Identifying Nonlinear Correlations in High Dimensional Data with Application to Protein Molecular Dynamics Simulations

REU Site: Interdisciplinary Program in High Performance Computing

William J. Bailey[1], Claire A. Chambless[2], Brandynne M. Cho[3], and Jesse D. Smith[4],
Graduate Assistant: Andrew M. Raim[4], Faculty Mentor: Kofi P. Adragni[4],
Client: Ian F. Thorpe[5]

[1]Department of Mathematics and Statistics, Kenyon College
[2]Department of Mathematical Sciences, Elizabethtown College
[3]Department of Mathematics and Computer Science, Saint Mary's College of California
[4]Department of Mathematics and Statistics, University of Maryland, Baltimore County
[5]Department of Chemistry and Biochemistry, University of Maryland, Baltimore County

Technical Report HPCF–2013–12, www.umbc.edu/hpcf > Publications

Abstract

Complex biomolecules such as proteins can respond to changes in their environment through a process called allostery, which plays an important role in regulating the function of these biomolecules. Allostery occurs when an event at a specific location in a macromolecule produces an effect at a location in the molecule some distance away. An important component of allostery is the coupling of protein sites. Such coupling is one mechanism by which allosteric effects can be transmitted over long distances. To understand this phenomenon, molecular dynamic simulations are carried out with a large number of atoms, and the trajectories of these atoms are recorded over time. Simple correlation methods have been used in the literature to identify coupled motions between protein sites. We implement a recently developed statistical method for dimension reduction called principal fitted components (PFC) in the statistical programming language R to identify both linear and non-linear correlations between protein sites while dealing efficiently with the high dimensionality of the data. PFC models reduce the dimensionality of data while capturing linear and nonlinear dependencies among predictors (atoms) using a flexible set of basis functions. For faster processing, we implement the PFC algorithm using parallel computing through the Programming with Big Data in R (pbdR) package for R. We demonstrate the methods' effectiveness on simulated datasets, and apply the routine to time series data from Molecular Dynamic (MD) simulations to identify coupled motion among the atoms.

Key words. Sufficient Reduction, Principal Fitted Components, parallel computing, Molecular Dynamics.

1 Introduction

Complex data are routinely generated by researchers across the applied sciences. These data create challenges for analysis due to their high dimensionality. As a prototypical example, we consider Molecular Dynamics (MD) simulations of complex biomolecules such as proteins. Steady improvements in computational power and algorithm efficiency have dramatically increased the capability and scope of MD simulations, facilitating their use in understanding important problems in chemistry and biology. In the face of this progress, a new challenge has emerged — that of analyzing the large amounts of data generated from such simulations. The high dimensionality of protein conformations sampled during MD simulations calls for efficient and adaptable methods for analysis. One widely studied problem in this scientific domain is that of identifying motional correlations that occur between distinct sites in proteins. Such correlations are known to play important functional roles in these molecules [1]. Of particular interest are correlations that occur between distant protein sites. These distant correlations form the basis of allostery, the process by which an event at one location perturbs the properties of another location in a biomolecule [28]. Allostery is important because it allows complex macromolecules such as proteins to sense and respond to changes in their environment. This phenomenon occurs because different regions of macromolecules are structurally or thermodynamically coupled, allowing information to be communicated over long distances [11]. There is still much that is not understood about allostery, particularly as it relates to regulation of protein function [33]. Open questions include: How can allosteric principles be used to rationally alter the functionality of proteins [16]? Which models of allostery are most relevant in biological systems [11, 19]? What are the evolutionary constraints that govern allostery [32, 5, 6, 21, 13]? Recently, there has been renewed interest in uncovering allosteric interactions in proteins and in discovering new allosteric sites due to their role in protein regulation and their potential as novel sites for drug discovery [18].

The standard tool to assess motional correlations in proteins has been the covariance map, in which correlations between motions in different sites on a molecule i and j are represented by the covariance computed for motions involving these sites. The result of this calculation is a matrix (M) in which each cell (m_{ij}) contains the covariance value computed for sites i and j. This simple, widely used tool has provided much insight into the role that correlated motions can play in protein function [34, 3, 14, 22, 12, 31]. Such maps compactly display motional correlations within a molecular structure and the degree of this correlation. Despite their utility, covariance maps do exhibit disadvantages. Primary among these is that correlation methods such as covariance, while well suited to detect linear relationships among sites, can fail to detect nonlinear statistical dependency [15, 24].

