J Biomol NMR (2017) 69:147–156
DOI 10.1007/s10858-017-0145-2

ARTICLE

# The PROSECCO server for chemical shift predictions in ordered and disordered proteins

**Máximo Sanz-Hernández**[1] · **Alfonso De Simone**[1]

Received: 30 August 2017 / Accepted: 12 October 2017 / Published online: 8 November 2017


**Abstract** The chemical shifts measured in solution-state and solid-state nuclear magnetic resonance (NMR) are powerful probes of the structure and dynamics of protein molecules. The exploitation of chemical shifts requires methods to correlate these data with the protein structures and sequences. We present here an approach to calculate accurate chemical shifts in both ordered and disordered proteins using exclusively the information contained in their sequences. Our sequence-based approach, protein sequences and chemical shift correlations (PROSECCO), achieves the accuracy of the most advanced structure-based methods in the characterization of chemical shifts of folded proteins and improves the state of the art in the study of disordered proteins. Our analyses revealed fundamental insights on the structural information carried by NMR chemical shifts of structured and unstructured protein states.

**Keywords** Biomolecular NMR · Chemical shift predictions · Disordered proteins

**Abbreviations**

| | |
|---|---|
| CS | Chemical shifts |
| IDPs | Intrinsically disordered proteins |
| IDRs | Intrinsically disordered regions |
| NMR | Nuclear magnetic resonance |
| PROSECCO | Protein sequences and chemical shift correlations |

**Electronic supplementary material** The online version of this article (doi:10.1007/s10858-017-0145-2) contains supplementary material, which is available to authorized users.

✉ Alfonso De Simone
adesimon@imperial.ac.uk

[1] Department of Life Sciences, Imperial College London, London SW7 2AZ, UK

## Introduction

Biomolecular nuclear magnetic resonance (NMR) spectroscopy has emerged as a powerful technique to accurately characterize the structure and dynamics of proteins and other biomacromolecules (Kay 2005). In the context of intrinsically disordered proteins (IDPs), NMR has a unique ability to quantify transient interactions and conformational preferences of such elusive protein states both in vitro (Jensen et al. 2010; Stollar et al. 2012) and in the cellular environment (Felli et al. 2014; Waudby et al. 2013). The quantitative interpretation of NMR spectra toward the characterization of the structural properties and atomic fluctuations of protein molecules has attracted considerable interest in the biochemical community. In this context, significant progress has been achieved by using statistical mechanics to analyze NMR databases, which enabled the definition of new approaches to study protein structure (Bouvignies et al. 2011; Cavalli et al. 2007; Hafsa et al. 2015; Shen et al. 2008) and dynamics (Berjanskii and Wishart 2005, 2013; Boulton et al. 2014; Kim et al. 2017; Krieger et al. 2014; Masterson et al. 2010; Neudecker et al. 2012; Robustelli et al. 2012; Selvaratnam et al. 2011) such as those employing exclusively chemical shifts (CS) from solution (Clore and Schwieters 2003; Kuszewski et al. 2004; Sgourakis et al. 2011) and solid state NMR (Mollica et al. 2012; Robustelli et al. 2008). Indeed CS are extremely precise probes of the secondary structures in folded and disordered proteins and have been largely used to map backbone dihedral angles (Neal et al. 2006; Shen and Bax 2015a).

The interpretation of CS of folded proteins is based on the ability to correlate these experimental data with protein three-dimensional structure (Han et al. 2011; Kohlhoff et al. 2009; Li and Brüschweiler 2012, 2015; Meiler 2003; Shen and Bax 2007, 2010; Xu and Case 2001) and vice versa (Berjanskii et al. 2015; Shen and Bax 2015b; Shen et al. 2009), whereas in the context of IDPs these observables are primarily correlated with the protein sequence (Kjaergaard and Poulsen 2011; Schwarzinger et al. 2001; De Simone et al. 2009; Tamiola et al. 2010; Wang and Jardetzky 2002; Wishart et al. 1995). In the present study, we defined a method to generate CS tables in structured and disordered proteins by using exclusively the information contained in their sequences. This structure-free approach, protein sequences and chemical shift correlations (PROSECCO), was derived from the analysis of large experimental datasets by using a unique statistical approach that in the case of IDPs enabled to improve the state of the art of the prediction of CS and in the case of structured proteins achieved the accuracy levels of powerful structure-based methods. In addition to providing a tool for predicting CS using as input exclusively the protein sequence, the parameterization of PROSECCO has revealed key insights into the structural dependences of protein chemical shifts.

## Results

We used sequence homology criteria to select a database of non-redundant proteins whose chemical shifts were deposited in the biological magnetic resonance data bank (BMRB) (Ulrich et al. 2008) (BMRB, see "Materials and methods"). In total, 20,154 experimental chemical shifts of atoms from disordered proteins were used to derive a CS predictor for IDPs—PROSECCO$_{IDP}$—and 3,953,878 experimental data were employed to generate a sequence-based CS predictor for structured proteins—PROSECCO$_{FOLDED}$—(see "Materials and methods" and supplementary Tables S1–S3).

### Gaussian kernel-based neighbor correction in CS prediction of IDPs

We previously showed that loop regions of natively folded proteins are excellent models to describe the CS of random coil protein states (De Simone et al. 2009). In defining PROSECCO$_{IDP,}$ we here identified a different approach that is based on the analysis of experimental CS from a pool of disordered proteins and on the use Gaussian kernel functions to generate density probability functions, $\widehat{d}_i^A(\delta)$:

$$\widehat{d}_i^A(\delta) = \frac{1}{n_i^A}\sum_{l=1}^{n_i^A} G_K(\delta - \delta_l) \quad (1)$$

In particular, for an atom of type $A$ of an amino acid of type $i$, a specific probability density function $\widehat{d}_i^A(\delta)$ was generated by summing $n_i^A$ Gaussian kernels centered at the CS values of $n_i^A$ experimental observations (Fig. S1). The density function provides an expectation value for the CS that, depending on the amount of statistics, is evaluated as the weighted δ in $\widehat{d}_i^A(\delta)$ or to the δ associated with highest probability in the function (Fig. S1).

