Protein NMR Reference Correction–A Statistical Approach for an Old Problem

Xi Chen[1,2,3,4], and Hunter N.B. Moseley[1,3,4,5]

[1]Department of Molecular & Cellular Biochemistry, [2]Department of Statistics, [3]Markey Cancer Center,
[4]Center for Environmental and Systems Biochemistry, [5]Institute for Biomedical Informatics, University of Kentucky

Introduction

NMR is a powerful and established tool for studying biomacromolecules. However, accurate chemical shifts assignments are a requirement for many aspects of biomacromolecular NMR, especially protein structure determination. While 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) is the established reference standard for biomacromolecular NMR, proper use and referencing of NMR spectra using DSS is a non-trivial step (especially for non-experts). Therefore, computational methods for accurately detecting and correcting referencing errors are needed.

We are developing a unique statistical-based method to refine reference values by comparing assigned amino acid composition probabilities with the known amino acid composition of the protein being investigated. The method estimates the probability that C_α and C_β resonance pairs from the NMR data arising from 19 standard amino acids (excluding glycine) and sums the probabilities across all resonance pairs to give an estimate of amino acid composition. Next, the method employs a simple grid search of the chemical shift reference value to find a minimum difference between predicted and actual amino acid composition.

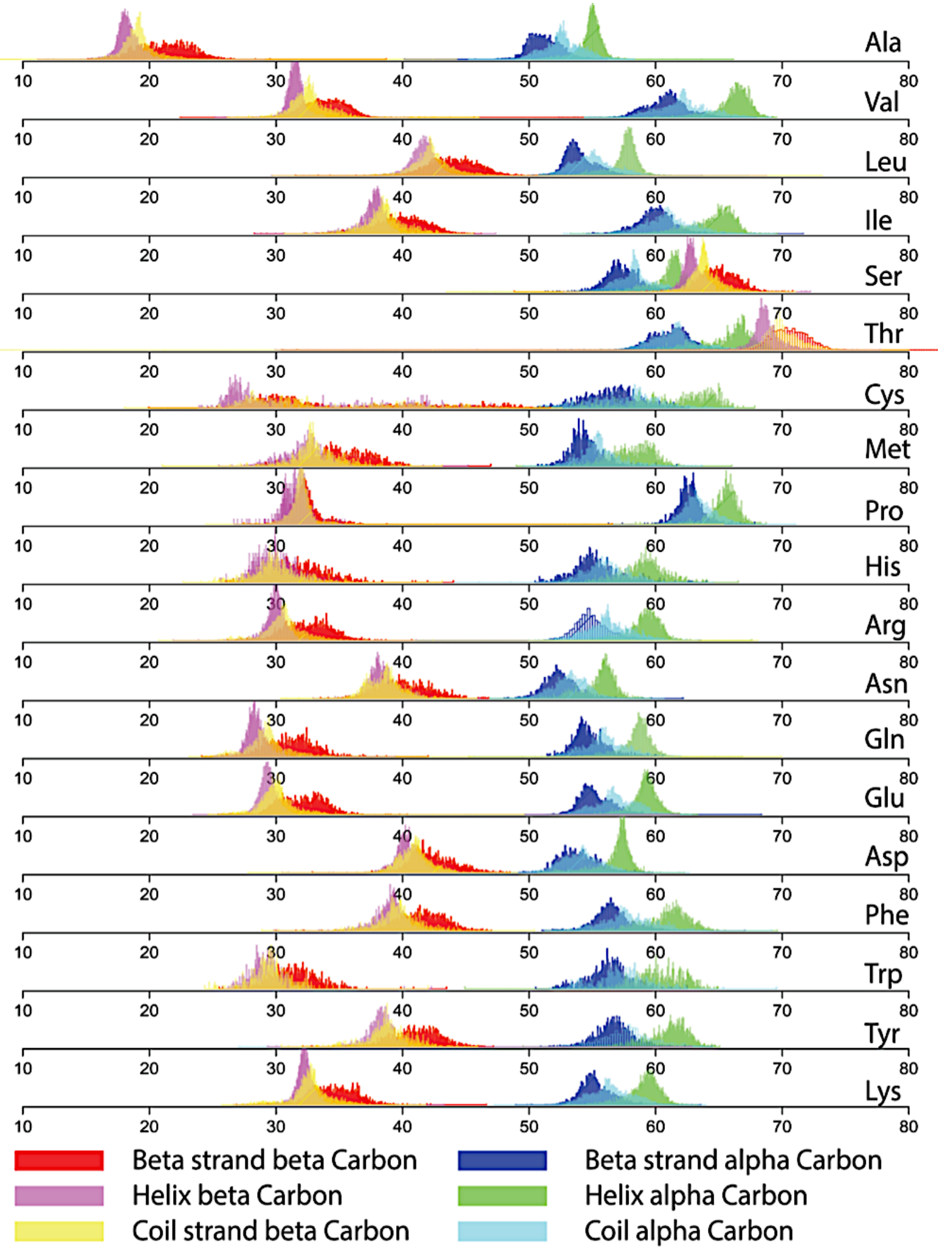


Figure 1. C_α and C_β chemical shifts distribution summary

Protein NMR Statistical Features

The key concept of this software package resides in the statistical features of the chemical shifts and the separate distributions due to amino acid type and secondary structure. Overall, chemical shifts distributions indicate that the alpha carbon (C_α) is in the range of 50-70 ppm and the beta carbon (C_β) is in the range of 15-45 ppm, with exceptions for glycine, threonine, and serine [Figure 1].

Figures 2 and 3 below are 2D scatter plots of C_α and C_β chemical shifts from the Re-referenced Chemical Shifts Database (RefDB) for a few amino acid types.

