Detection of methylation, acetylation and glycosylation of protein residues by monitoring 13C chemical-shift changes

Pablo G. Garay, Osvaldo A. Martin, Harold A. Scheraga, Jorge A. Vila

Post-translational modifications of proteins expand the diversity of the proteome by several orders of magnitude and have a profound effect on several biological processes. Their detection by experimental methods is not free of limitations such as the amount of sample needed or the use of destructive procedures to obtain the sample. Certainly, new approaches are needed and, therefore, we explore here, as a proof-of-concept, the feasibility of using 13C chemical shifts of different nuclei to detect methylation, acetylation and glycosylation of protein residues by monitoring the deviation of the 13C chemical shifts from the expected (mean) experimental value of the non-modified residue. As a validation test of this approach, we compare our theoretical computations of the 13Ce chemical-shift values against experimental data, obtained from NMR spectroscopy, for methylated and acetylated lysine residues with good agreement within ~1 ppm. Then, further use of this approach to select the most suitable 13C-nucleus, with which to determine other modifications commonly seen, such as methylation of arginine and glycosylation of serine, asparagine and threonine, shows encouraging results.

Detection of methylation, acetylation and glycosylation of protein residues by monitoring ^{13}C chemical-shift changes

Pablo G. Garay,[1] Osvaldo A. Martin,[1] Harold A. Scheraga[2] and Jorge A. Vila[1,§]

[1]*IMASL-CONICET, Universidad Nacional de San Luis, Italia 1556, 5700-San Luis, Argentina;*
[2]*Baker Laboratory of Chemistry, Cornell University, Ithaca, NY, USA.*

ABSTRACT

Post-translational modifications of proteins expand the diversity of the proteome by several orders of magnitude and have a profound effect on several biological processes. Their detection by experimental methods is not free of limitations such as the amount of sample needed or the use of destructive procedures to obtain the sample. Certainly, new approaches are needed and, therefore, we explore here, as a proof-of-concept, the feasibility of using ^{13}C chemical shifts of different nuclei to detect methylation, acetylation and glycosylation of protein residues by monitoring the deviation of the ^{13}C chemical shifts from the expected (mean) experimental value of the non-modified residue. As a validation test of this approach, we compare our theoretical computations of the $^{13}C^{\varepsilon}$ chemical-shift values against experimental data, obtained from NMR spectroscopy, for methylated and acetylated lysine residues with good agreement within ~1 ppm. Then, further use of this approach to select the most suitable ^{13}C-nucleus, with which to determine other modifications commonly seen, such as methylation of arginine and glycosylation of serine, asparagine and threonine, shows encouraging results.

Introduction

Since the pioneer observation of lysine methylation of a bacterial protein by Ambler & Rees (1959), there has been a rising interest in investigating protein post-translational modifications (PTMs) and their role as modulators of protein activity (Zobel-Tropp et al., 1998; Bienkiewicz & Lumb, 1999; Bannister et al., 2002; Paik et al., 2007; Bedford & Clarke, 2009; Kamieniarz & Schneider, 2009; Kamath et al, 2011; Luo, 2012; Theillet et al., 2012a; Theillet et al., 2012a 2012b; Evich et al., 2015; Rahimi & Costello, 2015; Schubert et al., 2015). As a consequence of their relevance, the development of fast and accurate experimental methods for detection of PTMs has also been an object of active research in the field (Kamath et al., 2011). In this regard, Schubert et al. (2015) recently proposed a novel NMR-based methodology to detect glycosylation of residues in proteins under urea-denaturing conditions. The advantages and disadvantages of the proposed new methodology against existing methods for the analysis and detection of PTMs, such as Mass Spectroscopy (MS) [Kamath et al., 2011; Luo, 2012], were discussed in detail by Schubert et al. (2015). One of the main disadvantages of this new methodology (Schubert et al., 2015) is the several larger orders of magnitude of sample required, compared to that needed in MS experiments. Despite this limitation, and the requirement of urea-denaturing conditions, the methodology of Schubert et al. (2015) presents interesting advantages over existing methods, such as analysis of intact proteins to allow the detection of glycosylation in proteins as well as to identify the composition of the attached glycans and the type of glycan linkages. The use of chemical-shift variations to sense PTMs is not a novel approach (Bienkiewicz & Lumb, 1999). In fact, it was used to analyze random-coil chemical-shift variations of the ^{1}N, ^{13}C and ^{15}N nuclei of serine, threonine and tyrosine upon phosphorylation (Bienkiewicz & Lumb, 1999). The results were very promising, e.g., up to ~4 ppm chemical-shift difference was observed

for the $^{13}C^{\beta}$ nucleus of Thr upon phosphorylation. However, to extend this approach to treat other PTMs would require the monitoring of chemical-shift changes in random-coil model peptides of these PTMs, and this is a costly and time consuming procedure. For this reason, we propose such monitoring *in silico,* rather than by using random-coil experiments, as a method to identify the most suitable ^{13}C nuclei with which to sense the existence of PTMs in proteins, e.g., to detect the states of arginine methylation, or glycosylation of serine, asparagine and threonine residues.

We propose to analyze the feasibility to detect PTMs by measuring the deviation of the ^{13}C chemical shifts of a given nucleus from its mean experimental value. To evaluate this idea, *in silico*, we proceed as follows. For a selected nucleus, we compute, at the DFT-level of theory for an ensemble of conformations, the ^{13}C chemical-shifts for the *non*-modified and the modified residues, respectively. Then, from each ^{13}C chemical-shift of the ensemble of modified residues, we subtract the mean of the ^{13}C chemical-shift of the ensemble from the *non*-modified residue. As a result, we obtain the computed chemical-shift difference or what we call, from here on, the Δ value. Naturally, the distribution of the Δ values for *non*-modified residues are, by definition, centered around 0.0 ppm, as can be seen for each of the blue-line curves in Figures 1 to 7 (except Figures 3 and 4).

The use of Δ values to detect PTMs does not require carrying out experiments under urea-denaturing conditions or the use of a large amount of sample, as with the recent NMR-based proposed methodology (Schubert et al., 2015). This is an advantage. As a disadvantage, it does require the labeling of the carbons of the protein, although ^{13}C-labeling is a widely and standard procedure used in both NMR-based experiments *and* theoretical studies because of the exquisite sensitivity of the chemical-shifts to: (*i*) identify flaws in protein structures (Martin et al., 2013); (*ii*) use as constraints during an NMR-based protein structure determination (Vila et al., 2008;

Rosato & Billeter, 2015); (*iii*) resolve local inconsistencies between X-ray crystal structures (Vila et al., 2012); (*iv*) determine the tautomer preference of histidine in proteins accurately (Sudmeier et al., 2003; Vila et al., 2011); (*v*) study sparsely populated, short-lived, protein states that could play a significant role in protein function (Hansen & Kay, 2014); etc.

MATERIAL AND METHODS

Preparation of the model tripeptides for the DFT calculations

DFT calculations were carried out for model tripeptides of the form Ace-Gly-**Yyy**-Gly-Nme, with **Yyy** being lysine (Lys) or arginine (Arg). The torsional angles for the tripeptides were taken from a data-base of a non-redundant set of 6,134 high-quality X-ray structures of proteins solved at resolution ≤1.8 Å, with *R* factor ≤ 0.25, and with less than 30% sequence identity. This ensures that the model tripeptides are a representative sample of the torsional angles observed in nature for the given amino acids. To test that 500 conformations are indeed representative of the conformational accessible-space we performed a preliminary test (data not shown) for 500 and 1000 conformations confirming that using 500 rather than 1,000 conformations leads to the same distributions of shielding values but with a considerable reduction in computational time. All the 500 conformations were free of atomic-overlaps.

For model Lys tripeptides, we generated a total of 5000 conformations, namely 500 for charged Lys (*i.e.* the unmodified amino acid), 1000 for acetylated Lys, 1500 conformations for *mono* methylated Lys, 1500 conformations for *di* methylated Lys *and* finally 500 for *tri* methylated Lys. The following is the reason for the need to compute more than 500 conformations for modified residues, except for tri-methylated Lys. The replacement of hydrogens by methyl or

acetyl groups introduces an asymmetry in the molecule that could influence the DFT computations; hence, rotamers must be generated and the DFT-computed shieldings have to be averaged over these rotameric states.

For Arg, we analyzed a total of 6,000 conformations, namely 500 for charged Arg (*i.e.* the unmodified amino acid), 2,500 conformations for *mono*-methylated Arg (Zobel-Thropp et al., 1998; Bedford & Clarke, 2009), i.e., 1,000 for the *mono*-methylation of each N^{η} of the guanidine nitrogens and 500 for the methylation of the N^{ε} side-chain nitrogen, respectively, and 3,000 for *di*-methylated Arg, i.e., 2,000 for asymmetric and 1,000 for symmetric *di*-methylation of Arg, respectively.

