Bayesian Model-comparison Of Biomolecular Structures

Agustina Arroyuelo[1*] and Osvaldo A Martin[1]

1 Instituto de Matemática Aplicada San Luis - CONICET, Italia 1556, 5700-San Luis, Argentina.

* aarroyuelo@unsl.edu.ar

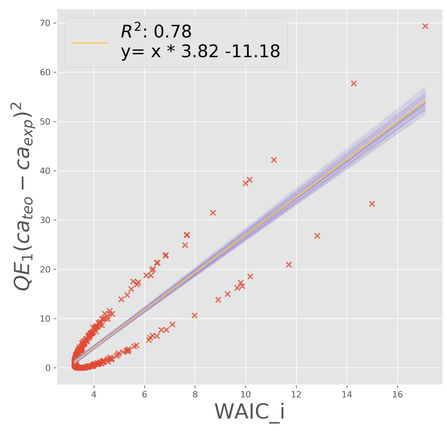
Abstract

Information criteria are often used for model comparison and averaging. Of the many information criteria we focus particulary on WAIC (Widely Applicable Information Criterion). WAIC presents the advantage of being *pointwise*, this is useful, because some observations are harder to predict than others and may also have different uncertainty [1]. Also WAIC is fully Bayesian in the sense that it's computation requires the whole *posterior*, and not just a single value, like the *Maximum a Posteriori* and its cheap to obtain, once we have computed the *posterior*. We propose to develop a metric based on WAIC for assessing the quality of biomolecular structures through Bayesian models. This metric, should be easy to interpret and take into account peculiarities of biomolecular structures like the different types of available experimental data. Additionally, WAIC is an approximation to the *out-of-sample error* and thus is conceptually similar to metrics like the *R-value* and *R-free-value* widely used in macromolecular crystallography. In this study we will evalute if $WAIC$ is an objective measure to assess the quality of Biomolecular structural models, specially those determined by Nuclear Magnetic Resonance (NMR).

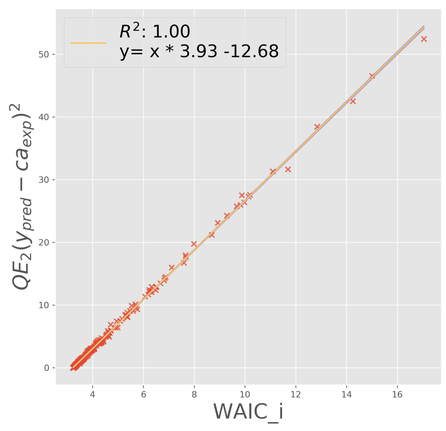
Methods

Our dataset consist in a pool of 111 high quality protein structures obtained from the Protein Data Bank [2]. Each protein in this set has a resolution < 2.0 Åand R-factors <= 0.25. The structures do not containing DNA and/or RNA molecules. Additionally, every protein in our data set has an entry at the Biological Magnetic Resonance Bank from which we obtained experimental chemical shift data for C_α [3]. Theoretical chemical shift data was obtained through computation with *CheShift2* [4]. A linear regression model for the chemical shifts was performed, both hierarchical and non-hierarchical. $WAIC$ by residue ($WAIC_i$) and Quadratic Error (QE) was computed. Once the bayesian model for the chemical shifts was defined, two different ways to compute the quadratic error were applied: $QE_1 = (theoretical - experimental)^2$ and $QE_2 = (teoretical_{predicted} - experimental)^2$. The second expression evaluates the data predicted by the bayesian model. Subsequently, another linear model to adjust $WAIC_i$ and QE was defined. The agreement between these observables and the impact of the hierarchical model were analysed. All bayesian models were computed using PyMC3 [5]. Plots for non-hierarchical models are shown on panels *a* and *b* and for hierarchical models on panels *c* and *d*.

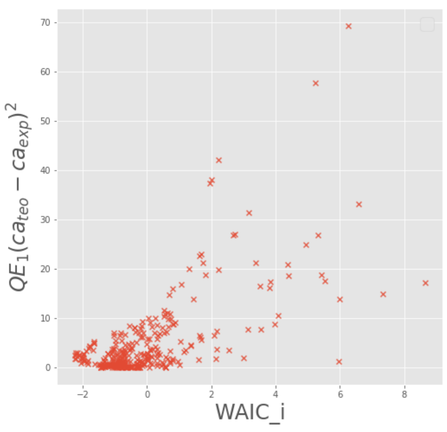
Results



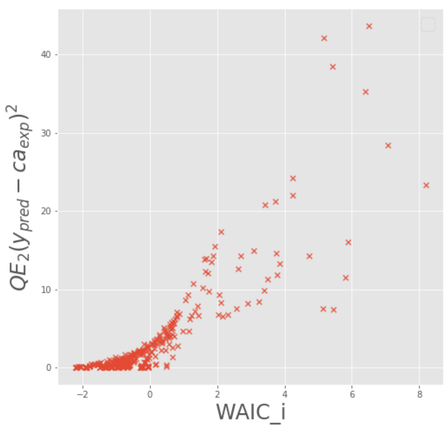
(a) Linear regression between $WAIC_i$ and QE_1.