To circumvent the limitations of simple covariance, more elaborate procedures have been proposed. Among these are the mutual information approach of McClendon et al. [18], which identifies statistically significant correlated motions from equilibrium molecular dynamics. It is an entropy-based method that requires large sample sizes and does not address specifically nonlinear relationships among molecules. Another approach is the quasi-anharmonic analysis of Ramanathan et al. [24], which uses higher-order statistics of protein motions to identify sub-states in the conformational landscape. Statistical coupling analysis (SCA) of Lockless and Ranganathan [17] is another approach that involves nonlinear dimensionality reduction of the protein coordinate space [27, 9, 10].

Other nonlinear dimensionality reduction methods have been widely used recently to determine the collective motions from molecular dynamics and to capture nonlinear relationships. These methods include kernel PCA [26], complete isometric feature mapping, or Isomap [29], Locally Linear Embedding or LLE [25], and the diffusion map [7]. However, these are essentially unsupervised methods that attempt to unfold curved manifolds into flat spaces. They often require large sample sizes and are typically not equipped to furnish explicit transformations between the input variables and the low-dimensional embedding, making interpretation difficult. Moreover, some of these methods employ non-orthogonal basis sets that further complicate comparison with conventional correlation analyses [24, 27, 9].

Recent development of dimension reduction methodologies in Statistics pioneered by Cook [8] capture linear and/or nonlinear dependencies between statistical variables through a likelihood-based approach. We adapt these principal fitted components methodologies to the complex, high-dimensional data of protein dynamics in molecular simulations. Two features of our approach make it relevant. First, linear dimension reduction procedures extract information contained in large datasets and avoid the difficulties of high dimensionality in the subsequent analysis while maintaining ease of interpretation. Second, the methodology is well equipped to uncover nonlinear motional relationships.

In the remainder of this report, we provide a description of the statistical methodology (Section 2), including results of the method on simulated small-scale data sets. In addition, we provide details on the parallel implementation; how the methodology was applied to the time series MD simulation data (Section 2.2). Finally, we review the specifics of PFC as applied to MD simulation data (Section 3).

2 Statistical Methods

2.1 Statistical Techniques: Principal Fitted Components

Principal fitted components is a statistical methodology with the primary goal of reducing the dimensionality of a data set. It is an inverse regression methodology with a set of p predictors $\mathbf{X} = (x_1, ..., x_p)^T$ and a response Y. The inverse regression model is concerned with the conditional distribution of predictors $\mathbf{X}$ on the response Y. That is, we model $\mathbf{X}|Y$ rather than $Y|\mathbf{X}$. The general form of the PFC model, as in [2], is

$$E(X|Y) - E(X) = \Gamma \nu_Y + \Delta^{-\frac{1}{2}}\epsilon. \tag{2.1}$$

In this equation, ν_Y is an unknown function of Y, $\epsilon \sim N(0, I)$, and Δ is the conditional variance function assumed to be independent of Y. The most important choice made when constructing the PFC model is in approximating ν_Y with a set of basis functions $\nu_Y \approx \beta \mathbf{f}_Y$, where $\mathbf{f}_Y$ is a collection of functions used to approximate the underlying structure of the data and β is an unconstrained parameter. A good basis approximation requires a flexible set of functions; common basis functions are polynomial, Fourier series, piecewise linear, and piecewise continuous functions. Here, we limit our analysis to polynomial basis functions.

In addition to primary structures in the data, we must identify the error structure the model will have. Several structures are possible; isotropic, anisotropic, and unstrutured error. Isotropic error structure assumes that each predictor is independent, and on the same measurement scale; thus $\Delta = \sigma^2 I_p$ [2, 8]. Anisotropic error structure assumes that the conditional response variables are not on the same measurement scales; so that $\Delta = [\delta_1, \ldots, \delta_p]^T I_p$ [2]. Unstructured error makes no assumptions of this kind, and allows error to be structured in any way; For our purposes, namely, analyzing MD simulation data, we will assume an isotropic error structure. Under the isotropic structure, $\Gamma^T X$ is the sufficient reduction of X, that is, $\Gamma^T X$ retains all the regression information about Y that is contained in X. Thus, X is not needed once $\Gamma^T X$ is obtained.

