxford

Goto NK, Kay LE (2000) New developments in isotope labeling strategies for protein solution NMR spectroscopy. Curr Opin Struct Biol 10:585–592

Güntert P (2009) Automated structure determination from NMR spectra. Eur Biophys J 38:129–143

Güntert P, Mumenthaler C, Wüthrich K (1997) Torsion angle dynamics for NMR structure calculation with the new program DYANA. J Mol Biol 273:283–298

Hefke F, Bagaria A, Reckel S, Ullrich SJ, Dötsch V, Glaubitz C, Güntert P (2011) Optimization of amino acid type- specific ^{13}C and ^{15}N labeling for the backbone assignment of membrane proteins by solution- and solid-state NMR with the UPLABEL algorithm. J Biomol NMR 49:75–84

Higman VA, Flinders J, Hiller M, Jehle S, Markovic S, Fiedler S, van Rossum BJ, Oschkinat H (2009) Assigning large proteins in the solid state: a MAS NMR resonance assignment strategy using selectively and extensively ^{13}C-labelled proteins. J Biomol NMR 44:245–260

Horst R, Damberger F, Luginbühl P, Güntert P, Peng G, Nikonova L, Leal WS, Wüthrich K (2001) NMR structure reveals intramolecular regulation mechanism for pheromone binding and release. Proc Natl Acad Sci USA 98:14374–14379

Kainosho M, Güntert P (2009) SAIL—stereo-array isotope labeling. Q Rev Biophys 42:247–300

Kainosho M, Torizawa T, Iwashita Y, Terauchi T, Ono AM, Güntert P (2006) Optimal isotope labelling for NMR protein structure determinations. Nature 440:52–57

Lian LY, Middleton DA (2001) Labelling approaches for protein structural studies by solution-state and solid-state NMR. Prog Nucl Magn Reson Spectrosc 39:171–190

López-Méndez B, Güntert P (2006) Automated protein structure determination from NMR spectra. J Am Chem Soc 128: 13112–13122

López-Méndez B, Pantoja-Uceda D, Tomizawa T, Koshiba S, Kigawa T, Shirouzu M, Terada T, Inoue M, Yabuki T, Aoki M, Seki E, Matsuda T, Hirota H, Yoshida M, Tanaka A, Osanai T, Seki M, Shinozaki K, Yokoyama S, Güntert P (2004) NMR assignment of the hypothetical ENTH-VHS domain At3g16270 from *Arabidopsis thaliana*. J Biomol NMR 29:205–206

Motáčková V, Nováček J, Zawadzka-Kazimierczuk A, Kazimierczuk K, Žídek L, Šanderová H, Krásný L, Koźmiński W, Sklenář V (2010) Strategy for complete NMR assignment of disordered proteins with highly repetitive sequences based on resolution-enhanced 5D experiments. J Biomol NMR 48:169–177

Mumenthaler C, Güntert P, Braun W, Wüthrich K (1997) Automated combined assignment of NOESY spectra and three-dimensional protein structure determination. J Biomol NMR 10:351–362

Oxenoid K, Kirn HJ, Jacob J, Sönnichsen FD, Sanders CR (2004) NMR assignments for a helical 40 kDa membrane protein. J Am Chem Soc 126:5048–5049

Pääkkönen K, Tossavainen H, Permi P, Rakkolainen H, Rauvala H, Raulo E, Kilpeläinen I, Güntert P (2006) Solution structures of the first and fourth TSR domains of F-spondin. Proteins 64:665–672

Pantoja-Uceda D, López-Méndez B, Koshiba S, Kigawa T, Shirouzu M, Terada T, Inoue M, Yabuki T, Aoki M, Seki E, Matsuda T, Hirota H, Yoshida M, Tanaka A, Osanai T, Seki M, Shinozaki K, Yokoyama S, Güntert P (2004) NMR assignment of the hypothetical rhodanese domain At4g01050 from *Arabidopsis thaliana*. J Biomol NMR 29:207–208

Pantoja-Uceda D, López-Méndez B, Koshiba S, Inoue M, Kigawa T, Terada T, Shirouzu M, Tanaka A, Seki M, Shinozaki K, Yokoyama S, Güntert P (2005) Solution structure of the rhodanese homology domain At4g01050(175–295) from *Arabidopsis thaliana*. Protein Sci 14:224–230

Press WH, Flannery BP, Teukolsky SA, Vetterling WT (1986) Numerical recipes. The art of scientific computing. Cambridge University Press, Cambridge

Reckel S, Sobhanifar S, Schneider B, Junge F, Schwarz D, Durst F, Löhr F, Güntert P, Bernhard F, Dötsch V (2008) Transmembrane segment enhanced labeling as a tool for the backbone assignment of α-helical membrane proteins. Proc Natl Acad Sci USA 105:8262–8267

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

Schmucki R, Yokoyama S, Güntert P (2009) Automated assignment of NMR chemical shifts using peak-particle dynamics simulation with the DYNASSIGN algorithm. J Biomol NMR 43:97–109

Scott A, Pantoja-Uceda D, Koshiba S, Inoue M, Kigawa T, Terada T, Shirouzu M, Tanaka A, Sugano S, Yokoyama S, Güntert P (2004) NMR assignment of the SH2 domain from the human feline sarcoma oncogene FES. J Biomol NMR 30:463–464

Scott A, Pantoja-Uceda D, Koshiba S, Inoue M, Kigawa T, Terada T, Shirouzu M, Tanaka A, Sugano S, Yokoyama S, Güntert P (2005) Solution structure of the Src homology 2 domain from the human feline sarcoma oncogene Fes. J Biomol NMR 31: 357–361

Szántay C Jr (2007) NMR and the uncertainty principle: how to and how not to interpret homogeneous line broadening and pulse nonselectivity. I. The fundamentals. Concepts in Magnetic Resonance Part A 30A:309–348

Szyperski T, Güntert P, Otting G, Wüthrich K (1992) Determination of scalar coupling constants by inverse fourier transformation of in-phase multiplets. J Magn Reson 99:552–560

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z, Nakatani E, Schulte CF, Tolmie DE, Wenger RK, Yao HY, Markley JL (2008) BioMagResBank. Nucleic Acids Res 36:D402–D408

Yamazaki T, Otomo T, Oda N, Kyogoku Y, Uegaki K, Ito N, Ishino Y, Nakamura H (1998) Segmental isotope labeling for protein NMR using peptide splicing. J Am Chem Soc 120:5591–5592