Structure based chemical shift prediction using Random Forests non-linear regression

K. Arun[†] **Christopher James Langmead**[*†]

July 2005
CMU-CS-05-163

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

[*]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA 15213.
[†] Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA, USA 15213

E-mail: cjl@cs.cmu.edu

This research is supported by a Young Pioneer Award to C.J.L. from the Pittsburgh Lifesciences Greenhouse.

Keywords: computational biology, structural biology, Nuclear Magnetic Resonance, NMR, chemical shift, regression, Random Forests

Abstract

Protein nuclear magnetic resonance (NMR) chemical shifts are among the most accurately measurable spectroscopic parameters and are closely correlated to protein structure because of their dependence on the local electronic environment. The precise nature of this correlation remains largely unknown. Accurate prediction of chemical shifts from existing structures' atomic co-ordinates will permit close study of this relationship. This paper presents a novel non-linear regression based approach to chemical shift prediction from protein structure. The regression model employed combines quantum, classical and empirical variables and provides statistically significant improved prediction accuracy over existing chemical shift predictors, across protein backbone atom types. The results presented here were obtained using the Random Forest regression algorithm on a protein entry data set derived from the RefDB re-referenced chemical shift database.

1 Introduction

Any nucleus with spin $I = 1/2$, when placed in an external magnetic field, will exhibit two spin states with an energy differential directly proportional to the strength of the applied magnetic field. Each nucleus, however, is influenced by the electrons in its vicinity and therefore the effective magnetic field at the nucleus is attenuated depending upon this electronic environment. The *chemical shift* (δ) is a measure of the electronic shielding that leads to this magnetic field attenuation, and therefore provides an accurate description of the local electronic environment. Thus, chemical shifts are among the most fundamental of nuclear magnetic resonance (NMR) spectral parameters. Chemical shifts are also among the most accurately measurable quantities in NMR spectroscopy (accuracy up to one part in a billion).

Given these properties of the chemical shift, there has long been an interest in understanding the nature of the relationship between molecular structure and shift, and applying said knowledge to infer additional structural information about the molecule under study. Protein molecules too give rise to NMR spectra in an applied magnetic field in a fashion intricately dependent on their three dimensional structures. The electronic environment around nuclei in the case of proteins is influenced by factors such as neighbor anisotropy, ring current anisotropy, hydrogen bond effects, and through-space electric field effects among others. A graphical representation of the chemical shift measurements from a standard protein NMR experiment (^{1}H/^{15}N HSQC) is depicted in Fig. 1. The center of each of the peaks observed in this two-dimensional plot represents two chemical shifts, the ^{1}H and ^{15}N shifts. The axes of the spectrum are in parts per million (ppm), the standard unit for chemical shifts.

Knowledge of the chemical shift and insight into the structure-shift relationship is useful in many contexts. The most obvious application is resonance assignment in the context of an protein NMR experiment where a model of the target protein's structure is available[10] (either via independent X-ray crystallography experiments or comparative modeling). Predicted shifts may also be used to refine existing structural models. There have also been efforts to infer low-resolution structure models given just the experimental chemical shifts. Examples include techniques for secondary structure prediction[16, 18], backbone torsion angle prediction[5], fold recognition[11, 13, 20], protein-protein docking[6] and modeling ligand interactions[15]. Predicted shifts, subject to their having acceptable accuracy, may be similarly employed.

Existing approaches to chemical shift prediction from protein structure apply quantum mechanical, classical and/or empirical methods to the atomic co-ordinate data. Examples of such algorithms include SHIFTS[19], SHIFTX[14] and PROSHIFT[12]. SHIFTS takes a quantum mechanical approach and employs a pre-calculated database of tri-peptide shifts (via density functional theory), while SHIFTX uses a hybrid empirical/semi-classical approach involving pre-calculated chemical shift hypersurfaces and equations for ring current, electric field, hydrogen bonding and solvent effects. PROSHIFT uses a neural network trained on $\sim$ 69,000 experimentally determined chemical shifts. Each of these shift prediction approaches has unique limitations either in terms of the size and composition of the training and/or test data sets, or due to the general tendency for