Clearly then, our primary goal is obtaining an accurate estimate of Γ. Such an estimate can be constructed using the eigenvectors of the fitted covariance matrix of X. We represent the data as the $n \times p$ matrix $\mathbf{X}$ where p is the number of predictor variables, and n is the number of observations. The fitted covariance matrix ($\hat{\Sigma}_{\text{fit}}$) is the result of projecting $\mathbf{X}$ onto the space spanned by the basis function $\mathbf{f}_Y$ and is given by

$$\hat{\Sigma}_{\text{fit}} = \frac{1}{n}\mathbf{X}^T P_f \mathbf{X}. \tag{2.2}$$

The term $P_{\mathbf{f}} = \mathbf{f}_Y({\mathbf{f}_Y}^T\mathbf{f}_Y)^{-1}{\mathbf{f}_Y}^T$ is the projection operator, and $\mathbf{f}_Y$ is an $n \times r$ matrix where r is the number of components of the basis function. Here, since we have limited ourselves to polynomial basis functions, r is the highest degree polynomial we will fit to the data. By choosing to express and reduce our data in terms of the dominant eigenvectors of $\hat{\Sigma}_{\text{fit}}$, we are making a classical assumption in dimension reduction methodology; we assume that the directions of greatest variance are the structures on interest.

2.2 Parallel Implementation of PFC

To implement the PFC model described in 2.1, the statistical programming language R was used. The advantages of R include its high level syntax and native plotting functions, and it is freely available. We adapted the PFC function initially written in the R library package 'ldr' that is available on CRAN to our problem for the parallel implementation.

The PFC implementation found in the CRAN library 'ldr' quickly produced a correct solution for small (e.g., 300×40) data sets, but for larger sets (e.g., 1500×100) the calculation took several hours. With this in mind a parallel wrapper was created in order to speed up the computation of correlations. The final program calculated correct correlation matrices were calculated within seconds rather than hours; thus providing a more efficient way to uncoverg accurate correlations, both linear and non-linear.

To enable analysis of large data sets, the program was implemented using the 'SNOW' and 'pbdMPI' libraries. The SNOW library [30] allows a cluster to be called in a master-slave configuration. While helpful for simple, repetitive computations, SNOW does not allow for direct communications between processes. The 'pbdMPI' library [4] provides a complete R interface for the Message Passing Interface, making it the more effective approach for our program.

Computations were carried out on the 86 node computing cluster tara located in the UMBC High Performance Computing Facility. Each node features two quad core Intel Nehalem X5550 processors (2.66 GHz, 8192 kB cache) and 24 GB of memory. Attached to the cluster is 160 TB of central storage. More information about the cluster and its use can be found at the webpage `www.umbc.edu/hpcf`. Documentation on running pbdMPI programs on tara is available in [23].

Since MD simulation data is a time series, is has some intrepretations that other data may not have. In particular, motional dependencies between protein sites may not be instantaneous. For example, there may be a slight delay between a molecule binding to a protein, and the change in motion in another part of the protein. To search for time delayed correlations, the observations must be shifted relative to one another. Simply by comparing response values with values of the predictors further forward in time, we can search for the same correlations while including a time lag. Our program is capable of a finding maximum correlations for all lags up to a specified value, and of finding correlations for a single, fixed time lag.

The primary output of the code is a scaled matrix Θ that describes the strength of correlation between the variables over over a span of time. This matrix can be represented as a false color plot; this representation makes interpretation fast and intuitive. Some challenges remain; the program's I/O and plotting operations are both serial. While some parallel I/O implementations such as 'pbdNCDF4' [20] exist, they are not commonly used.

2.3 Simulations

We performed a simulation study on smaller scaled data to assess the effectiveness of the implementation. We generated a sample data set with known correlations between each variables. The simulated data had $n = 300$ observations and $p = 100$ predictors. The data set was obtained as follows. We first generated the vector Y from the normal distribution with mean 0 and standard deviation 4. We then formed $\Gamma = (\Gamma_1, \Gamma_2, \Gamma_3)$, where

$$\begin{aligned}
\Gamma_1^T &= (u_1, ..., u_{20}, \mathbf{0}_{80}),\\
\Gamma_2^T &= (\mathbf{0}_{20}, v_1, ..., v_{20}, \mathbf{0}_{60}),\\
\Gamma_3^T &= (\mathbf{0}_{40}, w_1, ..., w_{20}, \mathbf{0}_{40}).
\end{aligned}$$

