Nonlinear Measures of Correlation and Protein Motion

UMBC REU Site: Interdisciplinary Program in High Performance Computing
Team members: Nancy Hong[1], Emily Jasien[2], Christopher Pagan[3], Daniel Xie[4]
Graduate assistant: Zana Coulibaly[5], Faculty mentor: Kofi P. Adragni[5], Client: Ian F. Thorpe[6]
[1]Stony Brook University, [2]Cal Poly Pomona, [3]CSEE, UMBC, [4]New College of Florida, [5]Math and Stat, UMBC, [6]Chemistry, UMBC

Problem

Allostery is a process in which an event that occurs at one region in a complex macromolecule can create a change at a distant, coupled region in that molecule. We look at this process in proteins and the motional correlations that result.

A commonly used measure of correlation, Pearson's correlation coefficient, is suitable only for measuring linear trends. When dealing with spatial information, however, we desire a measure of correlation that can detect non-linear relationships as well.

How can we find an accurate method to determine both linear and non-linear motional correlations in proteins?

Methodology

We use the qth order polynomial regression model of $Y|X = x$ of the form

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_q x_i^q + \varepsilon_i$$

to model our relationships.
To determine the optimal degree for the model, k-fold cross-validation is used. The data set is divided into k bins, with one bin used to test the model created from the other $k-1$ bins by calculating the mean squared prediction error

$$\mathrm{PE}_m = \frac{1}{n}\sum_{l=1}^{n}(Y_i - \hat{Y}_i)^2$$

The model with the smallest mean squared prediction error is chosen. We then estimate the strength of relationships between the variables by

$$R = r_{Y\hat{Y}} = \widehat{\mathrm{Corr}}(Y, \hat{Y})$$

which is our alternative to the Pearson correlation coefficient. Our final results are stored in a correlation matrix containing comparisons for all the variables of interest, $\widehat{\mathrm{Corr}}(Z_i, \hat{Z}_i)$, with Z_i as the regressand and Z_j forming the regressors.

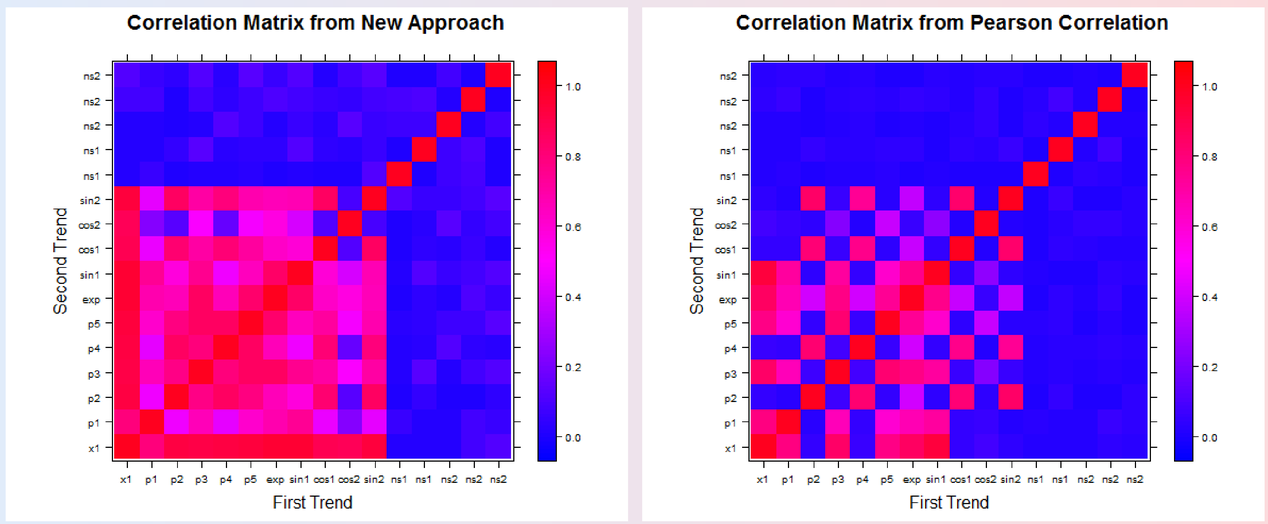
Implementation

We implemented the methodology by using the statistical programming language R. The code calculates the optimal degree of the linear models and produces a matrix of correlations between all random variables. The code was also made for parallel processing using the cluster maya at the High Performance Computing Facility (HPCF). The libraries used are 'MPV', 'Rmpi', and 'snow'.

In order to test our method, we simulated data of various functions that are correlated to see if our methodology produces correct and relatively consistent results. Sample noise data was also simulated to test whether or not the code will find correlations where there should not be. The code generated heatmaps to show correlations between the regressands and regressors.

Results

Heatmap of the correlation matrix generated by our method versus correlation matrix created from the absolute value of Pearson correlation coefficients:



Our measure of correlation picks up many more relationships without significantly picking up noise in the process.

Conclusions

- New method of determining correlation matrix is more effective than Pearson's.
- Using parallel computing methods produce results more quickly and efficiently.
- Can be used to better understand the effects of allostery and other problems where relationships are nonlinear.

References and Acknowledgments

Full technical report: HPCF–2014–11, `www.umbc.edu/hpcf` > Publications.

- REU Site: `www.umbc.edu/hpcreu`, funded jointly by NSF and NSA
- NSF, NSA, UMBC, HPCF, CIRC