Chemometrics and Intelligent Laboratory Systems 149 (2015) 132–139

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Conformational analysis from statistical treatment of ^{13}C NMR chemical shifts

Tomás R. Cotos-Yáñez [a], Patricia García-Domínguez [b], Beatriz Iglesias [b], Luis Muñoz [b,*]

[a] Departamento de Estatística e Investigación Operativa, Escola Superior de Enxeñaría Informática, Edificio Politécnico, 32004 Ourense, Spain
[b] Departamento de Química Orgánica, Facultade de Química, Edificio de Ciencias Experimentais, 36310 Vigo, Spain

ARTICLE INFO

Article history:
Received 29 July 2015
Received in revised form 22 October 2015
Accepted 23 October 2015
Available online 2 November 2015

Keywords:
^{13}C NMR chemical shift
Isotropic shielding constant
Conformational analysis
Ridge regression
Internal chemical shift

ABSTRACT

Statistical treatment of experimental ^{13}C NMR chemical shifts of different compounds and calculated isotropic shielding constants of their respective conformers has been carried out in two different ways. The first method was a ridge linear regression between experimental chemical shifts and calculated shielding constants where the calculated coefficients represent the mole fraction of each conformer. Consequently, the sum of all coefficients has to be restricted to 1. The second method was a linear regression between experimental and calculated internal chemical shifts using the same restriction. In general, both methods gave similar results although the second one had a larger standard deviation. The results showed that, in nearly all cases, there is a significant correlation between experimental and calculated data for, at least, one conformer, being this the major one present in the conformational equilibrium. For planar aromatic compounds the conformational equilibrium has been fully characterized when other conformers have been found statistically significant. The statistical analysis on cyclic aliphatic molecules always yielded a conformational composition comparable with the published values. The advantages and drawbacks of the methodology are discussed.

1. Introduction

Nuclear magnetic resonance (NMR) spectroscopy has developed into the technique of choice for the structural elucidation of organic molecules in solution. Comparison between experimental chemical shifts and/or coupling constants and calculated ones [1–3] for a proposed structure has been used as a means to validate it [4,5]. Among all the methods used to calculate chemical shifts and coupling constants, quantum mechanical approaches have become increasingly popular in the last years, as demonstrated by the widespread use of density functional theory (DFT), which allows the calculation of shielding and coupling constants at a reasonable cost. On the other hand, chemical shifts can be easily obtained from shielding constants in two ways [6]. The first one calculates the theoretical chemical shift of a given nucleus (δ_i) by relating its shielding constant (σ_i) to the calculated shielding constant of a reference compound (e.g. tetramethylsilane (TMS), σ_{ref}):

$$\delta_i = \sigma_{ref} - \sigma_i. \quad (1)$$

The second one uses a first degree equation obtained by fitting through a least-squares correlation the calculated shielding constants and the experimental chemical shifts for a set of nuclei.

$$\delta_i = a\sigma_i + b,$$

where a and b are unknown coefficients. From large data sets, regression formulae have been obtained for different types of magnetically active nuclei providing an easy way of estimating chemical shifts from theoretical shielding constants [7–9]. Since most of these equations have been obtained using DFT methods and different basis sets, results are only reliable when the same methods and basis sets are strictly used (for a comparison of methods and basis sets see [10]). In addition, other factors may affect the calculation results and consequently they must be carefully assessed. Thus, small differences in geometries can provide rather large differences in shielding constants. Results for nuclei bonded to heavy atoms have to be discarded if they are not properly taken into account in the calculation. Solvent effects yield smaller differences in calculations, however some authors state they should be considered, at least in an implicit way, by using the Polarizable Continuum Model (PCM) solvent model [11].

The most recent application of comparison between calculated and experimental chemical shifts is the establishment of the relative

---

* Corresponding author.
E-mail address: lmunoz@uvigo.es (L. Muñoz).

http://dx.doi.org/10.1016/j.chemolab.2015.10.009

configuration of two or more stereogenic centers in a molecule [12,13]. Other pieces of research have gone one step ahead and they have taken advantage of the power of the methodology to revise the proposed structure of a few molecules. Milestones on this kind of work are the revision of the structures of hexacyclinol [14] and aquatolide [15]. Some other authors have suggested the use of *ab initio* shielding calculations for protein structure determination and validation [16,17].

Unfortunately, molecules are not rigid structures and, at room temperature, single bonds have enough energy to rotate. Most of the molecules show a flexible behavior, exchanging their 3D structure among two or more conformers. Conformational analysis, a means for understanding the conformational behavior of molecules, relies on molecular modeling with computational methods – molecular mechanics, molecular dynamics, quantum chemical calculations, … – to explore energetically favorable spatial arrangements of a molecule. Once the conformational space has been explored, it is necessary to assess the results with experimentally determined structural data. Thus, nuclear Overhauser effect (NOE) restraints in molecular modeling of biomolecules have been widely used in the determination of their 3D structure in solution [18,19]. But its use is very limited for small molecules, for which NOE correlations are scarce and weak.

Alternatively, comparison can be done between calculated and experimental values for a given property. Obviously, the property needs to be very sensitive to the different spatial arrangements found in calculations. Amid those properties, NMR spectral parameters proved to be very valuable tools for conformational analysis.

The first NMR spectral parameter and probably the most used in the conformational analysis of small molecules was the 3-bond scalar coupling constant (J). This parameter, which is highly dependent on the dihedral angle defined by the two coupled nuclei [20], can be easily calculated either from an empirical Karplus-type equation or through *ab initio* calculations [21,22]. The main drawback in its use comes from the overlapping of signals in the 1D NMR spectra, which precludes very often the determination of the experimental values. Alternatively, the use of 2D sequences can help to obtain them with enough accuracy.