The terms u_i's, v_i's, w_i's are uniformly distriputed between 0.5 and 1 for all $1 \le i \le 10$ and uniformly distributed between -1 and -0.5 for all $11 \le i \le 20$. With this Γ, we then form the final simulated data matrix $\mathbf{X} = (Y, Y^2, Y\cos(6\pi Y))\Gamma^T + \epsilon$ where ϵ is a $n \times p$ matrix of observations from the standard normal distribution. In this data set, the columns, which are based on based on $\Gamma_1, \Gamma_2, \Gamma_3$, are correlated among themselves. In addition, there are correlations between the sets. For example, terms of Γ_1 and Γ_2 are quadratically correlated. These are the patterns the methodology is intended to show.

Once the data set was generated, we proceed as explained in Section 3.1. Each variable is used as a response and the remaining is used as predictors. We form the basis function using the response. Different basis functions were considered to help uncover the correlation among the variables. Among these were $\mathbf{f}_Y = Y, \mathbf{f}_Y = (Y, Y^2, Y^3), \mathbf{f}_Y = (Y, Y^2)$, and $\mathbf{f}_Y = (Y, Y^2, Y\cos(6\pi Y))$, and the basis functions corresponding to the correlation plots.

Figures 2.1–2.2 were obtained using the linear and a cubic polynomial bases $\mathbf{f}_Y = Y$ and $\mathbf{f}_Y = (Y, Y^2, Y^3)$, respectively. In Figure 2.1, there is a clear linear correlation shown

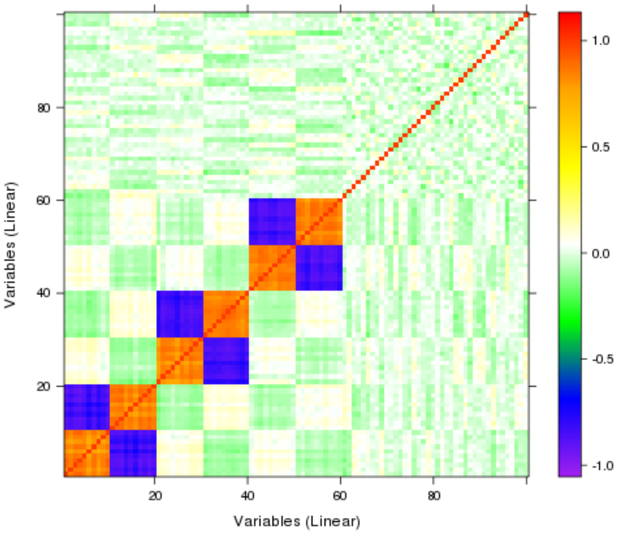


Figure 2.1: Correlation plot of simulated data with $\mathbf{f}_Y = Y$.

between variables, as shown by the blocks of stronger correlation along the diagonal. These correlations were expected due to the arrangement of Γ. The first 10 predictors are positively linearly correlated with each other and negatively linearly correlated with the next 10 predictors. This pattern continues with predictors 21 to 30 being positively linearly correlated with each other and negatively linearly correlated with predictors 31 to 40. The pattern continues in the next block. This is exactly what was expected due to the deliberate assignment of Γ.

In Figure 2.2, there is an association between variables 20 to 40 and variables 0 to 20 which was not seen when only looking at the linear correlation plot. Additionally, a very faint association can be seen between variables 20 to 40 and variables 40 to 60. These additional associations appear because of the existance of the y^2 term in the basis function here.

It appears that the adequate choice of the basis function is crucial is detecting the correlations among the variables. Other bases such as spline basis and trigonometrical basis will be explored in the future.

2.4 Parallel PFC Performance

A performance study was run using $n = 100$ observations and $p = 531$ predictors. Table 2.1 summarizes the program's performance on various numbers of processes.

Figure 2.3 (a) depicts a near linear speedup, substantially dropping off starting at 256 processes. Similarly, Figure 2.3 (b) shows that efficiency drops down below 50% only for 256 and 512 processes. This study demonstrates the effectiveness of the parallel implementation and its scalability.