In addition to the primary term of prediction, we introduced residue pair-wise correction terms that incorporate the effects of the local sequence in a window of five residues. In particular, these correction terms were derived from pair-wise density functions $\widehat{d}_{i,j}^A(\delta)$ (Fig. S2) corresponding to sub-datasets of our database featuring the specific pair of amino acids i and j of the input sequence.

$$\widehat{d}_{i,j}^A(\delta) = \frac{1}{n_{i,j}^A}\sum_{l=1}^{n_{i,j}^A} G_K(\delta - \delta_l) \quad (2)$$

The expectation values of the pair-wise density functions, $\delta_{i,j}^A$, are therefore used to calculate the corrections to the primary term of chemical shift prediction, $\Delta\delta_{i,j}^A = \delta_{i,j}^A - \delta_i^A$. In the cases of pairs of residues with limited statistics in the database (i.e. $<20$ observations), averaged nearest-neighbor correction terms were calculated $\Delta\delta_{i,j}^A$, corresponding to the overall effect that the amino-acid of type j exerts on all types of neighbor amino-acids in the position of the residue i.

The weights of the pair-wise correction terms were evaluated using the negative overlap between the primary and pair-wise density functions:

$$w = \left(1 - \int \min\left(\widehat{d}_i^A(\delta), \widehat{d}_{i,j}^A(\delta)\right) d\delta\right) \quad (3)$$

Finally, an overall normalization factor—$N_W$—was introduced to balance the contribution of the primary term of CS prediction and the nearest-neighbor corrections. $N_W$ values were calibrated for each atom by maximizing the agreement between the calculated and experimental CS (Fig. S3).

Taken together, these terms define PROSECCO$_{IDP}$, a method using the information of the protein sequence to calculate CS of disordered proteins. In particular, in the case of the atom $A$ of the residue $i$ in a local sequence $k$-$j$-$\underline{i}$-$l$-$m$, the overall equation of PROSECCO$_{IDP}$ is:

$$CS_i^A = \delta_i^A + \frac{1}{N_W}\left(w_{i,k-2}^A \Delta\delta_{i,k-2}^A + w_{i,j-1}^A \Delta\delta_{i,j-1}^A + w_{i,l+1}^A \Delta\delta_{i,l+1}^A + w_{i,m+2}^A \Delta\delta_{i,m+2}^A\right) \quad (4)$$

We compared the performance of the method described in our study with that of the current most accurate predictors

of CS of IDPs by Tamiola et al. (2010) and by Kjaergaard and Poulsen (2011) (Fig. 1). The benchmark indicates that PROSECCO$_{IDP}$ improves the overall prediction of CS in disordered proteins, with particular accuracy in the case of backbone carbon atoms, which are sensitive probes of the local secondary structures in proteins (Camilloni et al. 2012a; Maltsev et al. 2012; Shen and Bax 2012).

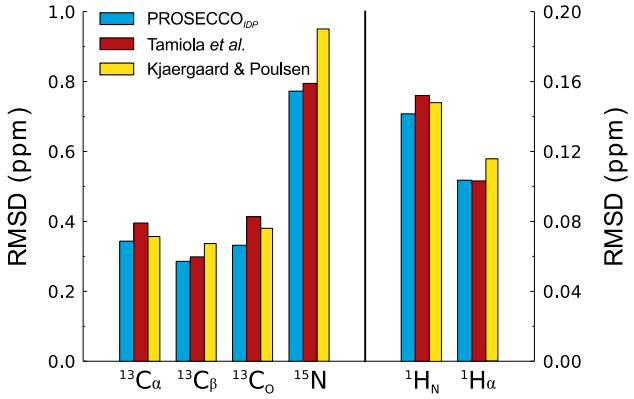


**Fig. 1** Sequence-based prediction of CS in IDPs. Root mean square deviations (RMSDs) are reported between experimental and predicted CS using PROSECCO$_{IDP}$ (cyan) and the methods by Tamiola et al. (2010) (red) and Kjaergaard and Poulsen (2011) (yellow). In the case of PROSECCO$_{IDP}$, the benchmark was performed using a "leave-one-out" approach, whereby when a BMRB entry is used to calculate the RMSD between experimental and calculated CS, the method is reparameterized by excluding this entry from the parameterizing dataset. The leave-one-out benchmark has been rotated on all the BMRB entries employed in PROSECCO$_{IDP}$. In the two other programs tested, the benchmark was performed on the whole dataset of BMRB entries used in PROSECCO$_{IDP}$. A web server for the PROSECCO method is available at http://desimone.bio.ic.ac.uk/prosecco/

## Prediction of chemical shifts in secondary structure elements of natively folded proteins

We then extended the kernel-based approach to the prediction of chemical shifts of secondary structure elements in natively folded proteins. To this end, we analyzed a database of nearly four millions CS of structured proteins to generate density probability functions and pair-wise correction terms for three types of structural elements, namely helixes (combined α-helixes and $3_{10}$-helixes), β-strands (combining both parallel and antiparallel β-sheets) and coils, in analogy with the Q3 classification that is common to many predictors of protein secondary structure. The resulting sequence-based prediction in the Q3 segments of folded proteins was benchmarked against a database containing 77 BMRB entries that were not included in the training dataset ("Materials and methods" and Table S4). The results showed a level of accuracy in the sequence-based CS prediction that is close to two powerful methods exploiting the analysis of protein three-dimensional structures, namely SPARTA+ (Shen and Bax 2010) and Camshift (Kohlhoff et al. 2009) (Fig. 2a). A higher accuracy was found when the benchmark of the sequence-based prediction excluded two residues for each N- and C- end of the Q3 regions, resulting in RMSD values that are similar to those obtained with SPARTA+ (Shen and Bax 2010) and better than those associated with Camshift (Kohlhoff et al. 2009) (Fig. 2b). This finding indicates that the protein sequence is the dominant factor in determining the CS values in internal regions of Q3 segments, but also evidenced that boundary regions between different Q3 segments require further terms of refinement to optimize the sequence-based CS prediction.

