A Pipeline for the training of NMR chemical shift prediction models

A.K. Dehof[1], H.P. Lenhof[1], A. Hildebrandt[2]

[1]Center for Bioinformatics, Saarland University,

[2]Johannes-Gutenberg-Universität Mainz

NMR chemical shift prediction plays an important role in many applications in computational biology [1]. Among others, structure determination, structure optimization, and the scoring of docking results can profit from efficient and accurate chemical shift estimation from a three-dimensional model of the molecule under consideration. The development of novel prediction techniques is a challenging task. The required information is spread over several databases and stored in hard-to-parse file formats which sometimes contain serious errors. In addition, the computation of physical terms or of molecular features for a heuristic approach requires complex molecular data structures and algorithms.

Here, we present a pipeline for developing hybrid NMR chemical shift prediction methods that combine physical terms – approximations to quantum mechanical effects – with a statistical model. The pipeline allows the simple import of data from diverse sources, such as the BMRB and the PDB. Several semi-classical terms for shift prediction are implemented and readily available. As of now, we include random coil contributions, aromatic ring current effects, electric field contributions, and hydrogen bonding effects. The feature set for the training of the statistical term encompasses sequential, structural (angles, surface, and density), force-field based, and experimental properties. All features are computed using our open source library BALL [2], and can be easily extended.

For the statistical contribution we propose a random forest model which has demonstrated in our experiments to yield very accurate and stable results. In general, however, the pipeline is model-agnostic and can be used with any regression technique implemented in R.

[1] D. S. Wishart, *Progress in Nuclear magnetic resonance spectroscopy*, **2011**, *58(1)*, 62-87

[2] A. Hildebrandt, A.K. Dehof, A. Rurainski, A. Bertsch, M. Schumann, N.C. Toussaint, A. Moll, D. Stockel, S. Nickels, S.C. Mueller, H.P. Lenhof, and O. Kohlbacher, *BMC Bioinformatics,* **2010**, *11*, 531