It is worth noting that, because we are interested *only* in the chemical-shift differences (Δ), the implicit assumptions, during the quantum-chemical calculation of the shieldings, are that errors associated with issues not-included in the calculations, such as those derived from a suitable selection of (*i*) dielectric solvent; (*ii*) geometry optimization; (*iii*) reference value, etc., should not affect the accuracy of the calculations of interest because *all* these effects are expected to be canceled-out during the computation of Δ.

Preparation of the glyco-amino acidic residue for the DFT calculations

From *all* possible tripeptides of the above mentioned data-base, i.e., of the *non*-redundant set of 6,134 high-quality X-ray structures of proteins, we randomly selected those containing serine (Ser), threonine (Thr) or asparagine (Asn) as residue **Yyy** in the sequence Ace-Xxx-**Yyy**-Zzz-Nme, with Xxx and Zzz being the nearest-neighbor residues of **Yyy** in the selected tripeptide. This procedure ensures that the model tripeptides are a representative sample of the torsional angles observed in nature for a given amino acid. At this point, it is worth noting that, for Lys and Arg,

the analysis was carried out on selected tripeptides with the sequence Ace-Gly-**Yyy**-Gly-Nme rather than on Ace-Xxx-**Yyy**-Zzz-Nme tripeptide, as for the glycosylated residues. The reason is that methyl and acetyl groups are small chemical groups while glycans are very bulky moieties and, hence, the degree of freedom of the glycosylated residue (Yyy) will be severely restricted depending on the identity of the nearest-neighbor Xxx and Zzz residues. Another peculiarity o the generation of model tripeptides for glycosylated residues is that, after glycosylation of the residue **Yyy,** new side-chain rotations must be explored because of the appearance of additional torsional angles, namely $\chi 2$, $\chi 3$ for Ser and Thr and $\chi 4$ for Asn (see Figure 4, and Figures S6 and S7 of the SI). Among all possible conformations only 500, showing non atomic-overlaps, were considered for the computation of the shieldings at the DFT-level of theory. To assure that the computed shielding differences (Δ) mirror *only* the presence of a glycan linked to an amino acid residue, the monosaccharide of each of the 500 chosen glycosylated conformations was removed and, for the remaining *non*-glycosylated residue, the shieldings were computed at the DFT-level of theory by using the same basis set and functional as for the glycosylated-residue. In this way, we have generated an ensemble of 500 conformations of glycosylated and 500 conformations of *non*-glycosylated residues, namely for Ser, Thr and Asn, that contain no atomic-overlapping and possessing identical backbone and side-chain torsional angles between the glycosylated and the non-glycosylated residue.

Computation of the shieldings, for the nuclei of interest, at the DFT level of theory

To compute the gas-phase ^{13}C-shielding values, at the DFT-level of theory, for any nucleus of interest we will follow the same approach used previously for proteins (Vila et al., 2009) and disaccharides (Garay et al., 2014), namely, the ^{13}C shielding value was computed, by using the Gaussian 09 package (Gaussian 09, 2010) by treating each nucleus, and their neighbors of interest,

at the OB98/6-311+G(2d,p) level of theory, while the remaining nuclei in the sequence were treated at the OB98/3-21G level of theory (Vila et al., 2009; Garay et al., 2014), i.e., by using the *locally-dense basis set* approach (Chesnut & Moore, 1989).

Computation of the standard deviation from the BMRB

On November 13, 2015, we downloaded *all* the chemical shifts deposited at the Biological Magnetic Resonance Bank (BMRB) [Ulrich et al., 2008]. We restricted the analysis to entries that were referenced to DSS, TMS or TSP. Then, we re-referenced the chemical shifts to DSS by adding 0.12 ppm to TSP and −1.7 ppm to TMS. All data points below Q1 − 1.5 x IQR or above Q3 + 1.5 x IQR were considered outliers and removed (with IQR being the Inter-Quartile Range between Q1 and Q3, where Q1 and Q3 are the first and third quartiles, respectively). After removing the outliers, we computed the mean and standard deviation of the distribution for each residue of interest.

Computation of the probability profiles

During the computation of the probability profile, we assume that each chemical shift belongs to one of two possible Gaussian distributions, for example, due to methylated and not methylated Arginine or lysine, respectively. Our aim is to compute a *probability profile* that indicates the probability of a chemical shift to belong to either distribution. For this purpose, we created a simple Bayesian model. In this model, we estimated the mean and standard deviation of the distribution of the chemical-shift differences (Δ), assuming that the Δ values are distributed approximately as Gaussian distributions with unknown mean and standard deviation.

The *prior* for the mean is a Student-*t* distribution with mean equal to the mean of the computed Δ values and a *scale* equal to 0.35. We assumed this *scale* of 0.35 from the lysine

analysis showing that the theoretical and experimental values are in very close agreement within ~1 ppm (see section *Validation Test on lysine derivatives* section below). In other words, we are confident that the theoretical chemical-shift distributions are an accurate representation of the experimental ones within ~1 ppm. Finally, the degrees of freedom of the Student-*t* distribution were estimated from the computed Δ values using as hyper-prior an exponential distribution with mean and standard deviation of 30. A Student-*t* distribution with degree of freedom about 30 or larger is almost indistinguishable from a Gaussian distribution. The *prior* for the standard deviation is a Gamma distribution with mean and standard deviation computed from the experimental values deposited in the BMRB, as explained in the section *Computation of the standard deviation from the BMRB.*

From the model described above, we computed the *posterior* distribution, and from the *posterior* distribution we computed the *posterior* predicted values, i.e. the values of chemical shifts, for each of the two given states, according to the Bayesian model. Given the *posterior* predicted values, it is straightforward to compute the probability of a residue to be in a given state as a function of the Δ values, essentially because we are assuming only two possible states, i.e., methylated and non-methylated. Figure 3 (and Figures S3 to S5 of the SI) shows the results of the analysis in red and blue, semitransparent, lines. Each of these blue lines corresponds to a possible occurrence of the probability profile and the red line, in each of these Figures, is the mean of *all* the blues lines. Thus, the red line represents the expected probability profile and the blue lines the uncertainty in the data according to the Bayesian model.

Data analysis and visualization

Data analysis and visualization were performed using Python (van Rossum, 1995), IPython (Perez & Granger, 2007), NumPy (van der Walt et al., 2011), Pandas (McKinney, 2010),

Matplotlib (Hunter, 2007), and Seaborn (Waskom et al., 2015); Bayesian computations were carried out with PyMC3 (Salvatier et al, 2016).

Results and Discussion

Validation Test on lysine derivatives

As a first step, it is necessary to validate the methodology. For this purpose, we started by analyzing the computed Δ values for the $^{13}C^{\varepsilon}$ chemical-shifts of Lys in a model tripeptide, Ace-Gly-Lys-Gly-Nme, for a total of 6,500 conformations of Lys with various degrees of acetylation or methylation (see details of the generation of the conformations in the Material and Methods section). By following this procedure, the resulting mean Δ values from the Kernel Density Estimation of the chemical-shift differences, shown in Figure 1, are 1.5 ppm, −10.1 ppm, −19.1 ppm and −25.8 ppm for acetylated, *mono*-, *di*- and *tri*-methylated Lys, respectively. A comparison of these computed mean Δ values with the observed $^{13}C^{\varepsilon}$ chemical-shift variations of charged Lys upon acetylation and methylation, namely, 0.0, 9.0, 18.0 and 26.5 ppm, respectively (Theillet et al., 2012a), enables us to conclude that very good agreement exists within ~1ppm between these theoretical predictions and experimental evidence. Overall, the $^{13}C^{\varepsilon}$ chemical-shifts are sensitive enough to detect methylation (see yellow-, violet-, red- and blue-lines in Figure 1) but not acetylation states of Lys; the superposition of the Δ values (see green- and blue-line in Figure 1) make the distinction between acetylated and *non*-modified Lys unfeasible.

In addition, the computed Δ-values for the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ nuclei upon acetylation and methylation of Lys are shown in Figures S1a,b of the Supplemental Information (SI). The superposition of these curves for the methylated with those of the *non*-methylated charged Lys (Figures S1a,b) indicates that these nuclei are not sensitive enough to detect methylation. From

Figure S1a,b (of SI) we also observe that the curves for acetylated Lys do not fully-overlap either the ones for methylated or the *non*-modified charged Lys and, hence, the origin of this behavior must be investigated. In this regard, it should be noted that acetylation, but not methylation, does not preserve the state of charge of Lys. Consequently, the change in protonation upon acetylation should be the reason for the above unexpected result. Indeed, the change of protonation for *non*-modified Lys, as occurs at a high pH value, leads to a Δ distribution (see Figure S1c,d) showing a very similar pattern to the one obtained after acetylation (see Fig. S1a,b). Taking all this together, these results indicate that the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ nuclei of Lys are not sensitive enough to detect either methylation or acetylation states of lysine in proteins.