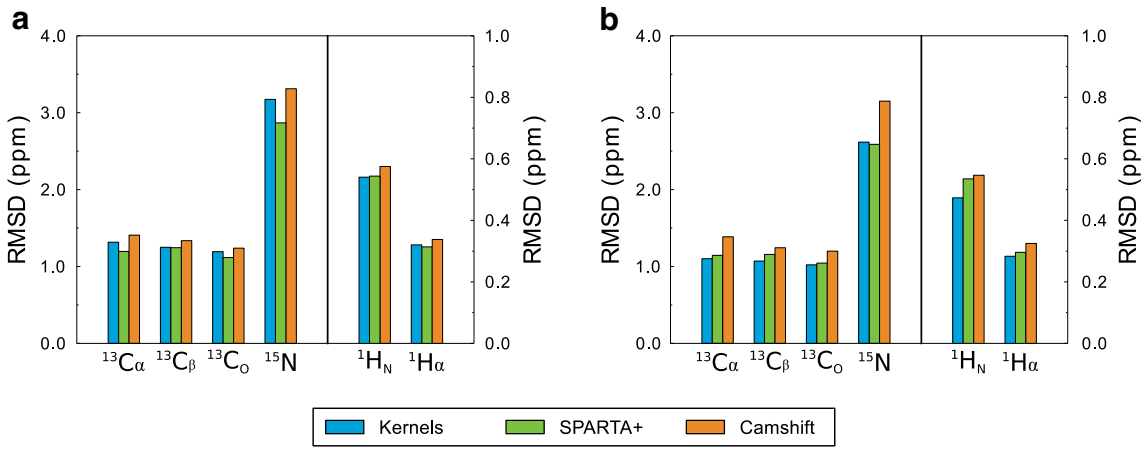


**Fig. 2** Gaussian-kernel prediction of chemical shifts in folded proteins classified in Q3 regions. The benchmark compares the performance of structure-based methods such as SPARTA+ (Shen and Bax 2010) and CamShift (Kohlhoff et al. 2009) with the prediction of CS using the Gaussian-kernels in indexed Q3 regions (helices, strands and coils) of the protein sequence. The dataset for this benchmark included 77 BMRB entries of structured proteins that were deposited from 2016 onwards (see Table S4 for the list BMRB entries and the corresponding PDB codes). **a** Benchmark performed including the whole protein sequences. **b** Benchmark performed by discarding two residues from each termini of the Q3 regions. A dissection of the accuracy in the different Q3 types is reported in Fig. S4

To identify an improved approach for treating the boundary regions, we calculated the distributions of the secondary shifts along the protein sequences, which are defined by the difference between experimental and the random coil CS. This analysis indicated that, in the boundary across two Q3 regions, the secondary shifts gradually morph from those of one segment into the characteristic values of the following (Fig. 3a). Some anomalous transitions, however, were observed such as for example the backbone amide $^{15}$N secondary shifts in the boundary region between helixes and coils (Fig. 3b). Overall, this analysis identified specific patterns of secondary shifts in the boundary regions between Q3 segments, which enabled to generate ad hoc corrections terms resulting in an improved performance of 7.7% in the boundary regions, with an overall improvement of 3.6% (Fig. S5).

## An accurate sequence-based CS predictor for structured proteins

The results of the kernel-based prediction in the Q3 segments indicated that an entirely sequence-based method to predict CS in folded proteins is possible, providing that the Q3 regions of the target proteins are known. Indeed in the above benchmarks, the Q3 regions were indexed from the analysis of the protein structures by DSSP (Kabsch and Sander 1983), however, the significant progresses made in the field of the secondary structure predictions enabled us to define a prediction of CS that is completely independent from experimental protein structures (Fig. 4). In this method, the kernel-based approach to generate CS tables is coupled with an estimation of the Q3 regions along the sequence by means of psipred (Jones 1999). The benchmark of this completely structure-free approach indicates that the uncertainty associated with the secondary structure prediction results in a minimal increase of RMSD values between calculated and experimental CS (only 5.5% compared to the case in which Q3 regions are identified from the experimental protein structures, Fig. S6). Remarkably, despite the such deterioration of performance, the sequence-based method was found to be more accurate than a structure-based predictor such as CamShift (Kohlhoff et al. 2009), although ultimately its performance resulted worse than that of SPARTA+ (Shen and Bax 2010) (Fig S6).

In order to minimize the error introduced when Q3 regions are estimated using psipred (Jones 1999), we introduced an artificial neural network. This network, which employs the information of the local sequences and local secondary structures, is defined with a single hidden layer and a single-node output layer, corresponding to the predicted CS values (see "Materials and methods"). Our benchmark shows that the network was successful in minimizing the difference in accuracy of the sequence-based prediction applied by using Q3 regions estimated with psipred or Q3 regions derived from the experimental protein structures (1.6% difference in RMSD values, Fig. S7).