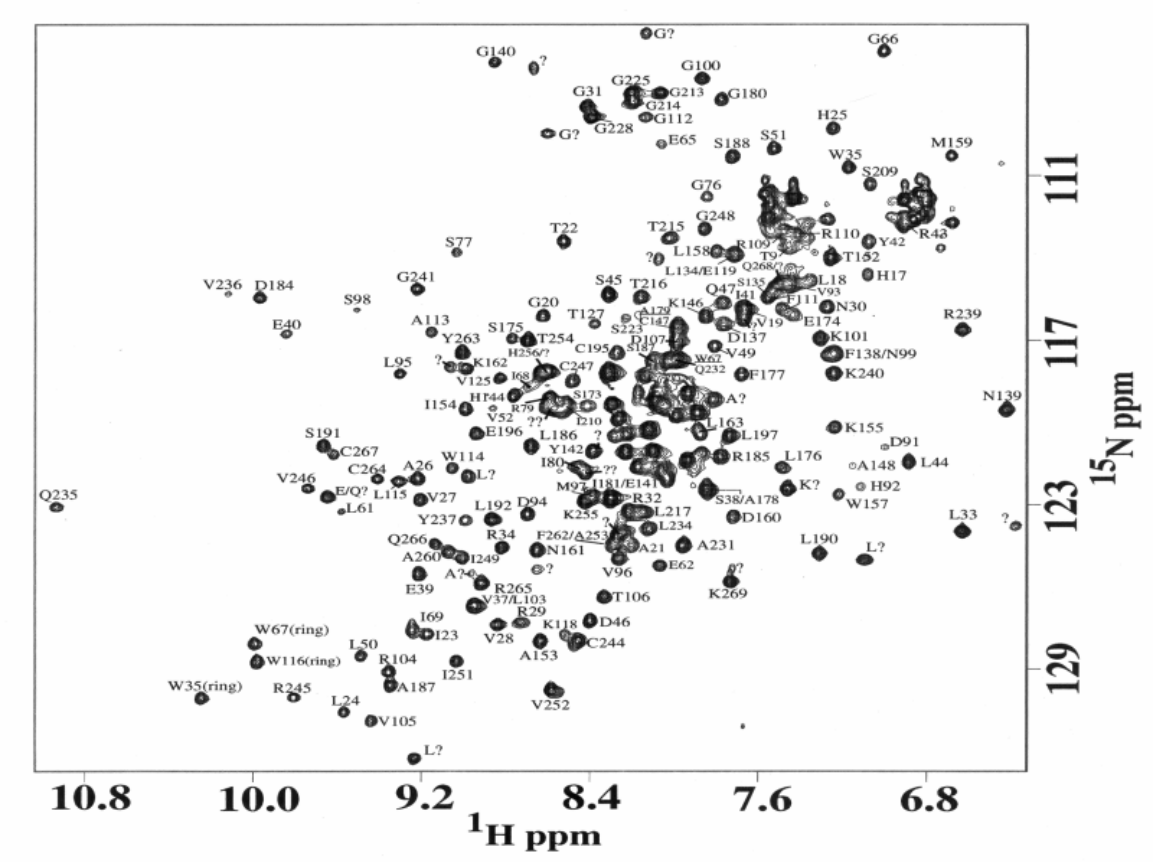

Figure 1: Two-dimensional ^{1}H/^{15}N Heteronuclear Single-Quantum Coherence (HSQC) NMR spectra of an *E. coli* DNA glycosylase, Fpg. (from `http://www.emsl.pnl.gov/homes/msd/bionmr/Buchko_Fpg_poster/Fpg.htm`)

learning methods such as neural networks to over-fit training data. We hypothesized that a better chemical shift predictor could be built by layering an ensemble machine learning algorithm (Random Forests[4]) capable of non-linear regression on top of these existing predictors in addition to expanding the feature set by taking into account numerous empirical structural features such as solvent accessibility, secondary structure and model quality. This paper presents the results of such an exercise.

In brief, the non-linear regression approach to chemical shift prediction employing the ensemble machine learning Random Forest algorithm outperformed each of the underlying shift prediction programs (*viz.* SHIFTS, SHIFTX, PROSHIFT) across all six backbone atom types. These improvements in prediction accuracies were measured in terms of root mean squared error from experimentally recorded shifts and in the case of the Random Forest algorithm, they ranged between 3% to $\sim$18% when compared to the best performer amongst the aforementioned prediction programs. The decrease in error observed was proven to be statistically significant by comparing the distribution of errors using a standard *t*-test. Across all atom types, *p*-values $\ll 0.001$ were observed.