Having presented the above test on lysine, we have a validation of our theoretical approach and have shown that computation of the Δ values, for a given nucleus, is a useful method with which to detect the methylation states of Lys (Theillet et al., 2012a, 2012b), but not acetylation (Theillet et al., 2012a). Consequently, we decided to extend this analysis to discuss methylation of Arg, and glycosylation of Ser, Asn and Thr, with model tripeptides, and the results are discussed below.

Methylation of Arginine

For Arg, with the sequence Ace-Gly-Arg-Gly-Nme, we have computed the chemical-shifts for the $^{13}C^{\zeta}$ nucleus in 6,000 conformations (see details of the generation of the conformations in *Materials and Methods* section). The Kernel Density Estimation (KDE) of the Δ, are shown in Figure 2a. This Figure shows that it is not possible to distinguish between *mono*-methylated, *i.e.*, between Arg methylated at the N^{ε} or N^{η} group, respectively, or *di*-methylated, *i.e.*, between symmetric or asymmetric *di*-methylated Arg. As a consequence, *all* 5 curves shown in Figure 2a

can be condensed into 3 curves, shown in Figure 2b. Each of the resulting 3 curves shows the Δ values for the *non*-, *mono*-, and *di*-methylated Arg, respectively. A comparison among the resulting distributions enables us to infer that *non*-, *mono*- and *di*-methylated Arg can be distinguished by monitoring the chemical-shift variations of the $^{13}C^{\zeta}$ nucleus. However, as shown in Figure 2b, there is still a small overlap between the distributions and, hence, regions of ambiguity. Because of this overlapping, we compute a probability profile *i.e*., the probability that Arg is in one of two possible states as a function of its Δ value. Computation of the Arg probability-profile is illustrated by a red line and their uncertainty as blue lines (See Figure 3). In general, the blue lines represent different occurrences of the probability-profiles, and the red line the average over all of them (see Figures 3 and Figures S3-S5 of the SI). Thus, chemical-shift differences (Δ) within the range −2 ppm to −3 ppm indicate a large probability (> 80%) that Arg is *mono*-methylated (see red-line of Figure 3a) and a very low probability (~0%) of being *di*-methylated (see Figure 2b). On the other hand, Δ values smaller than −4.8 ppm indicate a large probability (> 80%) that Arg is *di*-methylated (see red-line of Figure 3b) and very low probability of being *mono*-methylated (see Figure 2b). To proceed further with Arg analysis we find that the Δ values for the $^{13}C^{\alpha}$ and $^{13}C^{\beta}$ nuclei upon Arg methylation (shown in Figures S2a,b of SI) are superimposed among themselves indicating, as for Lys, that none of these nuclei is sensitive enough to detect methylation.

At this point, it is worth noting the following. First, there is no significant pK_a change within ~0.5 pK units among *mono*-methylated, *di*-methylated (symmetric or asymmetric) *and non*-methylated Arg (Evich et al., 2015). Therefore, perturbation of the pK_a upon methylation is not large enough to be used as a probe with which to sense Arg modification. Second, there are other nuclei, than carbons, of the Arg side-chain, such as N^{ε}, that show large chemical-shift dispersion upon methylation (Theillet et al., 2012b). However, as noted by Theillet *et al.* (2012b), there is

some limitation in using this nucleus to detect methylation: "*…NMR detection of solvent accessible protein arginine NHε and NHη resonances is only feasible at pH lower than 6.5, because of fast water/guanidinium proton chemical exchange…*" This drawback prevents the use of these nuclei to sense PTMs in proteins around physiological conditions, where most of the biological activities take place and which are conditions desirable for many experiments such as arginine methylase activity measurements. To end, the computation of the probability profiles was carried out taking into account the chemical shifts of *only* two states, such as *mono*- or *di*-methylation, but not other possible modifications, such as phosphorylation or citrullination, which were not considered in our analysis.

Glycosylation of Ser, Thr and Asn

Finally, we explore whether the computed Δ values upon-glycosylation, at the DFT-level of theory (Garay et al., 2014), for the ^{13}C nucleus closer to the glycosylation site, namely $^{13}C^{\beta}$ for the Ser and Thr and $^{13}C^{\gamma}$ for the Asn residue, respectively, can be used as a probe with which to sense the most commonly seen *O*- and *N*-glycosylation, namely the *O*-linked *N*-acetylglucosamine (GlcpNAc) and *N*-acetylgalactosamine (GalpNAc) glycosylation of Ser and Thr (Nishikawa et al., 2010), and the *N*-acetylglucosamine glycosylation of Asn (Chauhan et al., 2013). By focusing our attention on the Δ values upon glycosylation for some selected ^{13}C nuclei of the residue side-chain, we will be able to determine whether, first, the Δ-values can be used to determine glycosylation and, second, the type of glycosylated residue, e.g., GlcpNAc or GalpNAc for Ser and Thr. By focusing our analysis on some nuclei of the amino-acid residue side-chain, rather than on the ^{13}C nuclei of the monosaccharide, to which the residue is linked, would avoid comparing the computed ^{13}C chemical shifts of residue-linked glycans with those from *non*-linked glycans for which, as far we know, there is very sparse information.

O-glycosylation of Ser

We started the glycosylation analysis by computing the ^{13}C chemical-shift for an ensemble of 500 conformations, generated as a function of the torsional angles ϕ, ψ, $\chi1$, $\chi2$ and $\chi3$ (see Figure 4), at the DFT-level of theory, for both the glycosylated and the non-glycosylated Ser. Then, the ^{13}C chemical-shift differences, Δ, for Ser were computed, i.e., between the ^{13}C chemical shift**s?** for the non-glycosylated Ser in the tripeptide Ac-Xxx-Ser-Zzz-Nme, with Xxx and Zzz being the nearest-neighbor residues of Ser in the selected tripeptide, from a non-redundant set of high-quality 6,134 X-ray structures of proteins, and the corresponding ^{13}C chemical shift**s?** for glycosylated Ser, namely for α-D-GalpNAc-(1-O)-Ser and β-D-GlcpNAc-(1-O)-Ser, with Ser in the Ac-Xxx-Ser-Zzz-Nme tripeptide. The identical procedure, to that of Ser, was also carried out for 500 conformations of both the isolated Thr (Ac-Xxx-Thr-Zzz-Nme) and the glycosylated Thr, namely α-D-GalpNAc-(1-O)-Thr and β-D-GlcpNAc-(1-O)-Thr, with Thr in the Ac-Xxx-Thr-Zzz-Nme tripeptide, and 500 conformations of both the isolated Asn (Ac-Xxx-Asn-Zzz-Nme) and the glycosylated Asn, namely for β-D-GlcpNAc-(1-N)-Asn, with Asn in the Ac-Xxx-Asn-Zzz-Nme tripeptide. The resulting curves for the Δ values are shown in Figures 5 and 6 for the $^{13}C^{\beta}$ of Ser and Thr, respectively and Figure 7 for the $^{13}C^{\gamma}$ of Asn.

From Figure 5 we can see, first, large overlapping Δ values for glycosylated Ser, namely between the α-D-GalpNAc-(1-O)-Ser and the β-D-GlcpNAc-(1-O)-Ser (shown as green- and red-lines, respectively, in Figure 5) and, second, a broad distribution of the Δ values for glycosylated Ser with respect to non-glycosylated Ser (blue-line in Figure 5). The large overlapping of Δ values between the glycosylated curves for Ser enables us to represent both kinds of glycosylation as a single curve and, hence, a unique distribution of glycosylation probability (see Figure S3 of SI).

As a result, a Δ value smaller than –3 ppm indicates a large probability (> 80%) that Ser is glycosylated (see red-line in Figure S3 of SI). However, for Δ values above 2 ppm the uncertainty in the probability of glycosylation (represented by the blue-lines in Figure S3 of SI) grows, thus, preventing us from making an accurate assessment as to whether Ser is glycosylated. This is a consequence of the overlapping Δ values between the α-D-GalpNAc-(1-O)-Ser and *non*-glycosylated Ser (see Figure 5).