On the other hand, residual dipolar coupling constants (RDC) in oriented media have been successfully employed in the determination of the 3-dimensional structure of biomolecules [23–25]. But, for general use with small molecules in organic solvents, the obtention of an oriented medium has proved very difficult. To overcome this problem, the use of a polymeric matrix has been proposed, although the experimental setup can be challenging and the equilibration of solvent, matrix and solute is generally very slow [26,27]. Recently, the employment of liquid crystals has also been proposed, even though they can only be utilized in water [28].

Although chemical shifts are also very sensitive to the molecular environment and they can be easily obtained from NMR spectra, they have been barely used, since empirical algorithms [29] do not take into account geometry factors when evaluating them. On the other hand, *ab initio* methods provide the shielding constants for each conformer but their values do not seem to be accurate enough. Nevertheless, when isotropic shielding constants are calculated for two conformations of a molecule at the same level of theory, the small difference in values for each nucleus is mainly due to geometric factors. If data sets are treated properly, it should be possible to 'extract' conformational information from them, as demonstrated in the characterization of a tautomeric equilibrium with a statistical treatment of the small variations in ^{13}C NMR chemical shifts from variable temperature experiments [30–32].

In this paper, we present our results on the assessment of the conformational composition of a molecule by statistical analysis of its calculated and experimental ^{13}C NMR chemical shifts.

2. Methods

2.1. Chemical shifts

Experimental ^{13}C NMR chemical shifts and assignations were obtained either from a free access database [33] or from the research literature. Most of the selected spectra were recorded in deuterated chloroform solution. The spectra of the indole derivatives **9** and **10** were recorded in deuterated dimethylsulfoxide (CD_3SOCD_3) and the spectra of the 3-substituted cyclohexanols **13**–**15** were recorded in a 9:1 mixture of CS_2/CD_2Cl_2. Solution concentrations were very variable, ranging from a few tens to several hundred mg mL^{-1}. All spectra were assumed to be recorded at room temperature since it was not otherwise stated. Spectra were recorded at variable fields, ranging from 15 to 125 MHz.

Isotropic shielding constants were calculated through molecular modeling. Crude initial geometries were obtained through molecular mechanics methods implemented in package programs such as Chem3D [34] and Gabedit [35]. All DFT calculations were performed with the Gaussian 09 suite of programs [36]. The geometry optimization, and the values of energies, force constants and isotropic shielding constants were calculated at the B3LYP/6-311 + G(d,p) level of theory in chloroform using the PCM model with standard parameters. All the optimized geometries were characterized as true minima by checking the non existence of imaginary frequencies. Once the frequencies were obtained, a thermochemical analysis was carried out and the free energy value was obtained by modeling both the translational, rotational and vibrational entropies. Isotropic shielding constants were calculated in chloroform for each conformer using the Gauge-independent atomic orbital (GIAO) methodology.

2.2. Statistical analysis

In our analysis, it was supposed that the relationship between calculated isotropic shielding constants and experimental ^{13}C NMR chemical shifts is linear, given by

$$\delta_{\exp_i} = \beta_0 + \beta_1\sigma_{i1} + \ldots + \beta_p\sigma_{ip} + \epsilon_i, \quad 1 \le i \le n, \tag{2}$$

where $\delta_{\exp_i}$ is the experimental chemical shift, σ_{ij} is the j-th shielding constant, $(\beta_0, \beta_1, \ldots, \beta_p)^t$ is the unknown coefficients vector and ϵ_i is the random error with mean 0.

The basic requirement for the least squares standard estimation of Eq. (2) is that $(\mathbf{X}^t\mathbf{X})^{-1}$ exists, with $\mathbf{X}$ being the $n \times (p+1)$ design matrix given by row vectors $\mathbf{x_i}^t = (1, \sigma_{i1}, \ldots, \sigma_{ip})$ and $i = 1, \ldots, n$. However, highly correlated variables (multicollinearity) give an inaccurate inverse matrix which is an important problem in statistics and computer science (see 'badly conditioned linear regression problems' in [37]). The technique of ridge regression (RR) is one of the most popular and simple alternatives to ordinary least squares (OLS) [38]. To guarantee the inverse of $(\mathbf{X}^t\mathbf{X})$ in OLS, a diagonal matrix λI_{p+1}, where I_{p+1} is a $(p+1) \times (p+1)$ identity matrix, and λ is an unknown parameter, is added. The coefficients estimator is given by:

$$\hat{\beta}_{rr} = \left(\mathbf{X}^t\mathbf{X} + \lambda I_{p+1}\right)^{-1}\mathbf{X}^t\mathbf{y}, \tag{3}$$

with $\mathbf{y} = (\delta_{\exp_1}, \ldots, \delta_{\exp_n})^t$. If $\lambda = 0$, an unbiased OLS estimator is obtained. On the other hand, increasing the λ parameter makes the RR estimator more biased, but with a lower variance. So, introducing a little bias might lead to a substantial decrease in the prediction error.

Eq. (3) can be obtained as the solution of the following minimization with constraint:

$$\text{minimize } (\mathbf{y}-\mathbf{Z}\beta)^t(\mathbf{y}-\mathbf{Z}\beta) \; s.t. \; \sum_{i=1}^{p}\beta_i^2 \le t,$$