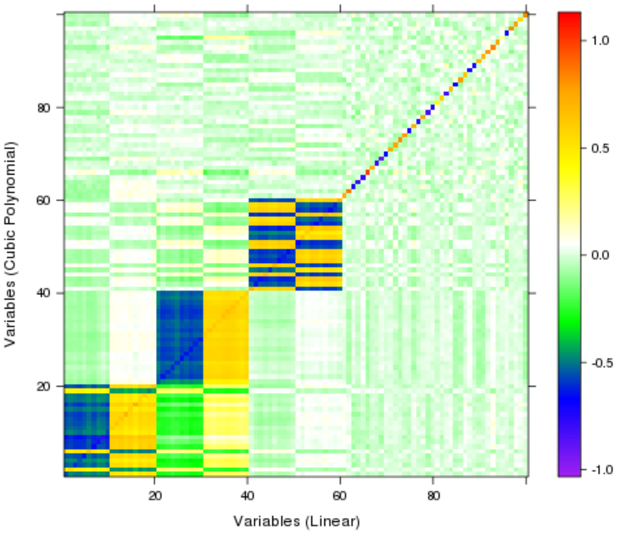


Figure 2.2: Correlation plot of simulated data with $\mathbf{f}_Y = (Y, Y^2, Y^3)$.

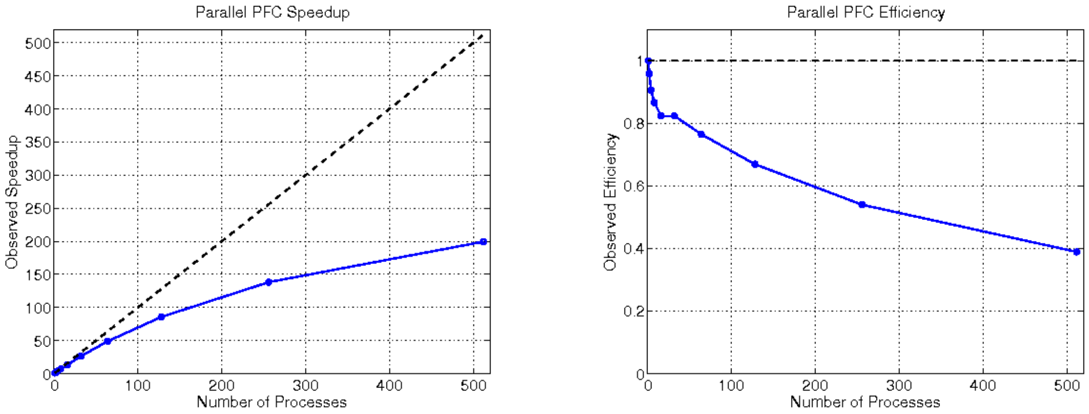
Table 2.1: Scalability study with $n = 100$ and $p = 531$. (a) Observed wall clock time in HH:MM:SS, (b) observed speedup, (c) observed efficiency.

No. Processes	(a) wall time	(b) speedup	(c) efficiency
1	02:22:52	1.00	1.00
2	01:14:28	1.92	0.96
4	00:39:24	3.63	0.91
8	00:20:36	6.94	0.87
16	00:10:50	13.19	0.82
32	00:05:25	26.38	0.82
64	00:02:55	48.98	0.77
128	00:01:40	85.72	0.67
256	00:01:02	138.26	0.54
512	00:00:43	199.35	0.39

3 Application to the Allosteric Data Set

3.1 PFC in the Context of Protein Dynamics

To describe the use of PFC in the context of protein dynamics, let $V_1, ..., V_p$ be p configurational variables. Consider V_k and its relationship to the remaining $p - 1$ variables. Let $\mathbf{X}$ be the $(p-1)$-vector of all variables except V_k, and let $Y = V_k$. The term Γ captures the statistical dependency between $\mathbf{X}$ and Y be it linear or nonlinear. For example, suppose



(a) (b)

Figure 2.3: Plots of (a) speedup and (b) efficiency with $n = 100$ and $p = 531$.

that $\nu_Y = Z$ and $\Gamma = (\gamma_1, 0, ..., 0)^T$. Then the expected value of V_1, $E(V_1) = \gamma_1 V_k$ and γ_1 is proportional to the correlation between V_1 and V_k. Similarly, if $\nu_Y = (V_k, (V_k)2)^T$ and Γ has two columns given by $(\gamma_{11}, 0, ..., 0)^T$ and $(\gamma_{22}, 0, ..., 0)^T$, then $V_1 = \gamma_{11} V_k + \gamma_{22}(V_k)^2$. Thus, the statistical dependency between V_1 and V_k can be readily established. The proper choice of basis function determines the adequacy of the method in capturing the relationship between $\mathbf{X}$ and Y. For example, if ν_Y is a quadratic function of Y and a linear approximation is used, the method may fail to capture the proper dependency.