O-glycosylation of Thr

A similar analysis for the Δ values of the $^{13}C^{\beta}$ of Thr, shown in Figure 6, indicates that *N*-acetylgalactosamine glycosylation of Thr can be detected, mainly because there is no strong overlapping between the glycosylated [α-D-GalpNAc-(1-O)-Thr] and the non-glycosylated (Ace-Xxx-Thr-Zzz-NMe) Δ-distribution for Thr. Indeed, if the computed Δ value is larger than ~ +3 ppm there is > 80% probability that an *N*-acetylgalactosamine glycosylation of Thr exists (see Figure S4 of SI). On the other hand, detection of *N*-acetylglucosamine glycosylation of Thr is not straightforward because of the strong overlapping of the Δ distributions between the α-D-GlcpNAc-(1-O)-Thr and the non-glycosylated Thr (see Figure 6).

The large $^{13}C^{\beta}$ chemical shift difference observed for Thr-106 upon glycosylation (Δ = +9.9 ppm), in the GalNAcα-IFNα2a glycoprotein (Ghasriani et al., 2013), is fully consistent with our prediction for the *N*-acetylgalactosamine glycosylation of Thr. Indeed, a Δ > 6 ppm (see Figure S4 of SI) reveals a high probability for Thr being glycosylated. However, it should be noted that Δ's > 8 ppm are missing from Figure 6, e.g., as for Thr-106 (Δ ~ 10 ppm). At this point, there are two problems associated with the analysis of Thr-106 glycosylation that needs to be clarified, namely the meaning of the computed Δ distributions and whether the observed Δ value for Thr-

106 can be reproduced by our calculations. Let us address each of them separately. First, the Δ distributions in Figure 6, like any other distributions inferred in this work (see Figures 1, 2, 5-7; S1 and S2 of SI) are meant to be representative of the chemical shift population for modified and unmodified residues, respectively, and, hence, Δ values out of range in these figures need to be interpreted as events with low probability, rather than null, occurrence. Second, to test whether the observed Δ value for Thr-106 can be reproduced we decided to (*i*) compute the chemical-shift values for the tripeptide Ac-Val-Thr_{106}-Glu-Nme, with Val and Glu being the nearest-neighbor amino-acid residues in the nonglycosylated and glycosylated GalNAcα-IFNα2a protein sequence; and (*ii*) adopt, for the tripeptide, the torsional angles defined for each of the 20 nonglycosylated conformations (PDB id 1ITF) and the 24 glycosylated conformations (PDB id 2MLS) of the protein. As a result, we obtain for Thr-106 a computed averaged Δ value (~12 ppm) in close agreement, within ~2 ppm, with the observed one (~10 ppm). At this point, is worth noting that the standard deviation (*sd*) of the computed chemical-shift**s** for the nonglycosylated conformations is quite large (~3 ppm); in fact, this is significantly larger than that the *sd* computed for the glycosylated conformations (~1 ppm) and, hence, consistent with the observation that nonglycosylated conformations are more flexible than that the glycosylated one (Ghasriani et al., 2013).

A comparison of the *N*-acetylgalactosamine and *N*-acetylglucosamine glycosylation of Ser and Thr (see red and green lines in Figures 5 and 6, respectively) highlight two very different behaviors, in terms of Δ, albeit Ser and Thr side-chains differ *only* by the attached chemical-group to the C^{β} nucleus, namely an H and a CH_3 group, respectively (see Figure 4 and S6 of SI). Actually, this fact can be understood in light of the differences, in term of the side-chain accessible conformational space, between glycosylated Ser and Thr. Indeed, the 500 conformations of

glycosylated Ser possess the side-chain $\chi 2$ torsional-angle equally clustered among -60^o, $+60^o$ and 180^o, respectively, independent of the nature of the attached glycan. On the contrary, the $\chi 2$ torsional-angles of the 500 conformations of either α-D-GalpNAc-(1-O)-Thr or β-D-GlcpNAc-(1-O)-Thr are mostly clustered around $+60^o$ or 180^o, respectively.

N-glycosylation of Asn

Finally, the Δ values for the $^{13}C^{\gamma}$ of Asn are shown in Figure 7. There is no full overlapping between Δ values computed from glycosylated and non-glycosylated Asn; hence, a chemical-shift difference larger than ~2 ppm indicates a large probability (> 80%) of Asn being glycosylated (see Figure S5 of SI).

Conclusions

Monitoring the Δ's of the $^{13}C^{\beta}$ nucleus of Ser and Thr and the $^{13}C^{\gamma}$ nucleus of Asn, can be used to detect the most commonly seen *O*- and *N*-glycosylations of these residues, except for the type of monosaccharide linked to Ser. Because the chemical-shift is a local property, the proposed detection method should be useful for any state of the protein, i.e., even for intrinsically disordered proteins.

Overall, with a test on lysine derivatives, the strategy proposed here to detect acetylation of Lys, methylation of Lys and Arg, and the *O*- and *N*-glycosylation of Ser, Thr and Asn residues has the potential to be used for recognition of posttranslational modifications within living cells (Doll et al., 2016), e.g., by using the proposed ^{13}C NMR spectroscopic methodology in cells for the study of intrinsically disordered proteins (Felli et al., 2014).

ASSOCIATED CONTENT

The information provided in the Supplemental Information includes: (*i*) the Kernel Density Estimation of the Δ values for Lys and Arg (Figures S1 and S2); (*ii*) the probability profile distribution to detect glycosylation of Ser, Thr and Asn (Figures S3 to S5), and (*iii*) the ball and stick representation of a glycan-amino acidic residue for Thr and Asn (Figures S6 and S7).

ACKNOWLEDGMENTS

This research was supported by grants from the U.S. National Institutes of Health (GM-14312), the U.S. National Science Foundation (MCB10-19767) (HAS), and PIP-112-2011-0100030 from CONICET-Argentina, Project 3-2212 from UNSL-Argentina, and PICT-2014-0556 from ANPCyT-Argentina (JAV).

AUTHOR INFORMATION

Corresponding Authors

§ vila@unsl.edu.ar; jv84@cornell.edu

REFERENCES

Ambler RP, Rees MW. 1959. Epsilon-N-Methyl-lysine in bacterial flagellar protein. *Nature* 184:56-57.

Bannister AJ, Schneider R, Kouzarides T. 2002. Histone methylation: dynamic or static? *Cell* 109(7):801-806. DOI : 10.1016/S0092-8674(02)00798-5.

Bedford MT, Clarke SG. 2009. Protein arginine methylation in mammals: who, what, and why. *Molecular Cell* 33(1):1-13. DOI: 10.1016/j.molcel.2008.12.013.

Bienkiewicz EA, Lumb KJ. 1999. Random-coil chemical shifts of phosphorylated amino acids. *Journal of Biomolecular NMR* 15(3):203-206. DOI: 10.1023/A:1008375029746.

Chauhan JS, Rao A, Raghava GPS. 2013. In silico platform for prediction of N-, O- and C-Glycosites in eukaryotic protein sequences. *PLoS ONE* 8(6):e67008. DOI: 10.1371/journal.pone.0067008.

Chesnut DB, Moore KD. 1989. Locally dense basis-sets for chemical-shift calculations. *Journal of Computationl Chemistry* 10:648-659. DOI: 10.1002/jcc.540100507.

Doll F, Buntz A, Späte A-K, Schart VF, Timper A, Schrimpf W, Hauck CR, Zumbusch A. 2016. Visualization of protein-specific glycosylation inside living cells. *Angewandte Chemie International* 55:2262-2266. DOI: 10.1002/anie.201503183.

Evich M, Stroeva E, Zheng YG, Germann MW. 2015. Effect of methylation on the side-chain pKa value of arginine. *Protein Science*. 25: 479–486. doi:10.1002/pro.2838

Felli IC, Gonnelli L, Pierattelli R. 2014. In-cell ^{13}C NMR spectroscopy for the study of intrinsically disordered proteins. *Nature Protocols* 9:2005-2015. DOI: 10.1038/nprot.2014.124

Garay PG, Martin OA, Scheraga HA, Vila JA. 2014. Factors affecting the computation of the ^{13}C shieldings in disaccharides. *Journal of Computational Chemistry* 35:1854-1864. DOI: 10.1002/jcc.23697

Gaussian 09, Revision C.01. 2010. MJ Frisch, GW Trucks, HB Schlegel, GE Scuseria, MA Robb, JR Cheeseman, G Scalmani, V Barone, B Mennucci, GA Petersson, H Nakatsuji, M Caricato, X Li, HP Hratchian, AF Izmaylov, J Bloino, G Zheng, JL Sonnenberg, M Hada, M Ehara, K Toyota, R Fukuda, J Hasegawa, M Ishida, T Nakajima, Y Honda, O Kitao, H Nakai, T Vreven, JA Montgomery, Jr, JE Peralta, F Ogliaro, M Bearpark, JJ Heyd, E

Brothers, K N Kudin, VN Staroverov, T Keith, R Kobayashi, J Normand, K Raghavachari, A Rendell, JC Burant, SS Iyengar, J Tomasi, M Cossi, N Rega, JM Millam, M Klene, JE Knox, JB Cross, V Bakken, C Adamo, J Jaramillo, R Gomperts, RE Stratmann, O Yazyev, AJ Austin, R Cammi, C Pomelli, JW Ochterski, RL Martin, K Morokuma, VG Zakrzewski, GA Voth, P Salvador, JJ Dannenberg, S Dapprich, AD Daniels, O Farkas, JB Foresman, JV Ortiz, J Cioslowski, and DJ Fox, Gaussian, Inc, Wallingford CT.