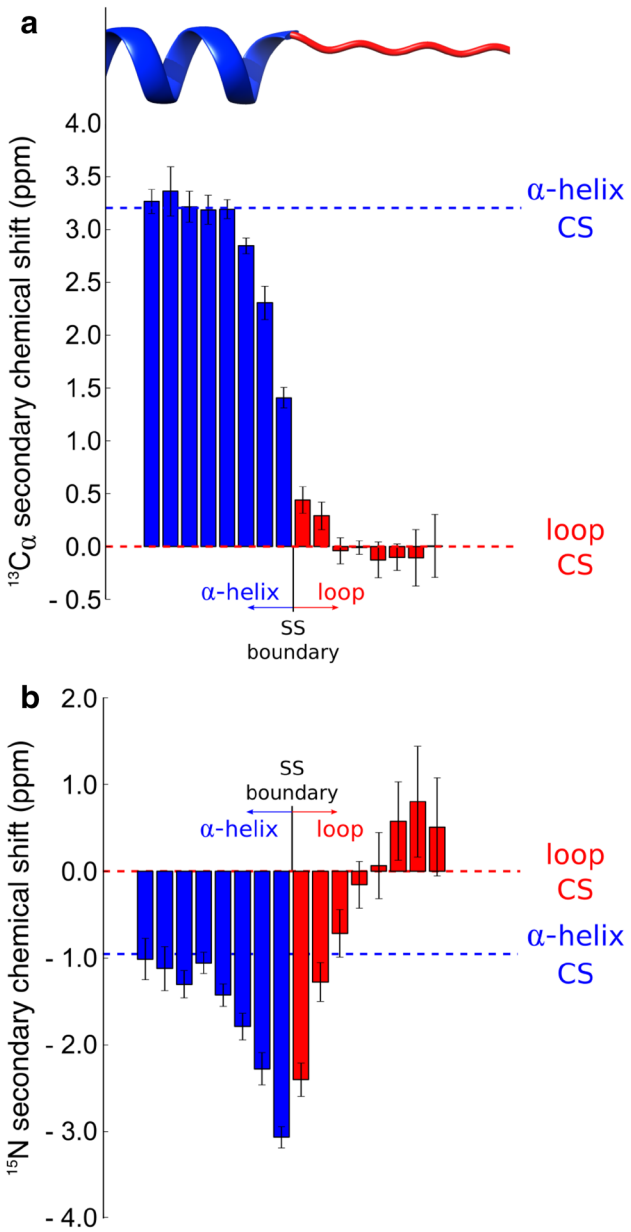


**Fig. 3** Secondary shifts in boundary regions between Q3 segments. The example of the boundary region between α-helixes and loops is shown. Bars report the average secondary shifts as a function of the distance from the boundary between the two Q3 segments, with error bars showing the standard deviations. **a** $^{13}$Cα secondary shifts gradually morph from the typical values adopted in α-helixes (+3.2 ppm) to those of loop regions (0.0 ppm). **b** Backbone amide $^{15}$N secondary shifts, however, exhibits anomalous trends. Starting from the typical values of α-helixes (−0.95 ppm), the secondary shifts augment to a maximum value of −3.0 ppm in correspondence of the last residue of the α-helixes to subsequently inverting toward positive values, with a maximum reached at the position 6 of the loop, and to finally fading to 0.0 ppm

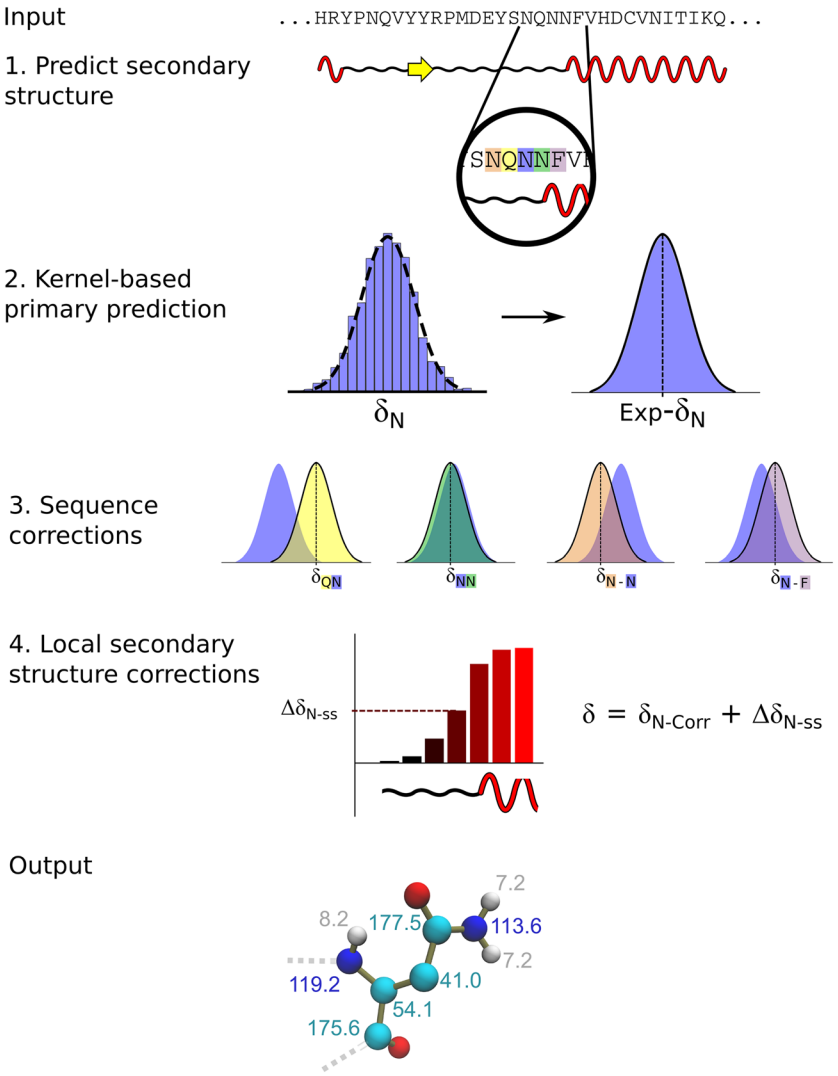


**Fig. 4** PROSECCO$_{FOLDED}$ Scheme of the structure-free prediction of PROSECCO$_{FOLDED}$ As example, a local segment of sequence "NQNNF" is used to illustrate how the prediction of the chemical shifts of the atoms in the central asparagine is generated. In the step 1, the protein sequence is analyzed using psipred (Jones 1999) to predict the secondary structure profile that provides the estimation of the Q3 regions (helixes, strands and coils) of the protein. In steps 2 and 3 the kernel-based prediction is applied to obtain CS tables in each of the Q3 segments, including the corrections of the boundary regions (step 4). The combination of all the prediction terms generates the output CS values (step 5)

Taken together, these analyses define a purely sequence-based method, PROSECCO$_{FOLDED}$, that is able to predict CS in folded proteins with a similar accuracy than methods exploiting structural-similarity criteria, SPARTA+ (Shen and Bax 2010), and higher accuracy than methods using first-principle analyses of the protein structures, Camshift (Kohlhoff et al. 2009) (Fig. 5). The excellent RMSD values for the sequence-based prediction of PROSECCO$_{FOLDED}$ are also reflected in the correlations showed in the scatter plots (Fig. S8).