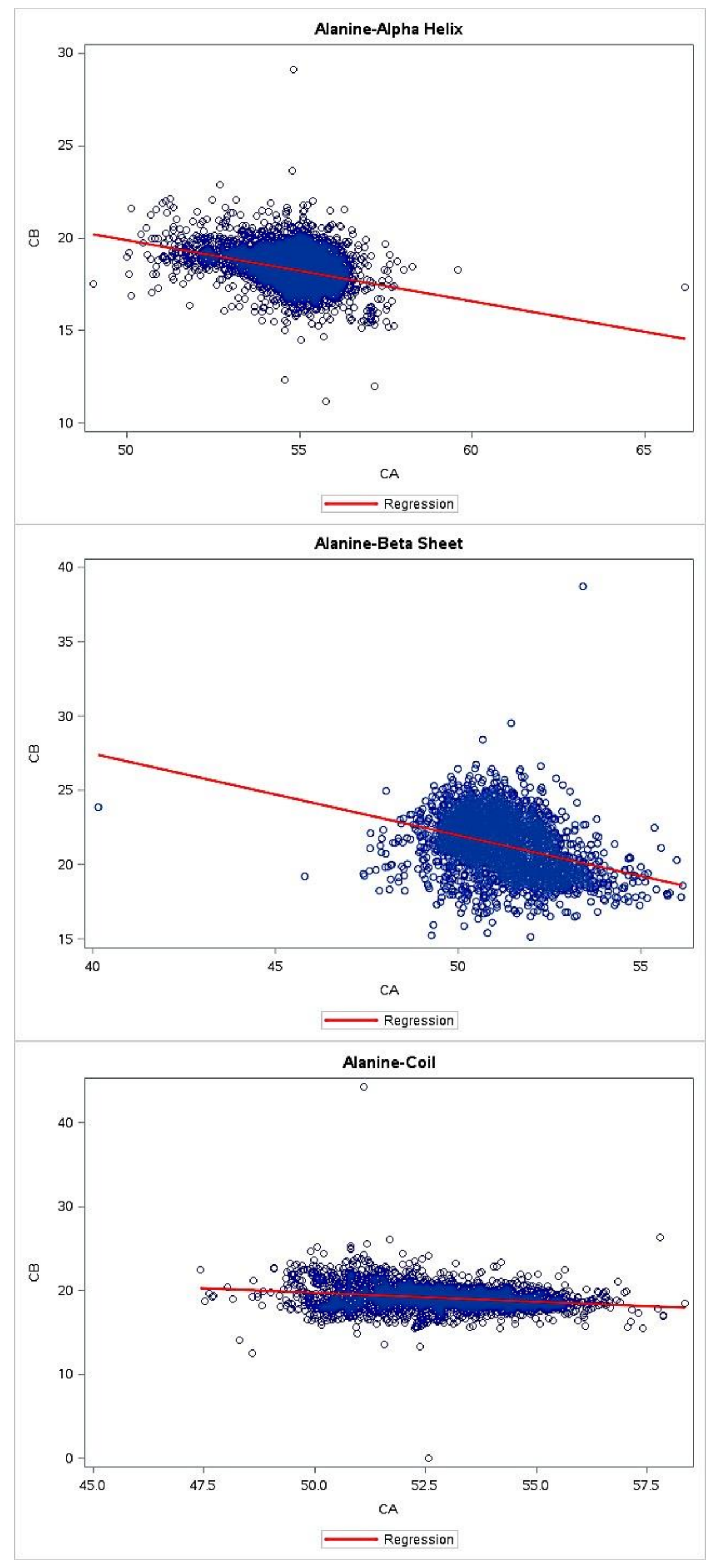


Figure 2. Alanine C_α and C_β chemical shifts distribution summary

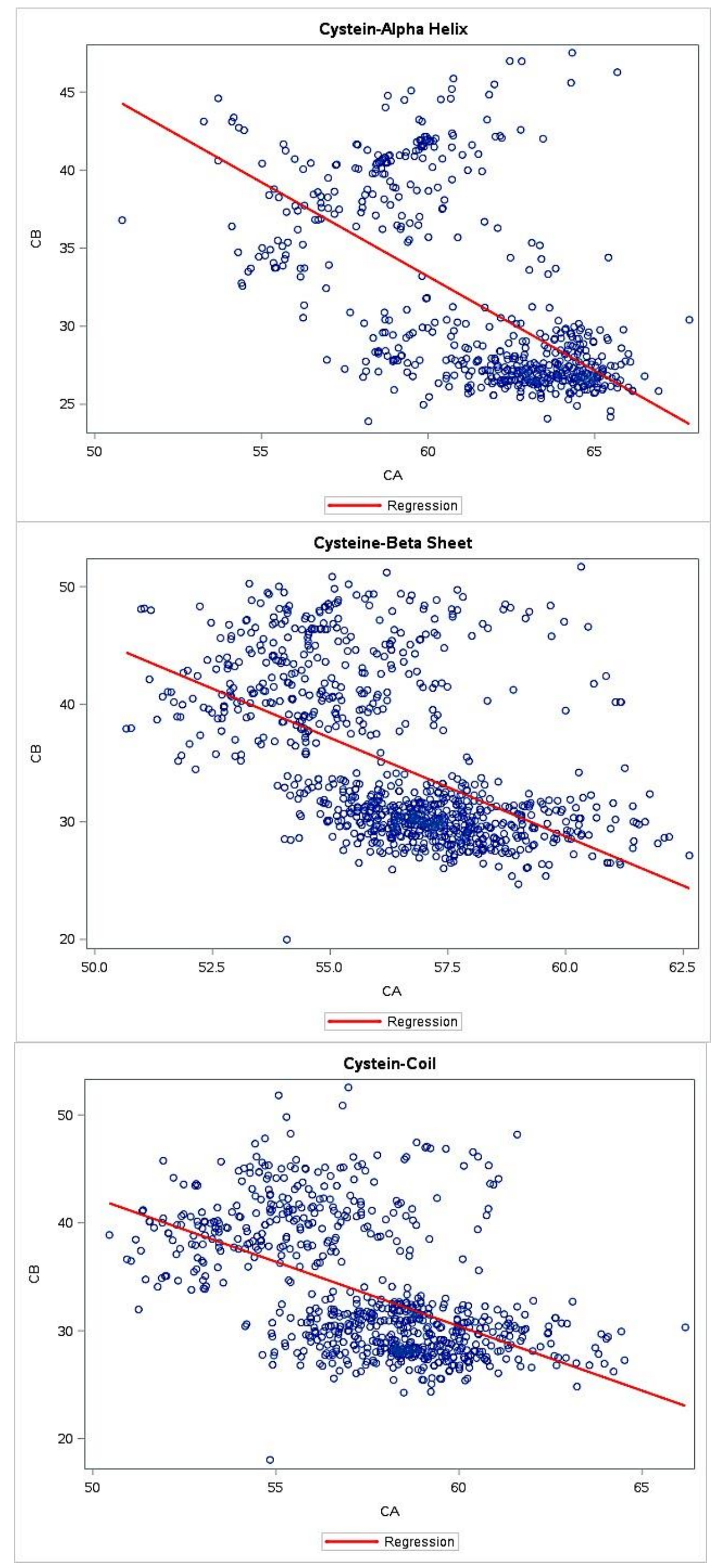


Figure 3. Cysteine Cα and Cβ chemical shifts distribution summary

Algorithm Formulism

The chemical shifts between C_α and C_β are not independent [Figure 2, 3]; they are correlated. This is an important observation that allows us to use statistical theorem to estimate the probability that each pair of C_α and C_β resonances represents each of the 19 amino acid types:

$$(X-\bar{X})^T\Sigma(X-\bar{X})\sim\chi_2^2$$

Covariance Matrix

During the analysis, we realized that the cysteines needed to be treated as separate states based on its two dominant chemical forms—oxidized and reduced. The overly-spread distribution of the cysteine chemical shifts in Figure 1 visualizes the fundamental problem caused by treating them as one state.