Ghasriani H, Belcourt PJF, Sauvé S, Hodgson DJ, Brochu D, Gilbert M, Aubin Y. 2013. A single N-acetylgalactosamine residue at threonine 106 modifies the dynamics and structure of interferon α2a around the glycosylation site. *Journal of Biological Chemistry* 288:247-254. DOI: 10.1074/jbc.M112.413252

Hansen AL, Kay LE. 2014. Measurement of histidine pKa values and tautomer populations in invisible protein states. *Proceedings of the Natural Academy of Sciences USA* 111:1705-1712. DOI: 10.1073/pnas.1400577111

Hunter JD. 2007. Matplotlib: A 2D Graphics Environment. C*omputing in Science & Engineering* 9:90-95 DOI:10.1109/MCSE.2007.55.

Kamath KS, Vasavada MS, Srivastava S. 2011. Proteomic databases and tools to decipher post-translational modifications. *Journal of Proteomic* 75:127-144. DOI: 10.1016/j.jprot.2011.09.014

Kamieniarz K, Schneider R. 2009. Tools to tackle protein acetylation. *Chemistry & Biology* 16:1027-1029. DOI: 10.1016/j.chembiol.2009.10.002.

Luo M. 2012. Current chemical biology approaches to interrogate protein methyltransferases. ACS *Chemical Biology* 7:443-463. DOI: 10.1021/cb200519y.

Martin OA, Arnautova YA, Icazatti AA, Scheraga HA, Vila JA. 2013. Physics-based method to validate and repair flaws in proteins structures. Proceedings of the National Academy of Sciences USA 110:16826-16831. DOI: 10.1073/pnas.1315525110.

McKinney W. 2010. Data Structures for Statistical Computing in Python, Proceedings of the 9th Python in Science Conference, 51-56.

Nishikawa I, Nakajima Y, Ito M, Fukuchi S, Homma K, Nishikawa K. 2010. Computational prediction of O-linked glycosylation sites that preferentially map on intrinsically disordered regions of extracellular proteins. *International Journal of Molecular Sciences* 11:4991-5008. DOI: 10.3390/ijms11124991.

Paik WK, Paik DC, Kim S. 2007. Historical review: the field of protein methylation. *Trends in Biochemical Sciences* 32:146-152. DOI: 10.1016/j.tibs.2007.01.006.

Pérez F, Granger BE. 2007. IPython: A System for Interactive Scientific Computing. *Computing Scientific & Engineering* 9:21-29, DOI:10.1109/MCSE.2007.53

Rahimi N; Costello CE. 2015. Emerging roles of post-translational modifications in signal transduction and angiogénesis. *Proteomics* 15:300-309. DOI: 10.1002/pmic.201400183.

Rosato A, Billeter M. 2015. Automated protein structure determination by NMR. *Journal of Biomolecular NMR* 62:411-412. DOI: 10.1007/s10858-015-9966-z

Salvatier J, Wiecki TV, Fonnesbeck C. 2016. Probabilistic programming in Python using PyMC3. *PeerJ Compututer Science* 2:e55. DOI: 10.7717/peerj-cs.5

Schubert M, Walczak MJ, Aebi M, Wider G. 2015. Posttranslational modifications of intact proteins detected by NMR spectroscopy: application to glycosylation. *Angewandte Chemie International* 54:7096-7100. DOI: 10.1002/anie.201502093.

Sudmeier JL, Bradshaw EM, Haddad EC, Day RM, Thalhauser CJ, Bullock PA, Bachovchin WW. 2003. Identification of histidine tautomers in proteins by 2D $^{1}H/^{13}C^{\delta 2}$ one-bond correlated NMR. *Journal of the American Chemical Society* 125:8430-8431. DOI: 10.1021/ja034072c.

Theillet FX, Liokatis S, Jost JO, Bekei B, Rose HM, Binolfi A, Schwarzer D, Selenko P. 2012a. Site-specific mapping and time-resolved monitoring of lysine methylation by high resolution NMR spectroscopy. *Journal of the American Chemical Society* 134:7616-7619. DOI: 10.1021/ja301895f

Theillet FX, Smet-Nocca C, Liokatis S, Thongwichian R, Kosten J, Yoon M-K, Kriwacki RW, Landrieu I, Lippens G, Selenko P. 2012b. Cell signaling, post-translational protein modifications and NMR Spectroscopy. *Journal of Biomolecular NMR* 54:217-236. DOI: 10.1007/s10858-012-9674-x.

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 H, Markley JL. 2008. BioMagResBank. *Nucleic Acids Research* 36:D402-D408. DOI: *10.1093/nar/gkm957*

van der Walt S, Colbert S, Varoquaux G. 2011. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering* 13:22-30. DOI:10.1109/MCSE.2011.37.

van Rossum G. 1995. Python tutorial, Technical Report CS-R9526, Centrum voor Wiskunde en Informatics (CWI), Amsterdam.

Vila JA, Sue S-C, Fraser JS, Scheraga HA, Dyson HJ. 2012. *Che*Shift-2 resolves a local inconsistency between two X-ray crystal structures. *Journal of Biomolecular NMR* 54:193-198. DOI: 10.1007/s10858-012-9663-0.

Vila JA, Aramini JM, Rossi P, Kuzin A, Su M, Seetharaman J, Xiao R, Tong L, Montelione GT, Scheraga HA. 2008. Quantum Chemical $^{13}C^{\alpha}$ Chemical Shift Calculations for Protein NMR Structure Determination, Refinement, and Validation. *Proceeding of National Academy of Science USA* 105:4389-14394. DOI: 10.1073/pnas.0807105105.

Vila JA, Arnautova YA,Vorobjev Y, Scheraga HA. 2011. Assessing the fractions of tautomeric forms of the imidazole ring of histidine in proteins as a function of pH. *Proceeding of National Academy of Science USA* 108:5602-5607. DOI: 10.1073/pnas.1102373108.

Vila JA; Arnautova YA; Martin OA; Scheraga HA. 2009. Quantum-Mechanics-Derived $^{13}C^{\alpha}$ Chemical Shift Server (CheShift) for Protein Structure Validation. *Proceedings of the National Academy of Sciences USA* 106:16972-16977. DOI: 10.1073/pnas.0908833106.

Waskom M, Botvinnik O, Hobson P, Warmenhoven J, Cole JB, Halchenko Y, Vanderplas J, Hoyer S, Villalba S, Quintero E, Miles A, Augspurger T, Yarkoni T, Evans C, Wehner D, Rocher L, Megies T, Coelho LP, Ziegler E, Hoppe T, Seabold S, Pascual S, Cloud P, Koskinen M, Hausler C, Kjemmett, Milajevs D, Qalieh A, Allan D, Meyer K. 2015. Seaborn: statistical data visualization. DOI: 10.5281/zenodo.19108.