With the atoms on the same measurement scale, the magnitude of the row elements of Γ indicates the strength of the correlation between individual predictors and the response. And thus, it suffices to estimate Γ and evaluate these dependencies.

Given the data with p predictors, each predictor is used once as the response, and Γ is estimated. The process is therefore run p times. Let $\widehat{\Gamma}_k$ be the estimate of Γ when the k-th predictor is used. We then form the $p \times p$ matrix $\Theta = (\widehat{\Gamma}_1, \widehat{\Gamma}_2, \cdots, \widehat{\Gamma}_p)$ that is scaled and plotted. When the dimension d of Γ is one and $\mathbf{f}_Y = Y$, then Θ can be obtained as the usual correlation matrix.

3.2 Molecular Dynamic Simulations and the Data Set

To obtain data for allosteric MD simulations, 400 ns simulations were run, with NS5B in explicit solvent with and without inhibitor VGI (PubChem ID 4177750) bound to the thumb domain. Covariance analysis was used to determine patterns of correlated motion. Distinct patterns of correlation were evident in the free enzyme that were eliminated when ligand is bound. This results in accord with other studies that suggest the function of related viral polymerases is mediated by specific long range correlations. Moreover, this observation supports the hypothesis that allosteric inhibitors such as VGI inhibit the enzyme by disrupting specific correlated motions. The current study strives to determine whether additional correlations exist that were not revealed using simple covariance.

Table 3.1: Molecular dynamics data structure.

Time Steps	$x_1, x_2, \ldots, x_{531}$	$y_1, y_2, \ldots, y_{531}$	$z_1,\ z_2, \ldots, z_{531}$
t_1	$6.22, 14.17, \ldots, 17.67$	$-10.52, -14.12, \ldots, -14.84$	$18.1\ 8, 14.69, \ldots, 15.99$
t_2	$10.07, 18.42, \ldots, 21.59$	$-10.26, -17.09, \ldots, -16.94$	$18.\ 06, 12.96, \ldots, 10.93$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_{100}	$12.82, 20.47, \ldots, 18.47$	$-10.44, -15.09, \ldots, -12.60$	$15.36, 7.76, \ldots, 9.95$

3.3 The Data Set

The MD simulation data is best represented as a large matrix. Each atom, or each coordinate of each atom is a variable of the data set, and each time step of the simulation is an observation furnishing a value for each variable. Every variable forms a column of the matrix, and the set of observations at each time step form the rows of the matrix. Typically, a simulation consists of 100s of atoms, and between 100 and 10000 time steps. In this case, the simulation has 531 atoms, and each coordinate is considered to be a predictor; thus the data matrix has $p = 1593$ columns. The entire simulation consists of 3000 time steps, but we will consider only the first $n = 100$ steps. Table 3.1 illustrates the data structure.

3.4 Application to MD Data Set

Because MD simulations simulate the motion of physical bodies, it should be noted that atoms repsonse to changes in ther surrondings may not be instantaneous. To address this issue from a data analysis perspective, we have implemented a method to systematically compare observations of the response variable with observations of the predictors at previous time steps. The PFC methodology is unaffected, it is simply applied to data that is shifted in this way.

Consider the kth column of the data matrix $\mathbf{X}$ to be the response, which is denoted by $V_k = [v_{1\,k}, v_{2\,k}, \ldots, v_{n\,k}]^T$. If we wish to know the correlations between the response V_k and the predictor values t time steps ago, then we shift each entry in V_k column up by one row, so that the value of v_i is now in the location of V_{i-t}. Before applying PFC, we remove the last $n - t$ rows from all columns other than V_k, these values have no corresponding observations in V_k, just as the first t observations of V_k have no corresponding predictors.

This shifting is carried out for each t up to some specified maximum t_f. For each predictor-response combination, the correlation value with the greatest magnitude over all time shifts is saved, so that if predictors produce effects with different lags, they will all be captured. Thus, the plot produced will represent the maximum correlations present over all time shifts up to t_f. This methodology has the unfortunate effect of acumulating noise; by taking the maximum in this way, it may appear that there are more meaningful relationships among the data than there actually is.