2 Systems and methods

2.1 Data assembly

Building a structure-based chemical shift prediction method requires a dataset of protein chains with experimentally recorded chemical shifts matched to structures solved by NMR or X-ray crystallography. The principal community repositories of chemical shift and structural (atomic co-ordinate) data are the BioMagResBank[3] (BMRB) and the Protein Data Bank[1] (PDB) respectively. However, it has been demonstrated that significant chemical shift referencing errors exist for a substantial portion of the BMRB data. Hence, the dataset used in this project is drawn from the RefDB[21] database - a carefully re-referenced set of chemical shifts derived originally from the BMRB. The RefDB also provides a sequence-based mapping to PDB entries for each set of re-referenced shifts. The sub-set of the RefDB entries selected was free of complexes and mapped to 454 PDB entries.

Metadata and structural information from each of the 454 PDB entries were extracted and each entry was split up into its constituent fragments. In this context, a fragment is defined as a single contiguous polypeptide chain present as part of a potentially larger protein structure with multiple such chains. These fragments were then processed through each of the three chemical shift predictors — PROSHIFT, SHIFTS and SHIFTX. STRIDE[7] secondary structural information was obtained for each fragment from the S2C[17] database. Additionally, a per-residue solvent exposure term was calculated using the half-sphere exposure $HSE - \beta$[8] measure. All structural information and predicted shifts partitioned by protein backbone atom type were stored in a relational database using appropriate schema.

A mapping between the residues in a PDB fragment and those in a RefDB entry with experimental shifts is required to be able to compare the predicted shifts with experimental shifts. Alignments between the corresponding residue sequences were generated using a simple pairwise

Feature	Description
aa	Amino acid residue
sec_str	STRIDE secondary structure
solv_exp	Half-sphere solvent exposure ($HSE - \beta$) terms
$q_{i-1}^{\phi,\psi}$	Contribution from preceding residue's backbone torsion angles
$q_{i}^{\phi,\psi}$	Contribution from target residue's backbone torsion angles
$q_{i+1}^{\phi,\psi}$	Contribution from succeeding residue's backbone torsion angles
q_{i-1}^{χ}	Contribution from preceding residue's type and χ_1 torsion angle
q_{i}^{χ}	Contribution from target residue's type and χ_1 torsion angle
e^{HB}	Hydrogen bond contributions
rand_coil	Random coil reference shift value
pred_shifts	Predicted shifts from SHIFTS, SHIFTX and PROSHIFT

Table 1: Feature set employed in regression for protein backbone heavy atoms

dynamic programming alignment algorithm provided by Biopython[2].

2.2 Feature extraction

Chemical shifts can be predicted from structural models in three ways: using quantum mechanics, classical mechanics, and empirical models. A purely quantum approach is theoretically possible but, in the case of most macromolecules the size of typical protein structures, computationally infeasible. Thus, most protein chemical shift prediction methods employ hybrid techniques, combining quantum, classical and empirical approaches in various ways. Examples of such algorithms include SHIFTS (combines quantum and empirical methods), SHIFTX (combines classical and empirical methods) and PROSHIFT (maps a variety of empirically-determined structural features to chemical shifts using neural networks). Our approach employs each of these individual predictors' final predicted shifts as input to a non-linear regression algorithm. Also, the per-residue quantum mechanical contributions calculated by SHIFTS via density functional analysis of tri-peptides are independently included in the feature array. Additionally, the secondary structural assignments and solvent exposure information obtained in the manner described earlier are incorporated on a per-residue basis. Tables 1 and 2 enumerate the specific features employed in predicting backbone heavy atom and proton shifts respectively. Fig. 2 is a flowchart depicting the assembly of data and feature extraction described herein.

2.3 Regression using Random Forests

The proposed regression model has the form :

$$\delta_i = f(\vec{x_i}) \tag{1}$$

where δ_i is the estimated chemical shift for the *i*th nucleus, $f(\cdot)$ is a non-linear regression function and $\vec{x_i}$ is a vector whose components encode the variables of the regression model. These variables

Feature	Description
aa	Amino acid residue
sec_str	STRIDE secondary structure
solv_exp	Half-sphere solvent exposure ($HSE - \beta$) terms
e^{RC}	Ring current contributions from neighboring aromatic rings
e^{E}	Electrostatic contributions from nearby point charges
e^{PA}	Peptide group anisotropy
rand_coil	Random coil reference shift value
pred_shifts	Predicted shifts from SHIFTS, SHIFTXand PROSHIFT

Table 2: Feature set employed in regression for protein backbone protons

correspond to computable properties in each nucleus' environment and are essentially the features described in the section above. The algorithm selected for implementing the regression function in this set of experiments is Random Forest regression[4]. Random Forest consists of a collection of regression trees, each regression tree itself being a regression function. Each of these trees predicts a real value by querying a set number of variables and instances within the regression model. Each regression tree is thus trained on a different bootstrap sample of both training instances and features. The Random Forest then averages the predictions made by the trees in the forest to produce the final output. Random Forest is thus an example of an ensemble method of machine learning.