Zobel-Thropp P, Gary JD, Clarke S. 1998. Delta-N-methylarginine is a novel posttranslational modification of arginine residues in yeast proteins. *Journal of Biological Chemistry* 273:29283-29286. DOI: *10.1074/jbc.273.45.29283.*

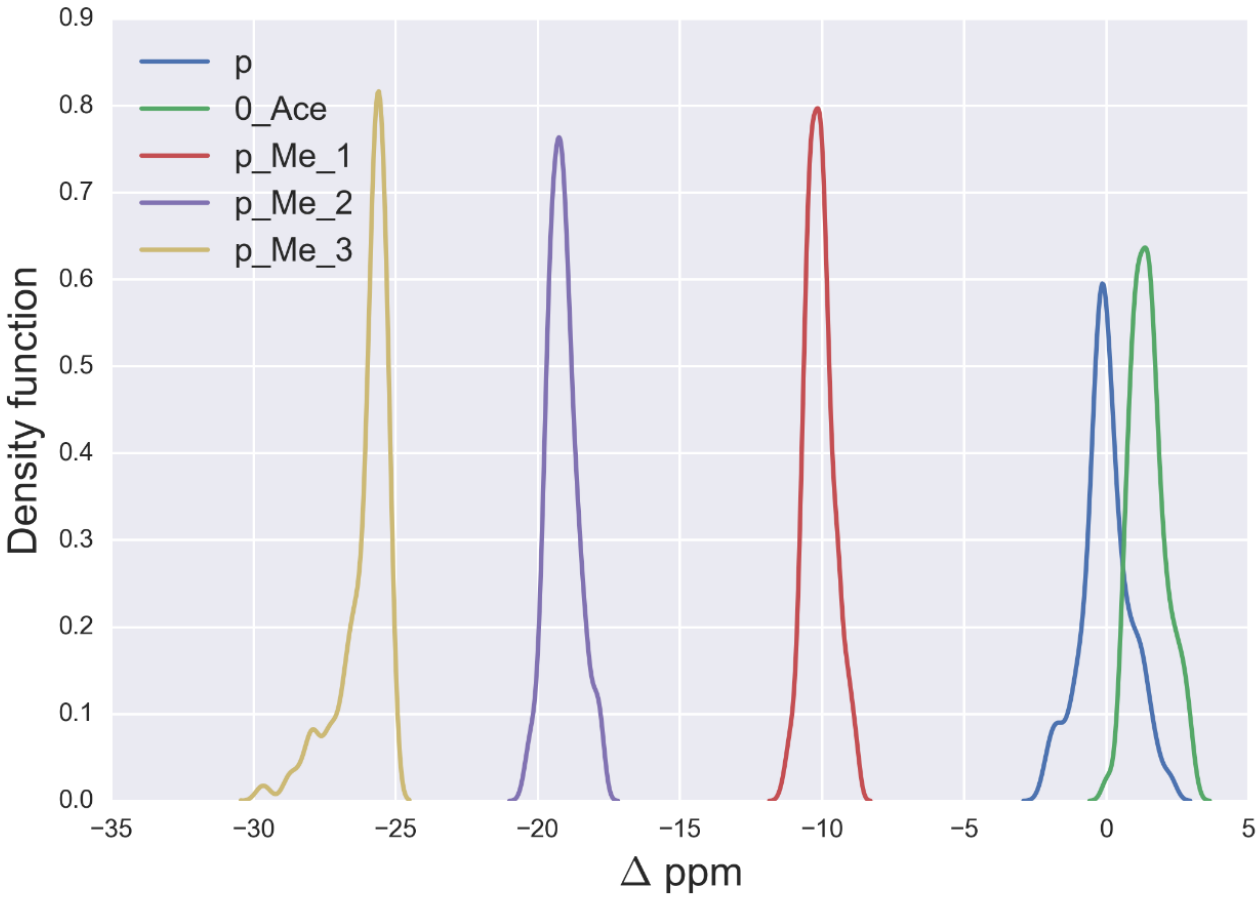
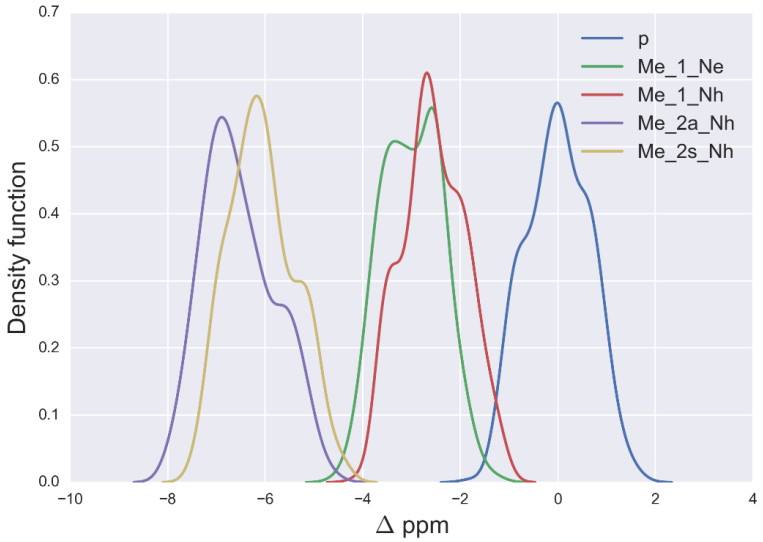
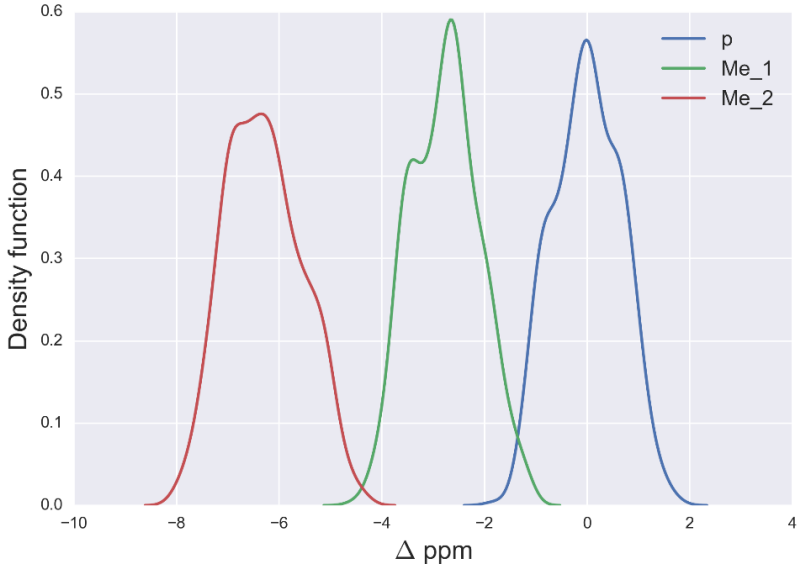


Figure 1.- Kernel Density Estimation of the computed Δ values for the $^{13}C^{\varepsilon}$ nucleus of *non*-modified charged (blue-line), acetylated (green-line), mono- (red-line), di- (violet-line), and tri-methylated (yellow-line) Lys.

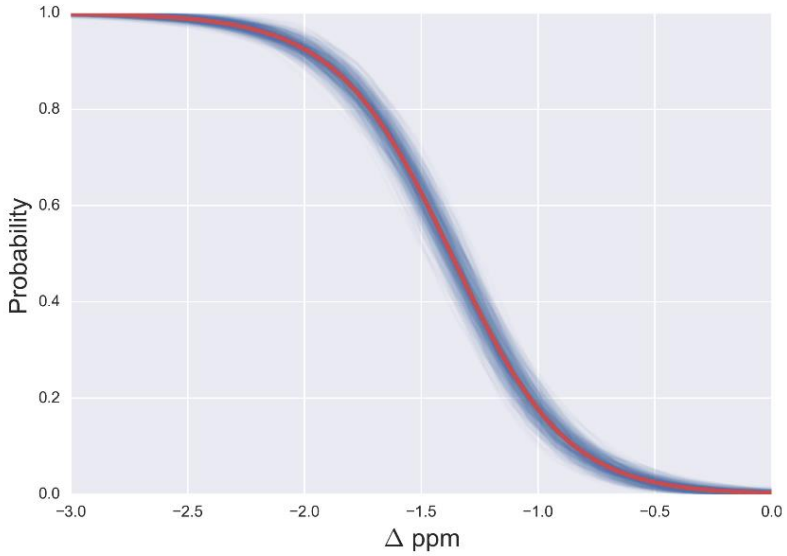


(a)

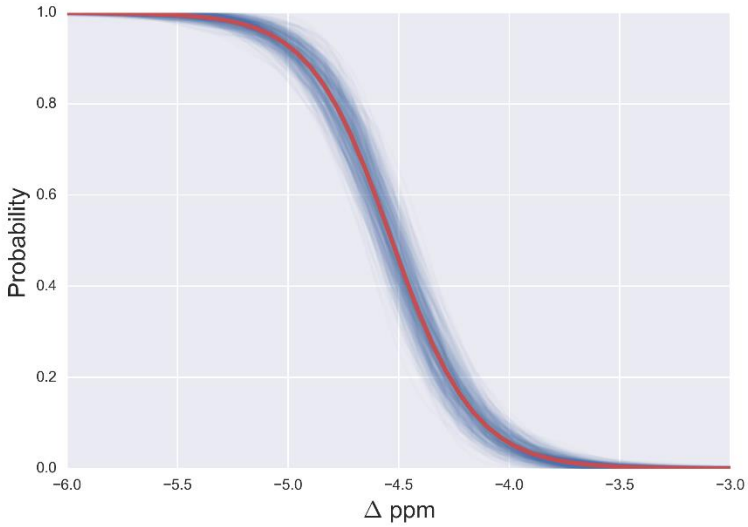


(b)

Figure 2.- (**a**) Kernel Density Estimation of the computed Δ values for the $^{13}C^{\zeta}$ nucleus of *non*-methylated charged (blue-line), mono-methylated [N^{ε} (green-line) and N^{η} (red-line)] and *di*-methylated [symmetric (yellow-line) and asymmetric (violet-line)] Arg; (**b**) all 5 curves shown in (**a**) are condensed in 3 curves.