While the benchmark on the whole database suggests that PROSECCO$_{FOLDED}$ can reach similar accuracy than SPARTA+ (Fig. 5), the case-by-case comparisons indicate that some proteins are better predicted than others (Fig. S9). More specifically, we found that in general PROSECCO$_{FOLDED}$ has better performances with proteins rich in α-helix than β-sheet rich systems, a finding that reflects the performance of PROSECCO in the specific secondary structure elements (Fig. S4).

It is worth noting that the success of the sequence-based prediction by PROSECCO$_{FOLDED}$ depends on the quality of the secondary structure prediction that generates the Q3 indexing. While the accuracy of the current alghoritms such as psipred is generally extremely high, some isolated cases are known where the quality of the secondary structure prediction is low. In these cases, the accuracy of the chemical shift prediction by PROSECCO$_{FOLDED}$ can be severely affected. We illustrate this point using the example of the GA$^{95}$/GB$^{95}$ proteins, two systems designed to have 95% sequence identity but to fold into very different protein topologies (Shen et al. 2009). The secondary structure prediction of these two proteins, which differ for only three residues of their sequences, is extremely challenging for the secondary structure predictors, including psipred. When using PROSECCO$_{FOLDED}$ to predict the chemical shifts of GA$^{95}$/GB$^{95}$, the error associated with psipred, which predicts essentially the same incorrect secondary structure pattern in both proteins (Fig. S10), is propagated to the chemical shift

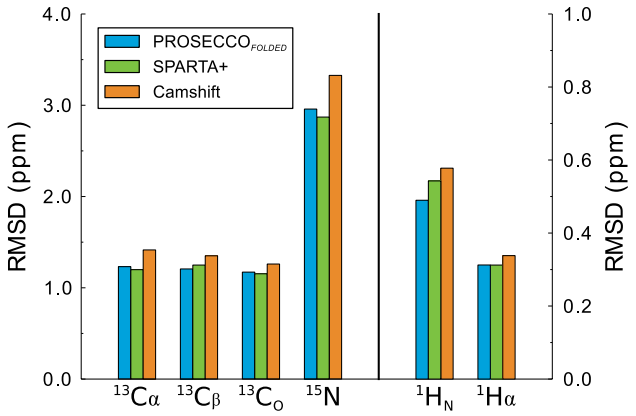


**Fig. 5** Accuracy of PROSECCO$_{FOLDED}$ coupled with an artificial neural network. In its final version, PROSECCO$_{FOLDED}$ was coupled with an artificial neural network to minimize the uncertainty introduced by using predicted secondary structure elements to index the Q3 regions. RMSD values between experimental and calculated CS showed that the sequence-based approach defined in this way reached a similar accuracy of predictors relying on structural similarity criteria such as SPARTA+ (Shen and Bax 2010) and a better performance than methods employing first principle approaches to analyze protein structures such as Camshift (Kohlhoff et al. 2009)

prediction, resulting in worse performance than SPARTA+ (Fig. S10). However, if the correct secondary structure pattern is provided, PROSECCO$_{FOLDED}$ generates highly accurate chemical shifts with a similar precision to SPARTA+.

### Extending the prediction to side chains and different experimental conditions

The large datasets of chemical shifts employed in this investigation enabled a variety of atom types and experimental conditions to be parameterized. In particular, in addition to backbone atoms, PROSECCO generates accurate CS of side chain atoms (Table S5). Among these CS, a particular relevance is given to methyl groups, as these are accurate probes of structure and dynamics of large structured proteins via TROSY experiments (Ollerenshaw et al. 2003). The prediction of chemical shifts of methyl groups by PROSECCO resulted more accurate in the case of IDPs than in the case of folded proteins (Fig. S11a), which for the second is likely due to the lack of information on the tertiary structure, including effects such as the current ring shifts that strongly affects the CS of methyl protons. When compared with structure-based methods such as CH3SHIFT (Sahakyan et al. 2011) (Fig. S11b), the sequence-based prediction of methyl CS in PROSECCO$_{FOLDED}$ resulted more accurate in the case of $^{13}$C atoms (3.5% lower RMSD values) and only 9.5% less accurate in the case of $^{1}$H atoms, which is remarkable considering that the CS of these atoms are significantly influenced by the protein tertiary structure.

Other features of PROSECCO include the ability to distinguish between oxidized and reduced cysteine residues, *cis*- and *trans*-proline residues, and protonated and non-protonated histidine residues. Finally, the distribution of pH values in the CS databases enabled to calibrate two different pH values (Fig. S12) in PROSECCO, namely 6.4 and 2.8.

## Discussion

The approach that we have illustrated has the unique ability to characterize NMR chemical shifts in both structured and disordered proteins using exclusively the information contained in their amino acid sequences. The statistical approach generating PROSECCO has taken advantage of the information contained in large datasets of experimental chemical shifts, which enabled a variety of experimental conditions and system properties to be parameterized, including two types of native states of proteins (folded and intrinsically disordered), two pH values (6.4 and 2.8), main chain and side chain atoms, *cis* and *trans* proline residues and oxidized/reduced cysteine residues. In predicting CS of IDPs, PROSECCO improves the accuracy of the current sequence-based predictors, whereas in the study of folded proteins it achieves similar levels of accuracy of SPARTA+ (Shen and Bax 2010) and an overall higher precision than Camshift (Kohlhoff et al. 2009), but in contrast to these methods without exploiting the information contained protein three-dimensional structures.

