CheShift-2: graphic validation of protein structures

Osvaldo A. Martin1, Jorge A. Vila1,2,∗ and Harold A. Scheraga2,∗

1Universidad Nacional de San Luis, IMASL-CONICET, Ejército de Los Andes, 950-5700 San Luis, Argentina and 2Baker Laboratory of Chemistry and Chemical Biology, Cornell University, Ithaca, NY 14853-1301, USA

Associate Editor: Anna Tramontano

ABSTRACT
Summary: The differences between observed and predicted 13Cα chemical shifts can be used as a sensitive probe with which to detect possible local flaws in protein structures. For this reason, we previously introduced CheShift, a Web server for protein structure validation. Now, we present CheShift-2 in which a graphical user interface is implemented to render such local flaws easily visible. A series of applications to 15 ensembles of conformations illustrate the ability of CheShift-2 to locate the main structural flaws rapidly and accurately on a per-residue basis. Since accuracy plays a central role in CheShift predictions, the treatment of histidine (His) is investigated here by exploring which form of His should be used in CheShift-2.

Availability: CheShift-2 is free of charge for academic use and can be accessed from www.cheshift.com

Supplementary information: Supplementary data are available at the Bioinformatics online.

Received on January 4, 2012; revised on March 27, 2012; accepted on March 31, 2012

1 INTRODUCTION

Chemical shifts provide important information about the conformations of proteins in solution (see, for example, Wishart, 2011, and references therein). For this reason, we developed CheShift-2, a Web server for protein structure validation based on a quantum mechanics database of 13Cα chemical shifts (Vila et al., 2009). CheShift was originally developed to return a list of predicted values of 13Cα chemical shifts. It was the user’s responsibility to compare the predicted with the observed 13Cα chemical shifts to assess the global quality of a protein. However, it is a highly desirable goal of any accurate validation method that receives a new PDB ID.

2 METHODS

For each amino acid μ, it is possible to define the difference between observed and predicted 13Cα chemical shifts as:

$$\Delta_{\mu} = \frac{13C_{\mu,\text{observed}}}{\mu} - \frac{13C_{\mu,\text{predicted}}}{\mu} Ω$$

where, 13Cμ, is the chemical shift of residue μ in conformation Ω out of Ω conformations. The average of the predicted chemical shifts over the Ω conformations is evaluated because proteins in solution exist as an ensemble of conformations.

The following procedure for mapping the ∆μ values onto a 3D protein model was formulated. First, the ∆μ value computed for each residue μ is smoothed by averaging it over the values of the two nearest-neighbor residues (see Supplementary Material for details). Second, the resulting averaged ∆μ value is discretized according to the following rule:

$$\Delta_{\mu}^{\text{range}} = \begin{cases} 1, & (\Delta_{\mu}^{\text{mean}}) < 1.70 \text{ppm} \leq 2 \text{ppm} \leq 3.40 \text{ppm} \leq 5 \text{ppm} \leq 7 \text{ppm} \leq 10 \text{ppm} \\ 0, & (\Delta_{\mu}^{\text{mean}}) < 1.70 \text{ppm} \leq 2 \text{ppm} \leq 3.40 \text{ppm} \leq 5 \text{ppm} \leq 7 \text{ppm} \leq 10 \text{ppm} \end{cases}$$

The selection of the cut-off σ value of 1.7 ppm is explained in the Supplementary Material. Third, the ∆μ > 10 value is discretized according to the following rule:

The authors have collected new data or had re-refined the structure. The obsolete entry is usually replaced by a new (superseding) entry that receives a new PDB ID.

3 RESULTS

We found evidence (see Supplementary Material) indicating that the protonated form of histidine (His) characterizes the neutral ones, namely, the Nδ1-H or Nδ2-H tautomers form, respectively, leading to a better representation of the observed 13Cα chemical shifts. This observation, together with the well-documented effect of proline on chemical shifts that are within ±18 ppm (blue) are considered small; within ±2 (red) they are considered medium, i.e. being both blue and white considered as acceptable differences; and beyond ±2 (red), they are considered large differences and, hence, special attention should be attached to those residues. In addition, the color yellow was adopted to indicate the absence of the observed or computed 13Cα chemical shift value.

To whom correspondence should be addressed.
The reason for this drawback of the method is due to the fact that the most important difference between 1BKQ and 1F3C is the sheet arrangement of Figure 1B (I) is better than that of Figure 1B from 1BKQ and 1F3C could mislead the user to conclude that the β in 1F3C, strand 3B is part of the conformation, strand 3B is part of a with respect to the rest of the protein. Specifically, in the 1BKQ chain protein (Fig. 1B). This protein was first, incorrectly, solved as a dimer by X-ray crystallography, at 2.5 Å resolution, is enriched in blue/white regions, confirming that the 1CMI protein is indeed a very good structure [see Fig. 1B (III)].

4 CONCLUSIONS
CheShift-2 constitutes a fast and accurate validation tool with which to determine the existence of local flaws in protein models. Examples analyzed in the present study show that, if the NMR-determined ensemble had not been solved at a high-quality level, a comparison with the corresponding structure determined by X-ray crystallography reveals that the X-ray structure is almost flawless and, hence, indicates that the detected flaws in the NMR-determined ensemble are not a bias of the method but a warning that the NMR-derived structure may benefit from further structural refinement.

This new physics-based validation tool, CheShift-2, should be used as a complementary one to other existing knowledge-based methods, such as WHAT IF (Vriend, 1990) and PROCHECK (Laskowski et al., 1993), or combined knowledge-based and physics-based methods, such as the PSVS package (Huang et al., 2005; Bhattacharya et al., 2007).

ACKNOWLEDGEMENTS
We thank P. Serrano at the Scripps Research Institute for valuable discussions regarding protein structure validation methods. The research was conducted by using the resources of Pople, a facility of the NSF Terascale Computing System at the Pittsburgh Supercomputer Center.

Funding: NIH (GM-14312) and NSF (MCB10-19767), USA; CONICET and UNSL (P-328402), Argentina.

Conflict of Interest: none declared.

REFERENCES