(a)



(b)

Figure 3. (**a**) Probability profile of the Arg residue to be *mono*-methylated (instead of being non-modified) as function of the Δ values for the $^{13}C^{\zeta}$ nucleus; with data from Figure 2b; (**b**) same as (**a**) for the *di*-methylated Arg. The red line represents the expected probability-profile and the blue lines the uncertainty in the data according to the Bayesian model.

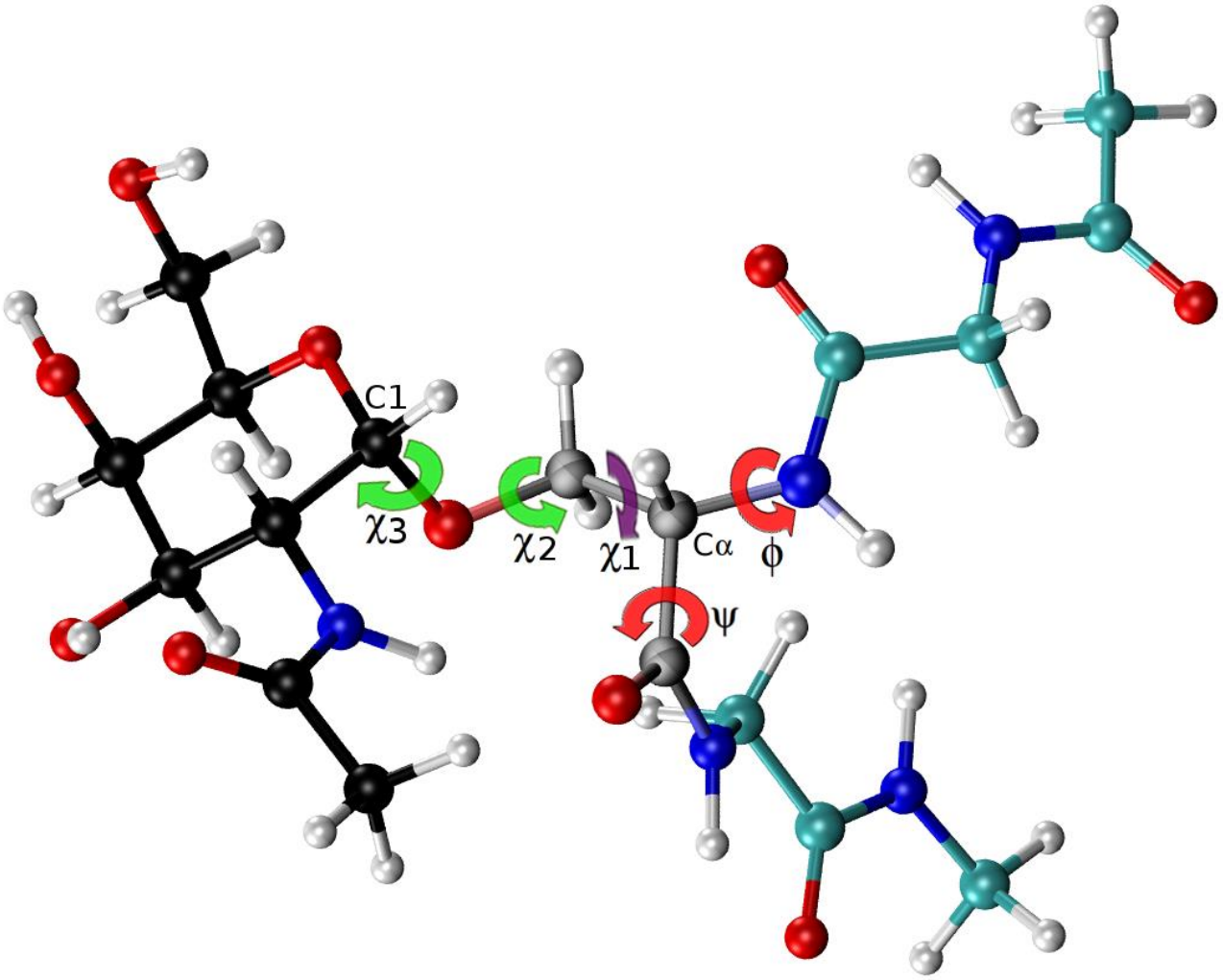


Figure 4.- Ball and stick representation of a glycan-amino acidic residue, namely for α-D-GalpNAc-(1-O)-Ser with "1" representing **C1** of the glycan, and "O" representing the oxygen of the side-chain of Ser in an Ac-Gly-Ser-Gly-Nme tripeptide, in an arbitrary conformation. The χ2 and χ3 torsional angles, of the carbohydrate group (α-D-GalpNAc), are highlighted in green, while the corresponding one for the amino-acidic residue (Ser) are in red, for ϕ,ψ, and purple, for χ1.

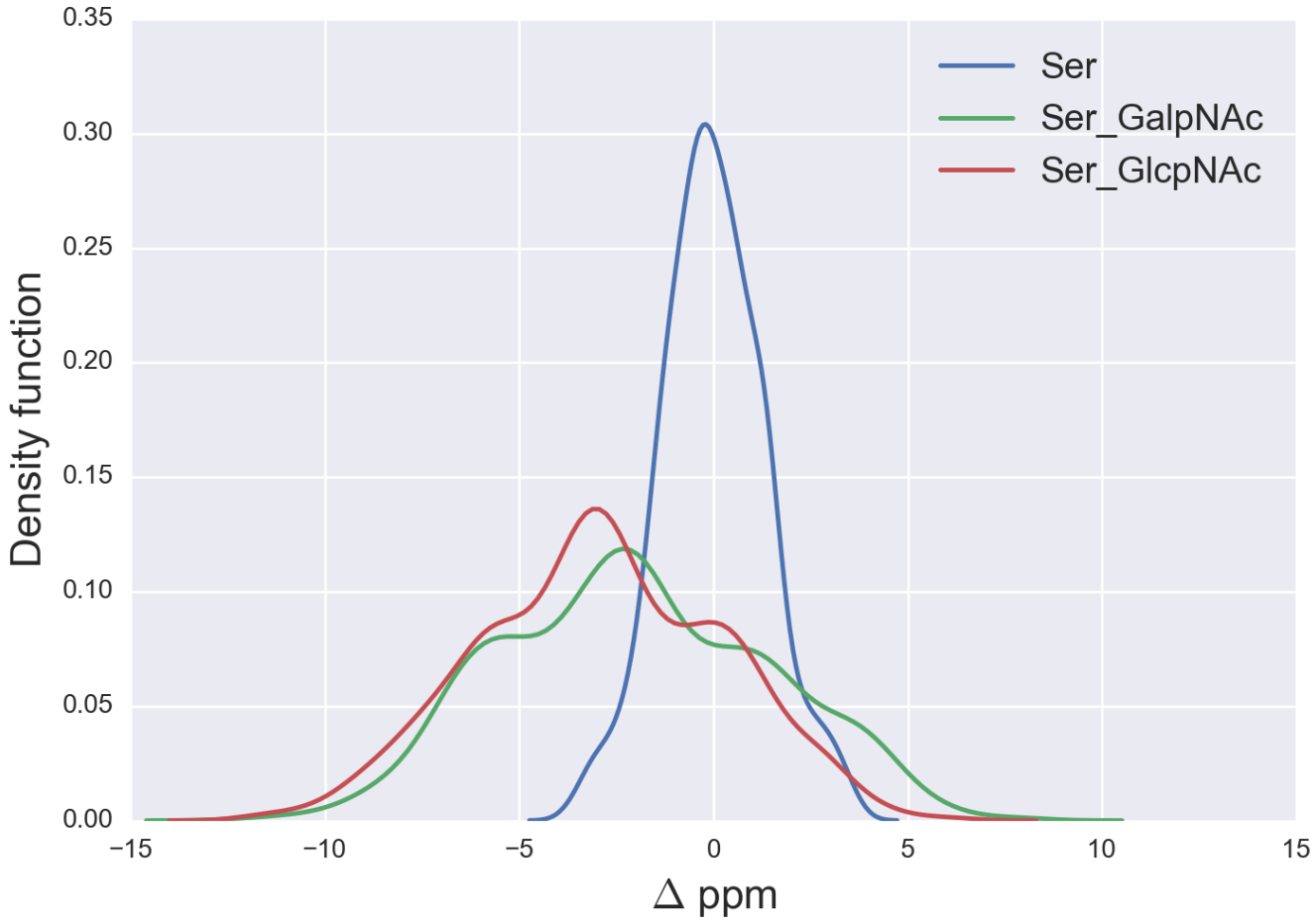


Figure 5.- Kernel Density Estimation of the computed Δ values for the $^{13}C^{\beta}$ nucleus of Ser for: Ace-Xxx-Ser-Zzz-NMe (blue-line), α-D-GalpNAc-(1-O)-Ser (green-line) and α-D-GlcpNAc-(1-O)-Ser (red-line).