Since each base regressor (tree) in the forest trains on a unique randomized sub-set of data and variables, no single tree over-fits the training data. More formally, statistical learning theory[9] decomposes error into bias and variance. The goal of every machine learning algorithm is to reduce both these quantities. Unfortunately, there is a fundamental trade-off between the two, and most algorithms opt to reduce bias at the expense of variance. High error variance however is a sign of over-fitting and any algorithm that over-fits its training data will not generalize well to novel test instances (test error will be high). Random Forest is a variance reduction technique and has provable properties with regard to resisting over-fitting. In contrast, algorithms such as the neural networks employed by PROSHIFT have no such guarantee. Additionally, Random Forests are very efficient to train and test, and have built-in mechanisms for estimating test error and confidence in each prediction made.

In the experiments described, Random Forests were trained for each nucleus type on the given set of features and the accuracy of the final predicted shifts was estimated using 10-fold cross-validation. Chemical shift prediction accuracies are reported for the H^{α}, H^{N}, ^{15}N, $^{13}C^{\alpha}$, $^{13}C^{\beta}$and $^{13}C'$ backbone atom types in terms of root mean squared error (RMSE) from the experimental value. These RMSE values are compared to similar values obtained for the three component chemical shift predictors, PROSHIFT, SHIFTS, and SHIFTX. The *p*-values of decreases in RMSE are calculated using a standard *t*-test to assess the significance of improvements in prediction accuracy.

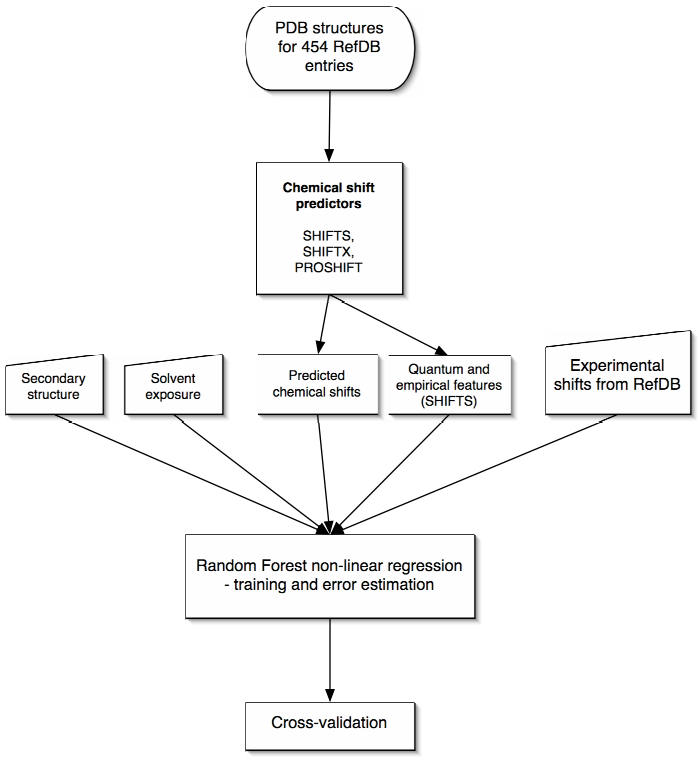

Figure 2: Flowchart depicting the experimental procedure involved in training Random Forest regressors

		SHIFTS	SHIFTX	PROSHIFT	RANDOM FOREST
Nucleus	Instances	RMSE (ppm)	RMSE (ppm)	RMSE (ppm)	RMSE (ppm)
H^N	46,991	0.66	0.63	*0.58*	**0.49** (15.5%)
H^α	38,767	0.79	0.36	*0.34*	**0.28** (17.7%)
^{15}N	40,166	5.29	3.51	*3.44*	**2.93** (14.8%)
$^{13}C^\alpha$	37,006	1.86	*1.64*	2.59	**1.51** (7.9%)
$^{13}C^\beta$	29,809	3.13	*3.02*	3.75	**2.93** (3%)
$^{13}C'$	24,253	1.89	*1.40*	2.34	**1.19** (14.9%)