In addition to providing a sequence-based tool for treating chemical shifts of both structured and IDPs, the analyses leading to PROSECCO have revealed some key structural dependencies of protein chemical shifts. In particular, they showed that the backbone chemical shifts of residues in internal regions of Q3 segments are mostly independent from the tertiary and quaternary structure of the protein, and rely mainly on the local sequences. Moreover, the boundaries between Q3 segments were shown to adopt secondary shifts that morph between the values of adjacent segments, but in some cases anomalous patterns, such as for example the transition between helix and coil regions, reveal insights to account in structure-based CS predictors as well as methods to generate CS restraints to guide structural refinements in proteins.

## Conclusions

In conclusion, our data show that the sequence-based prediction of chemical shifts in folded proteins can be as accurate as that achieved by structure-based methods, and this feature enables to define a method for treating both structured and disordered proteins. As NMR is assuming a primary

role in characterizing the structure and dynamics of proteins that cannot be studied using other approaches of structural biology, and for which often there is no structure available, an accurate sequence-based prediction of CS will support a large number of NMR investigations. The ability to treat with the same server both structured and disordered proteins is also convenient in the study of proteins that possess both types of characters, and for which current programs can only account either of the folded or of the disordered regions. Our results also suggest that the prediction of chemical shifts in biomacromolecules will be further improved by exploiting the increasingly growing information contained in databases of NMR chemical shifts. The PROSECCO method is available as a web server at http://desimone.bio.ic.ac.uk/prosecco/.

## Materials and methods

### Construction of the database of chemical shifts of proteins

In this investigation we employed an initial dataset of 9514 chemical shift assignments of proteins in the BMRB (Ulrich et al. 2008). This database was subjected to a series of filtering steps to remove sequence redundancies and poorly referenced entries. In particular, using the USEARCH algorithm (Edgar 2010), the initial database was filtered to ensure a maximum sequence identity between protein entries of 90%, which under these criteria resulted in 5140 non-redundant protein entries. The dataset selection was further refined by discarding the entries that showed an incorrect referencing, as assessed by comparing individual chemical shifts with the overall distribution in the BMRB for the specific type of atom and amino acid in the associated secondary structure element. Using this benchmark, entries that had overall CS deviations from the expected value (the center of the distribution) of more than $\delta^{tolerance}$ (tolerance values of 1.5, 0.5 and 3.5 ppm for $^{13}C$, $^{1}H$ and $^{15}N$ respectively) were deemed as incorrectly referenced. The referencing analysis resulted in the removal of 147 BMRB entries. Additionally, specific CS values in otherwise well-referenced BMRB entries were removed if deviating of more than 3.5 standard deviations from their expected values (center of the corresponding distribution), amounting to a total of 0.51% values of the dataset.

### Dataset for disordered proteins

Forty-four entries of our CS database were from IDP or proteins containing an intrinsically disordered region (IDR) of at least 20 residues. IDPs and IDRs were identified by cross-referencing with Disprot (Piovesan et al. 2017) or specific annotations in the literature. The resulting database contained a total of 20,154 chemical shifts from IDPs or IDRs, which were used to parameterize $PROSECCO_{IDP}$. To avoid reductions in the number of experimental data, the whole set of CS was employed in the parameterization, and the *leave-one-out* approach was used to compute the RMSDs between experimental and calculated CS during the benchmarks as previously done (De Simone et al. 2009). A list of BMRB codes for the database employed in the parameterization of $PROSECCO_{IDP}$ is provided in the supplementary materials (Table S1).

### Dataset for folded proteins

The final version of the database of CS contained 4993 BMRB entries of structured proteins, which resulted in 3,953,878 experimental data composing the parameterizing dataset for $PROSECCO_{FOLDED}$. Within this database, however, only 2943 entries were associated with a protein structure in the protein data bank (PDB), which flags the importance of sequence-based CS predictors such as PROSECCO. As a result, in order to assign the regions of helixes, strands and coil (the Q3 segments) for the whole set of 4993 entries, we employed the δ2D (Camilloni et al. 2012b), which is an accurate method to individuate secondary structure elements from protein chemical shifts. Additional entries to the parameterizing dataset were employed exclusively for benchmarking $PROSECCO_{FOLDED}$, SPARTA+ (Shen and Bax 2010) and CamShift (Kohlhoff et al. 2009). In particular, the selected benchmark database contained 77 BMRB entries having all been deposited after 2016, which ensured that these data were not part of the parameterization of SPARTA+ (Shen and Bax 2010) and CamShift (Kohlhoff et al. 2009). The lists of BMRB codes composing the databases employed for the parameterization (Tables S2 and S3) and the benchmarks (Table S4) of $PROSECCO_{FOLDED}$ are provided in the supplementary materials.

### Neural network architecture

The final implementation of $PROSECCO_{FOLDED}$ included a feedforward single-layer regressor neural network model composed of 143 input nodes. This neural network has proved to be a powerful tool in a variety of prediction algorithms across a spectrum of scientific disciplines (Papik et al. 1998; Zupan and Gasteiger 1999), and have also been previously employed in the prediction of chemical shifts based from three-dimensional structures (Li and Brüschweiler 2012; Meiler 2003; Shen and Bax 2010).

In particular, we used a neural structure of a directed mathematical graph, where an initial layer of input data is coded by assigning the data on individual nodes of the graph. Our inputs include local sequence and local secondary structure (Q3 classification of helixes, strands and coils).

The neural network therefore converts the nodes from the input layer into a single output value, the predicted chemical shift, via a series of mathematical operations. The first step is to connect the input layer with a hidden layer of nodes, where hidden nodes store the results of the weighted sum of input nodes. The weights of these sums have been calibrated during the training of the network to model complex relationships between the inputs and the output (Hornik 1991). The calibration revealed an optimal amount of nodes in the hidden layer of 50. In the second step, the results of each hidden node are transformed using an activation function, typically a sigmoid or a hyperbolic tangent function, to enable the algorithm to model nonlinear relationships between the data and the output variable. In our implementation, the hyperbolic tangent function generated more accurate results than sigmoid and rectified linear functions. Finally, a single output node, which contains the predicted chemical shift, receives the weighted sum of the hidden nodes, with weights that are also calibrated during the training of the network. The calibration of all the weights of the network was achieved by minimizing the mean squared error between the outputs and the experimental chemical shifts in the training dataset. The algorithm used for weight optimisation was Adam (Kingma and Ba 2015), a stochastic variation of the gradient descent algorithm, whereby the gradient of the mean squared error is evaluated at each step with respect to the weights, and these are then modified in order to minimize the error.