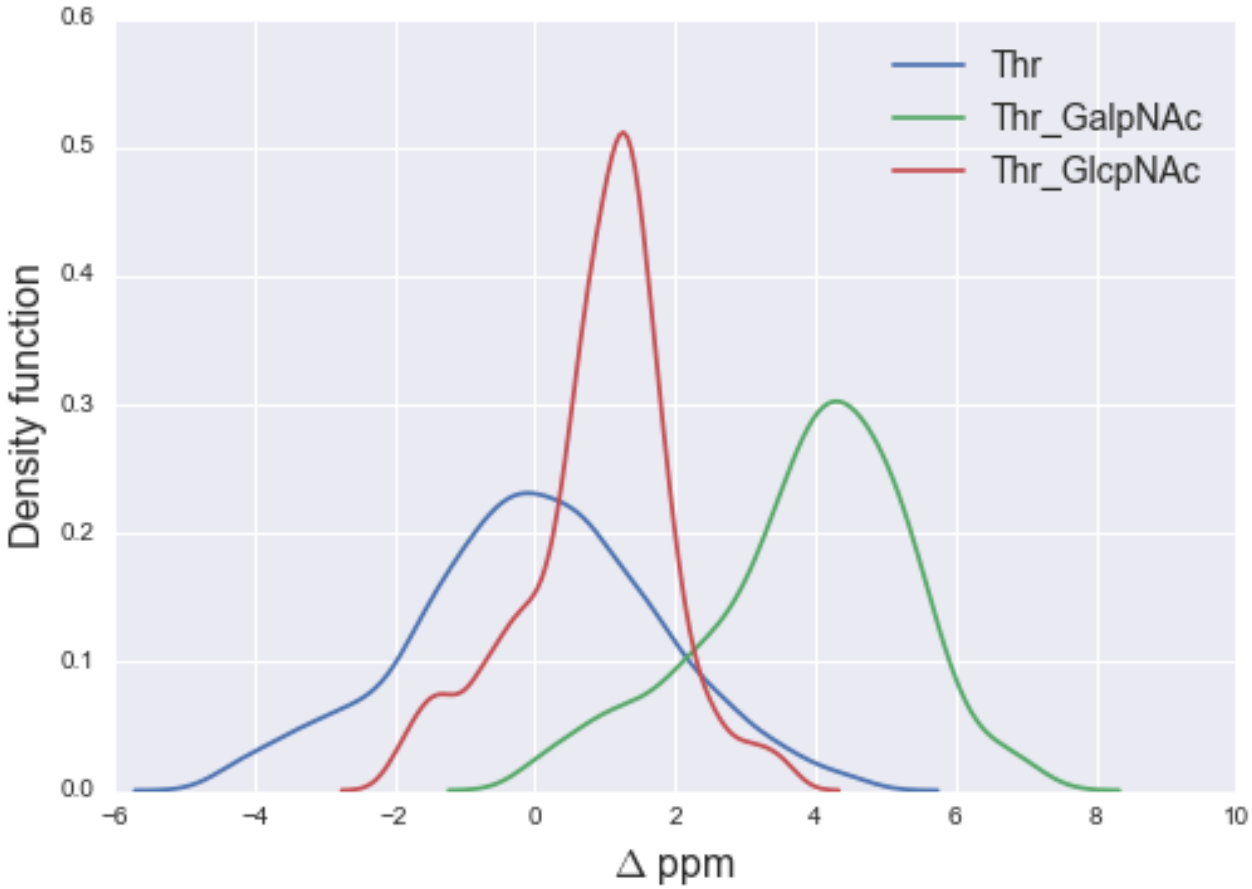


Figure 6.- Kernel Density Estimation of the computed Δ values for the $^{13}C^{\beta}$ nucleus of Thr for: Ace-Xxx-Thr-Zzz-NMe (blue-line), α-D-GalpNAc-(1-O)-Thr (green-line) and α-D-GlcpNAc-(1-O)-Thr (red-line).

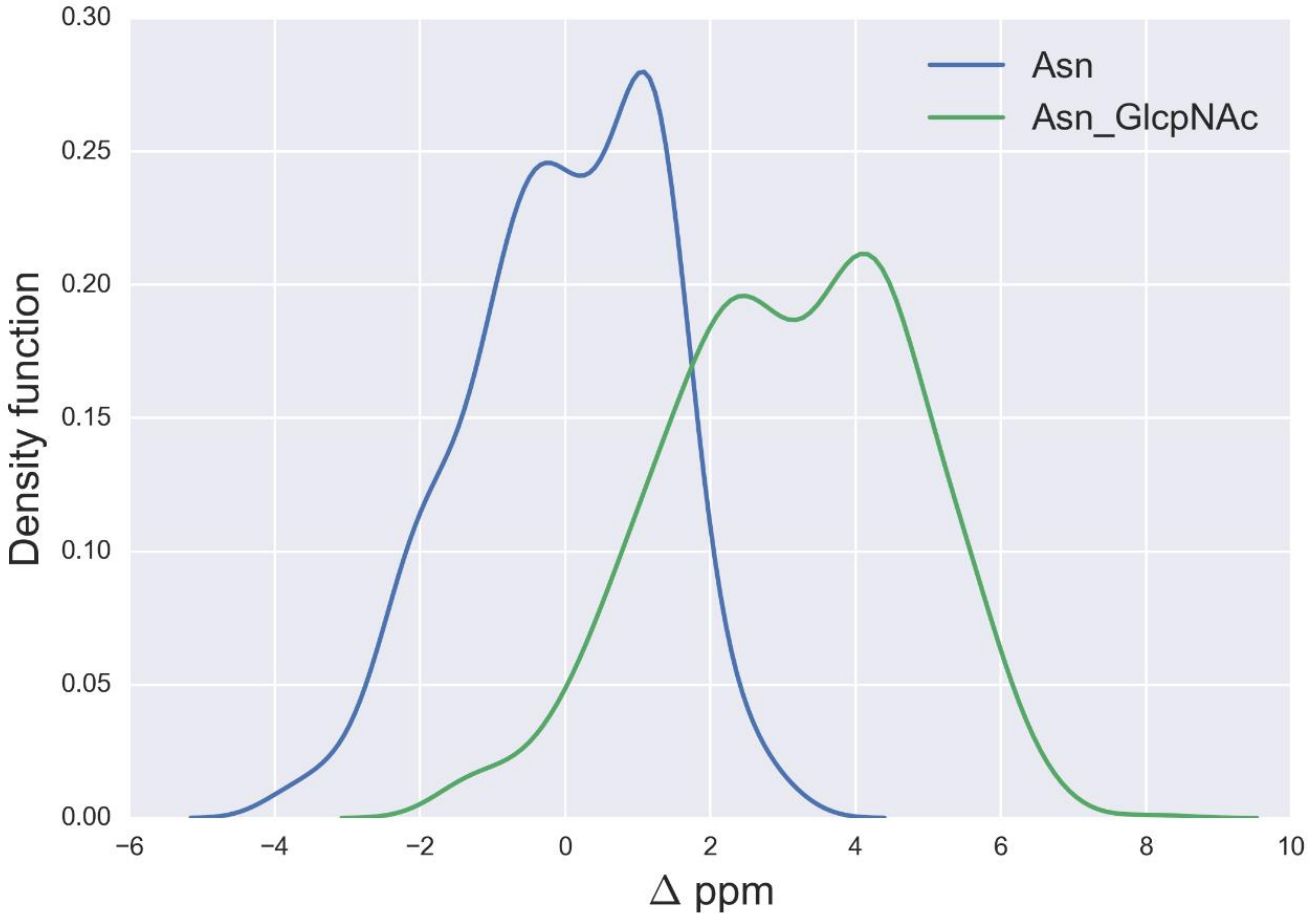


Figure 7.- Kernel Density Estimation of the computed Δ values for the $^{13}C^{\gamma}$ nucleus Asn for: Ace-Xxx-Asn-Zzz-NMe (blue-line) and α-D-GlcpNAc-(1-N)-Asn (green-line)