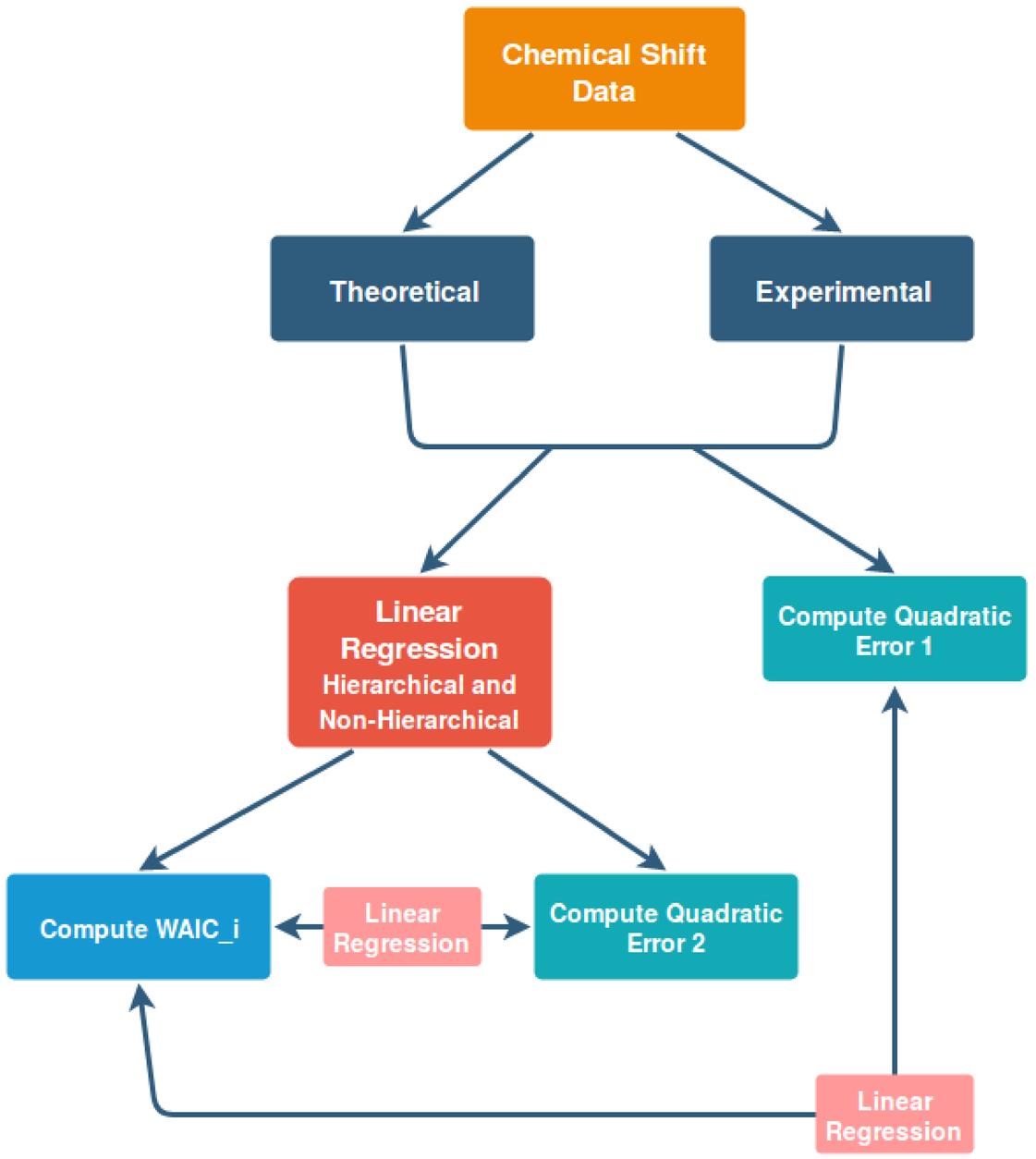
(b) Linear regression between $WAIC_i$ and QE_2.



(c) Scatter plot of $WAIC_i$ and QE_1.



(d) Scatter plot of $WAIC_i$ and QE_2.



Conclusions

In our study we observed that QE_1 as a function of $WAIC_i$ displays an unexpected behaviour. This behaviour seems to be masked when the same regression analysis is performed for the entire data set of proteins (not shown), due to observation superposition. This is related to the references used to define and compute theoretical chemical shifts. We conclude that the references must be updated for every protein and that is not convenient to use the same reference for the entire data set. On non-hierarchical models, QE_2 and $WAIC_i$ contain the same information about a model's fit, as expected. Otherwise, in hierarchical models, QE_2 is not very informative compared to $WAIC_i$ for a range of observations. In consequence, $WAIC_i$ and hierarchical models could be a sound alternative to QE for the evaluatuion of biomolecular structural data.

References

[1] Richard McElreath. *Statistical Rethinking: a Bayesian Course with Examples in R and Stan. McElreath, 2015 ISBN 978-1-482-25344-3.*

[2] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[3] Eldon L. Ulrich, Hideo Akutsu, Jurgen F. Doreleijers, Yoko Harano, Yannis E. Ioannidis, Jundong Lin, Miron Livny, Steve Mading, Dimitri Maziuk, Zachary Miller, Eiichi Nakatani, Christopher F. Schulte, David E. Tolmie, R. Kent Wenger, Hongyang Yao, and John L. Markley. Biomagresbank. *Nucleic Acids Research*, 36(suppl_1):D402–D408, 2008.

[4] Osvaldo A. Martin, Jorge A. Vila, and Harold A. Scheraga. Cheshift2: graphic validation of protein structures. *Bioinformatics*, 28(11):1538–1539, 2012.

[5] John Salvatier, Thomas V. Wiecki, and Christopher Fonnesbeck. Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2:e55, 2016.

Acknowledgements

This work was supported by grant PICT-0218 PICT-Joven Plan Argentina Innovadora 2020: Probabilistic programing for Structural Bioinformatics; PICT-0767 FONCyT: Validation and Determination of nucleic acid structures from NMR 13C chemical shifts and PICT-0556 FONCyT: Determination, Validation and Refinement of Glycans and Glycoprotein structures