In the specific implementation of our network, we used 143 input nodes. These nodes include information for each residue in the sequence. In particular, four nodes were used to describe the secondary structure content (one node for each Q3 category and an extra node to account for terminal regions in the segments) in a window of seven residues of the protein sequence, amounting to a total of 28 nodes. The remaining 115 nodes describe the sequence effects on the chemical shifts by including a window of five residues (each residue described by 23 nodes: 20 standard amino acids and additional nodes for oxidized cysteines, *cis* prolines and protonated histidines). Instead of using binary values (0 or 1) on the 23 types of nodes describing the residue identities we employed the score from the BLOSUM-62 amino-acid similarity score matrix (Henikoff and Henikoff 1992), which increases the performance of the network as implemented in neural network employed in SPARTA+ (Shen and Bax 2010) and in other contexts (Berry et al. 2004; Lundegaard et al. 2011).

### Root mean square deviation (RMSD)

As a primary tool to compare the predicted and experimental chemical shifts, we used in our benchmarks the root mean square deviation between n calculated and experimental chemical shift values for a specific spin system:

$$RMSD = \sqrt{\frac{1}{n}\sum_{l=1}^{n}\left(\delta_{l,exp} - \delta_{l,calc}\right)^2} \quad (5)$$

**Acknowledgements** This work was supported by the UK EPSRC (A.D. and M.S) and the Leverhulme Trust (A.D.) and UK BBSRC (A.D.).



## References

Berjanskii MV, Wishart DS (2005) A simple method to predict protein flexibility using secondary chemical shifts. J Am Chem Soc 127:14970–14971

Berjanskii MV, Wishart DS (2013) A simple method to measure protein side-chain mobility using NMR chemical shifts. J Am Chem Soc 135:14536 – 14539

Berjanskii M, Arndt D, Liang Y, Wishart DS (2015) A robust algorithm for optimizing protein structures with NMR chemical shifts. J Biomol NMR 63:255–264

Berry EA, Dalby AR, Yang ZR (2004) Reduced bio basis function neural network for identification of protein phosphorylation sites: comparison with pattern recognition algorithms. Comput Biol Chem 28:75–85

Boulton S, Akimoto M, Selvaratnam R, Bashiri A, Melacini G (2014) A tool set to map allosteric networks through the NMR chemical shift covariance analysis. Sci Rep 4:7306

Bouvignies G, Vallurupalli P, Hansen DF, Correia BE, Lange O, Bah A, Vernon RM, Dahlquist FW, Baker D, Kay LE (2011) Solution structure of a minor and transiently formed state of a T4 lysozyme mutant. Nature 477:111–114

Camilloni C, Schaal D, Schweimer K, Schwarzinger S, De Simone A (2012a) Energy landscape of the prion protein helix 1 probed by metadynamics and NMR. Biophys J 102:158–167

Camilloni C, De Simone A, Vranken WF, Vendruscolo M (2012b) Determination of secondary structure populations in disordered states of proteins using nuclear magnetic resonance chemical shifts. BioChemistry 51:2224–2231

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

Clore GM, Schwieters CD (2003) Docking of protein—protein complexes on the basis of highly ambiguous intermolecular distance restraints derived from 1H/15N chemical shift mapping and backbone 15N–1H residual dipolar couplings using conjoined rigid body/torsion angle dynamic. J Am Chem Soc 125:2902–2912

De Simone A, Cavalli A, Hsu S-TD, Vranken W, Vendruscolo M (2009) Accurate random coil chemical shifts from an analysis of loop regions in native states of proteins. J Am Chem Soc 131:16332–16333

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461

Felli IC, Gonnelli L, Pierattelli R (2014) In-cell 13C NMR spectroscopy for the study of intrinsically disordered proteins. Nat Protoc 9:2005–2016

Hafsa NE, Arndt D, Wishart DS (2015) CSI 3.0: A web server for identifying secondary and super-secondary structure in proteins using NMR chemical shifts. Nucleic Acids Res 43:W370–W377

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

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci 89:10915–10919

Hornik K (1991) Approximation capabilities of multilayer feedforward networks. Neural Netw 4:251–257

Jensen MR, Salmon L, Nodet G, Blackledge M (2010) Defining conformational ensembles of intrinsically disordered and partially folded proteins directly from chemical shifts. J Am Chem Soc 132:1270–1272

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. Biopolymers 22:2577–2637

Kay LE (2005) NMR studies of protein structure and dynamics. J Magn Reson 173:193–207

Kim J, Ahuja LG, Chao F, Xia Y, McClendon CL, Kornev AP, Taylor SS, Veglia G (2017). A dynamic hydrophobic core orchestrates allostery in protein kinases. Sci Adv 3:e1600663

Kingma DP, Ba J (2015). Adam: a method for stochastic optimization. In: Proceedings of the 3rd international conference for learning representations, San Diego, pp 1–15

Kjaergaard M, Poulsen FM (2011) Sequence correction of random coil chemical shifts: correlation between neighbor correction factors and changes in the Ramachandran distribution. J Biomol NMR 50:157–165

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

Krieger JM, Fusco G, Lewitzky M, Simister PC, Marchant J, Camilloni C, Feller SM, De Simone A (2014) Conformational recognition of an intrinsically disordered protein. Biophys J 106:1771–1779

Kuszewski J, Schwieters CD, Garrett DS, Byrd RA, Tjandra N, Clore GM (2004) Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear overhauser enhancement spectra and chemical shift assignments. J Am Chem Soc 126:6258–6273