Table 3: Chemical shift prediction accuracies for individual shift predictors and Random Forest regression in terms of root mean squared error (RMSE) from experimental values. The values in italics identify the least RMSE value amongst the SHIFTS, SHIFTX and PROSHIFT predictors for that atom type. The values in bold type identify the best overall predictor, which is the Random Forest approach for all nuclei. The percentage figures in parentheses in the Random Forest column represent the decrease in RMSE as a percentage of the least RMSE value amongst the underlying predictors.

3 Results and discussion

The database of chemical shifts employed in this exercise consisted of between 24,000 to 47,000 separate chemical shifts depending on the nucleus type. These were mapped to 454 different protein structures from the PDB. The results obtained by training Random Forest regressors for each nucleus type (subject to 10-fold cross-validation) are shown in table 3. Prediction accuracies are reported in terms of root mean squared error (RMSE) from experimental shift values. It is seen that the Random Forest predictions are 15.5%, 17.7%, 14.8%, 7.9%, 3% and 14.9% more accurate than the best of SHIFTS, SHIFTX, and PROSHIFT for H^α, H^N, ^{15}N, $^{13}C^\alpha$, $^{13}C^\beta$ and $^{13}C'$ nuclei respectively. The *p*-values of these decreases in RMSE, based on *t*-tests on the residuals, are each $\ll 0.001$, thereby indicating that the decreases in error are statistically significant. Note, that although the $^{13}C^\beta$ RMSE value shows only modest improvement (3%) when predicted using the Random Forest algorithm, a separate experiment (data not shown) where rotameric configurations served as a feature resulted in an RMSE drop of greater than 7% for the same nucleus. This is to be expected since the configuration of the sidechain and the resultant distribution of the sidechain electrons likely have a significant influence on the $^{13}C^\beta$ chemical shift. This also indicates that the same set of regression features may not be optimal for every type of nucleus.

It is clear from these results that the Random Forest-based non-linear regression approach to shift prediction promises significant improvements in prediction accuracy over existing methods. Apart from the resistance of the technique to over-fitting, it is to be noted that the size of the training data set employed in this exercise is significantly larger than any prior comparable effort. This, in turn, will allow this prediction method to better generalize to novel protein structures. Also, given that Random Forests are extremely efficient to train and each tree in the forest can be grown in parallel, additional structural variables may be rapidly tested for their contribution to improvement in shift prediction accuracy. Experiments using B-factors from X-ray crystallographic structures and discrete per-residue rotamer library categories as additional features are currently in progress.

The method reported here is also notable for the fact that it is a hybrid meta-prediction approach, combining quantum, classical and empirical information about protein structures. Purely quantum mechanical approaches to shift prediction work well for small molecules but are computationally infeasible for anything the size of a typical protein structure. Conversely, purely empirical approaches are unlikely to capture all the complexity inherent in the factors affecting the electronic environment which finally dictates the chemical shift. The meta-prediction aspect, wherein predictions from multiple underlying chemical shift predictors (PROSHIFT, SHIFTS and SHIFTX in this case) are incorporated as input to the regression algorithm, allows for a judicious combination of information from both approaches to be incorporated into a single prediction technique. Meta-prediction approaches have been successfully used in secondary and tertiary structure prediction and ligand docking. The results obtained indicate that chemical shift prediction is also a suitable candidate for this approach.

4 Conclusion

We have shown that a non-linear regression approach to chemical shift prediction employing a ensemble machine learning approach has the potential to improve chemical shift prediction accuracy significantly. The ensemble Random Forest algorithm employed is provably resistant to over-fitting the test data and generalizes well to novel test instances. This is demonstrated by the improvement in shift prediction accuracy seen in the 10-fold cross-validation exercise over existing chemical shift predictors across all six protein backbone nuclei. Random Forests allow for rapid training of regressors and are eminently parallelizable, therefore permitting one to explore the protein structural variable space efficiently. They make feasible the potential training of separate regressors for varied partitions of the training data set (all NMR structures versus all X-ray structures, per amino acid type regressors, per secondary structure type regressors etc.). It is possible that a future variant on this method will render predictions by using such different regressors internally to predict on different partitions of the test data.