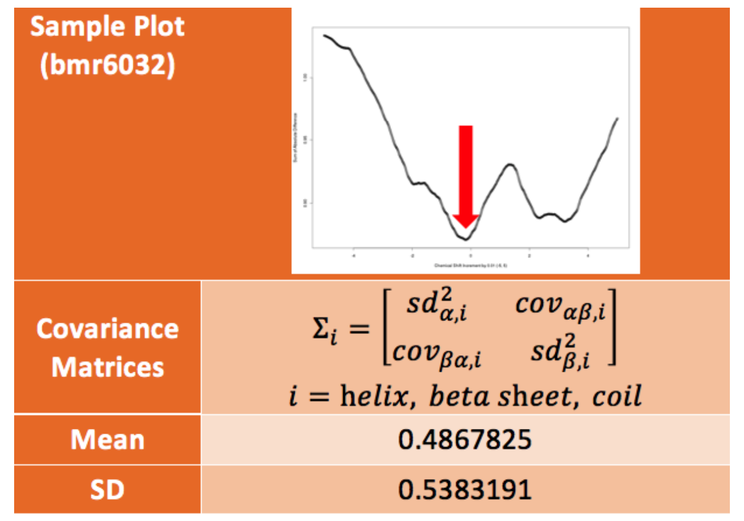
Highlight Core SAS® Software Code

```
/* Structure the chi_square function.*/

proc iml;
        start Chi_star_f(r, secondaryStructure);
                use invcov;
                read all var{SD_CA COV_1 COV_2 SD_CB} into X where(SS = secondaryStructure);
                Chi_star = j(nrow(X), 1);
                do i = 1 to nrow(X);
                        invMat = j(2,2, (X[i]));
                        chiStr = r*invMat*r`;
                        Chi_star[i] = chiStr;
                end;
                close invcov;
                return (Chi_star);
        finish;
        store module=Chi_star_f;
quit;
```

Preliminary Results

A prototype of the method has been implemented. We have estimated covariance matrices from an analysis of the RefDB, including separate matrices for the two states of cysteines. Figure 4 below summarizes the testing of our method using all 3000 datasets in the RefDB. The top plot shows the sum of absolute differences between predicted and actual amino acid frequencies for the bmr6032 dataset as chemical shifts are shifted in a grid search. The plot shows two deep local minima and a middle weak local minimum corresponding to three major types of secondary structure: helix, beta-sheet, and coil. The mean absolute correction across all RefDB datasets is 0.49 ppm with a standard deviation of 0.54 ppm, Which is robust enough for detecting and correcting significant referencing errors.



Sample Plot (bmr6032)	
Covariance Matrices	$\Sigma_i=\begin{bmatrix} sd_{\alpha,i}^2 & cov_{\alpha\beta,i} \\ cov_{\beta\alpha,i} & sd_{\beta,i}^2 \end{bmatrix}$ $i = helix,\ beta\ sheet,\ coil$
Mean	0.4867825
SD	0.5383191

Figure 4. Sample Results and Algorithm Performance across all the RefDB datasets

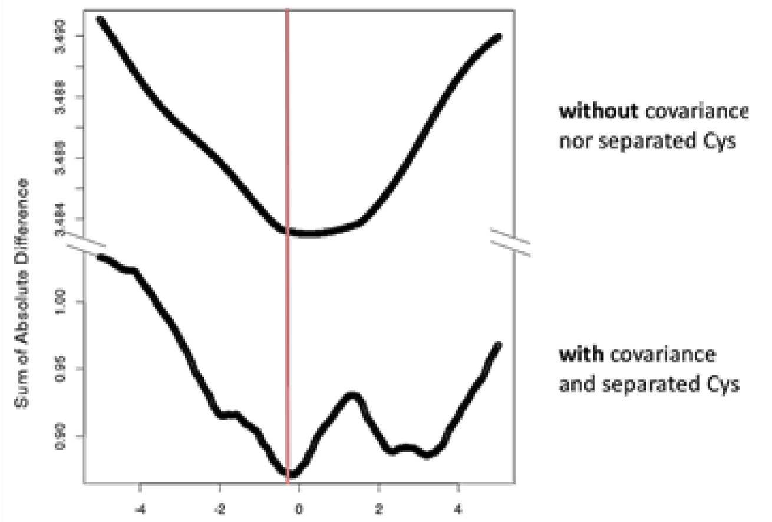


Figure 5. Residuals decrease with incorporation of covariance and Cys treated as two separate AA types

Figure 5 below shows the sum of absolute differences for the bmr6032 dataset for two variations of our method: with and without covariance and separate covariance for reduced and oxidized cysteines. The figure demonstrates a much better prediction of amino acid type when both covariance and separate cysteine states are used. The plot also demonstrates how much more sensitive the covariance with separate cysteine states implementation is to referencing error over the less sophisticated implementation.

Conclusion

This project will provide the biomolecular NMR field with a unique tool that allows the spectral referencing to be corrected at the beginning of protein NMR data analysis. Current reference correction methods rely on later retrospective analyses that requires assigned chemical shifts or even protein structure. Our method should result in more accurate spectral referencing at the beginning of data analysis, improving both the speed and quality of protein resonance assignment and downstream NMR-based analyses including structure determination.