Li DW, Brüschweiler R (2012) PPM: A side-chain and backbone chemical shift predictor for the assessment of protein conformational ensembles. J Biomol NMR 54:257–265

Li D, Brüschweiler R (2015) PPM_One: a static protein structure based chemical shift predictor. J Biomol NMR 62:403–409

Lundegaard C, Lund O, Nielsen M (2011) Prediction of epitopes using neural network based methods. J Immunol Methods 374:26–34

Maltsev AS, Ying J, Bax A (2012) Impact of N-terminal acetylation of α-synuclein on its random coil and lipid binding properties. BioChemistry 51:5004–5013

Masterson LR, Cheng C, Yu T, Tonelli M, Kornev A, Taylor SS, Veglia G (2010) Dynamics connect substrate recognition to catalysis in protein kinase A. Nat Chem Biol 6:821–828

Meiler J (2003) PROSHIFT: protein chemical shift prediction using artificial neural networks. J Biomol NMR 26:25–37

Mollica L, Baias M, Lewandowski JR, Wylie BJ, Sperling LJ, Rienstra CM, Emsley L, Blackledge M (2012) Atomic-resolution structural dynamics in crystalline proteins from NMR and molecular simulation. J Phys Chem Lett 3:3657–3662

Neal S, Berjanskii M, Zhang H, Wishart DS (2006) Accurate prediction of protein torsion angles using chemical shifts and sequence homology. Magn Reson Chem 44:158–167

Neudecker P, Robustelli P, Cavalli A, Walsh P, Lundström P, Zarrine-Afsar A, Sharpe S, Vendruscolo M, Kay LE (2012) Structure of an intermediate state in protein folding and aggregation. Science 336:362–366

Ollerenshaw JE, Tugarinov V, Kay LE (2003) Methyl TROSY: explanation and experimental verification. Magn Reson Chem 41:843–852

Papik K, Molnar B, Schaefer R, Dombovari Z, Tulassay Z, Feher J (1998) Application of neural networks in medicine—a review. Med Sci Monit 4:538–546

Piovesan D, Tabaro F, Mičetić I, Necci M, Quaglia F, Oldfield CJ, Aspromonte MC, Davey NE, Davidović R, Dosztányi Z et al (2017) DisProt 7.0: A major update of the database of disordered proteins. Nucleic Acids Res 45:D219–D227

Robustelli P, Cavalli A, Vendruscolo M (2008) Determination of protein structures in the solid state from NMR chemical shifts. Structure 16:1764–1769

Robustelli P, Stafford KA, Palmer AG III (2012) Interpreting protein structural dynamics from NMR chemical shifts. J Am Chem Soc 134:6365–6374

Sahakyan AB, Vranken WF, Cavalli A, Vendruscolo M (2011) Structure-based prediction of methyl chemical shifts in proteins. J Biomol NMR 50:331–346

Schwarzinger S, Kroon G.J.A., Foss TR, Chung J, Wright PE, Dyson HJ (2001) Sequence-dependent correction of random coil NMR chemical shifts. J Am Chem Soc 123:2970–2978

Selvaratnam R, Chowdhury S, VanSchouwen B, Melacini G (2011) Mapping allostery through the covariance analysis of NMR chemical shifts. Proc Natl Acad Sci 108:6133–6138

Sgourakis NG, Lange OF, Dimaio F, André I, Fitzkee NC, Rossi P, Montelione GT, Bax A, Baker D (2011) Determination of the structures of symmetric protein oligomers from NMR chemical shifts and residual dipolar couplings. J Am Chem Soc 133:6288–6298

Shen Y, Bax A (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. J Biomol NMR 38:289–302

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

Shen Y, Bax A (2012) Identification of helix capping and beta-turn motifs from NMR chemical shifts. J Biomol NMR 52:211–232

Shen Y, Bax A (2015a) Protein structural information derived from NMR chemical shift with the neural network program TALOS-N. Methods Mol Biol 1260:17–32

Shen Y, Bax A (2015b) Homology modeling of larger proteins guided by chemical shifts. Nat Methods 12:747–750

Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A et al (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci 105:4685–4690

Shen Y, Vernon R, Baker D, Bax A (2009) De novo protein structure generation from incomplete chemical shift assignments. J Biomol NMR 43:63–78

Stollar EJ, Lin H, Davidson AR, Forman-Kay JD (2012) Differential dynamic engagement within 24 SH3 domain: peptide complexes revealed by co-linear chemical shift perturbation analysis. PLoS ONE 7:e51282

Tamiola K, Acar B, Mulder FAA (2010) Sequence-specific random coil chemical shifts of intrinsically disordered proteins. J Am Chem Soc 132:18000–18003

Ulrich EL, Akutsu H, Doreleijers JF, Harano Y, Ioannidis YE, Lin J, Livny M, Mading S, Maziuk D, Miller Z et al (2008) BioMagResBank. Nucleic Acids Res 36:402–408

Wang Y, Jardetzky O (2002) Investigation of the neighboring residue effects on protein chemical shifts. J Am Chem Soc 124:14075–14084

Waudby CA, Camilloni C, Fitzpatrick A.W.P., Cabrita LD, Dobson CM, Vendruscolo M, Christodoulou J (2013) In-cell NMR characterization of the secondary structure populations of a disordered conformation of α-synuclein within *E. coli* cells. PLoS ONE 8:e72286

Wishart DS, Bigam CG, Holm A, Hodges RS, Sykes BD (1995) 1H, 13C and 15N random coil NMR chemical shifts of the common amino acids. I. Investigations of nearest-neighbor effects. J Biomol NMR 5:67–81

Xu X-P, Case DA (2001) Automated prediction of 15 N, 13 C α, 13 C β and 13 C chemical shifts in proteins using a density functional database. J Biomol NMR 29:309–318

Zupan J, Gasteiger J (1999) Neural networks in chemistry. Angew Chem Int Ed 32:503–527