The availability of a rapid, accurate and easily adapted method of chemical shift prediction will make it easier to study the relationship between shift and structure. Any technique that incorporates chemical shift prediction, such as NMR resonance assignment, low resolution structure prediction, fold recognition, protein docking and ligand interaction modeling, will benefit from the increased accuracy provided by this method. Additionally, the speed of training of the Random Forests will permit domain-specific regressors to be trained in these endeavors.

5 Acknowledgment

The authors would like to thank Drs. Robert Murphy and Gordon Rule for enlightening discussions on topics relevant to this work, and Tal Blum for a critical insight into the cross-validation process. K.A. was partially supported by a Merck Computational Biology fellowship for the duration of this work. This work is supported by a Young Pioneer Award to C.J.L. from the Pittsburgh Life Sciences Greenhouse.

References

[1] H.M. Berman, J. Westbrook, Z. Feng, T.N. Bhat G. Gilliland, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235–242, 2000.

[2] The BIOPYTHON project. URL: `http://www.biopython.org`.

[3] BioMagResBank (BMRB), a NIH funded bioinformatics resource, Department of Biochemistry, University of Wisconsin-Madison, Madison, WI USA (URL: `http://www.bmrb.wisc.edu`) Grant: LM05799-02.

[4] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[5] G. Cornilescu, F. Delaglio, and A. Bax. Protein backbone angle restraints from searching a database for chemical shift and sequence homology. *J. Biomol. NMR*, 13(3):289–302, 1999.

[6] C. Dominguez, R. Boelenc, and A. Bonvin. Haddock: a protein-protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.*, 125(7):1731–1737, 2003.

[7] D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins*, 23(4):566–579, 1995.

[8] T. Hamelryck. An amino acid has two sides: a new 2D measure provides a different view of solvent exposure. *Proteins*, 59(1):38–48, April 2005.

[9] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer-Verlag, 2001.

[10] J. Hus, J. Prompers, and R. Brüschweiler. Assignment strategy for proteins of known structure. *J. Mag. Res.*, 157:119–125, 2002.

[11] C.J. Langmead and B.R. Donald. High-throughput 3D homology detection via NMR resonance assignment. In *Proc. IEEE Computer Society Bioinformatics Conference (CSB)*, pages 278–289, 2004.

[12] Jens Meiler. PROSHIFT: Protein chemical shift prediction using artificial neural networks. *Journal of Biomolecular NMR*, 26:25–37, 2003.

[13] S. Mielke and V. Krishnan. Protein structural class identification directly from NMR spectra using averaged chemical shifts. *Bioinformatics*, 19(16):2054–2064, 2003.

[14] Stephen Neal, Alex M. Nip, Haiyan Zhang, and David S. Wishart. Rapid and accurate calculation of protein 1H, 13C and 15N chemical shifts. *J. Biomol. NMR*, 26:215–240, 2003.

[15] C. Peng, S. Unger, F. Filipp, M. Sattler, and S. Szalma. Automated evaluation of chemical shift perturbation spectra: new approaches to quantitative analysis of receptor-ligand interaction NMR spectra . *J. Mol. Biol.*, 29(4):491–504, 2004.

[16] AB. Sibley, M. Cosman, and VV. Krishnan. An empirical correlation between secondary structure content and averaged chemical shifts in proteins. *Biophys J*, 84(2 Pt 1):1223–1227, Feb 2003.

[17] G. Wang, J.W. Arthur, and R.L. Dunbrack. S2C: A database correlating sequenceand atomic co-ordinate numbering in the Protein Data Bank. URL: `http://dunbrack.fccc.edu/Guoli/s2c/`, 2002.

[18] D. Wishart and B. Sykes. The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data. *J. Mol. Biol.*, 4(2):171–180, 1994.

[19] XP. Xu and DA. Case. Automated prediction of 15N, 13Calpha, 13Cbeta and 13C' chemical shifts in proteins using a density functional database. *J Biomol NMR*, 21:321–333, Dec 2001.

[20] H. Zhang, A. Leung, and D. Wishart. THRIFTY. URL: `http://redpoll.pharmacy.ualberta.ca/thrifty`, 2005.

[21] Haiyan Zhang, Stephen Neal, and David S. Wishart. RefDB: A database of uniformly referenced protein chemical shifts. *J. Biomol. NMR*, 25:173–195, 2003.