This shifting methodology was applied to our molecular dynamics simulation data, and we show the result of the first 531 variables in Figure 3.1. This correlation plot uses a basis

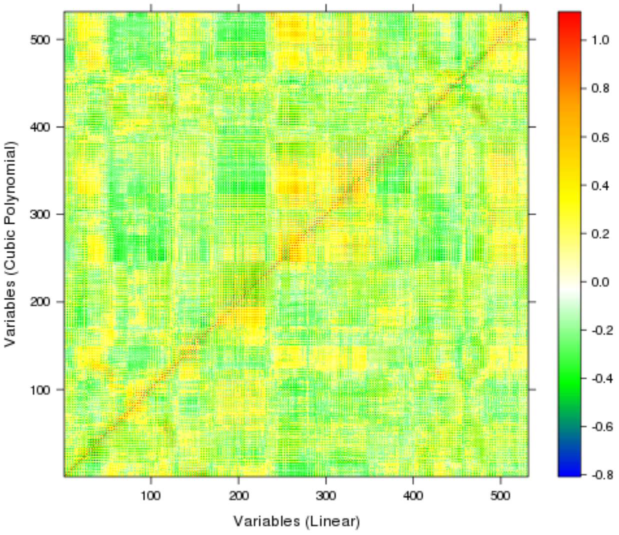


Figure 3.1: Correlation plot of MD simulation data with $\mathbf{f}_Y = Y^3$.

function of $\mathbf{f}_Y = Y^3$.

4 Conclusions

PFC methodology was found to be quite amenable to parallel computation. A wrapper, written using the 'pbdMPI' R library used existing R code and provided significant speedup. The resulting program and allowed both instantaneous and delayed correlations to be uncovered, making PFC methodology useful in the context of MD simulation analysis.

PFC demonstrated these advantages on data with known correlations, successfully identifying linear and polynomial associations, supplanting some popular regression methods (e.g., PCA). The final program also produced correlations maps that provided an effective qualatative tool for data analysis.

These correlation maps are useful for analyzing MD simulation data, but previously have been available only for linear correlations. Producing these maps for higher-order correlations represents a significant improvement in our ability to quickly analyze MD simulation data for allosteric sites. Because the code may be run in parallel on many different size systems, the program also represents an improvement in our ability to analyze even larger data sets, consisting of even 1000s of atoms.

While our program repersents an improved method for analyzing MD simulation data, it has significant drawbacks. Data I/O and plotting operations are bottlenecks, and increase total program runtime. Implementing parallel file I/O is possible using existing R libraries, and would avoid one of these bottlenecks.

The biochemical implications of these improvements are all the implications that an improved understanding of allostery itself provides; the oppertunity for better understanding of how allosteric sites affect protein structure and function. This would present oppertunities for novel drug design using these sites.

Acknowledgments

These results were obtained as part of the REU Site: Interdisciplinary Program in High Performance Computing (www.umbc.edu/hpcreu) in the Department of Mathematics and Statistics at the University of Maryland, Baltimore County (UMBC) in Summer 2013. This program is funded jointly by the National Science Foundation and the National Security Agency (NSF grant no. DMS–1156976), with additional support from UMBC, the Department of Mathematics and Statistics, the Center for Interdisciplinary Research and Consulting (CIRC), and the UMBC High Performance Computing Facility (HPCF). HPCF (www.umbc.edu/hpcf) is supported by the National Science Foundation through the MRI program (grant nos. CNS–0821258 and CNS–1228778) and the SCREMS program (grant no. DMS–0821311), with additional substantial support from UMBC. Co-author Jesse D. Smith was supported, in part, by the UMBC National Security Agency (NSA) Scholars Program though a contract with the NSA. Graduate RA Andrew M. Raim was supported by UMBC as HPCF RA.

References

[1] C.F. Abrams and E. Vanden-Eijnden. Large-scale conformational sampling of proteins using temperature-accelerated molecular dynamics. *Proceedings of the National Academy of Sciences of the United States of America*, 107(11):4961—4966, 2010.

[2] Kofi P. Adragni and Andrew Raim. LDR: An R software package for likelihood-based sufficient dimension reduction. Submitted.

[3] N. Boekelheide, R. Salomon-Ferrer, and T.F. Miller. Dynamics and dissipation in enzyme catalysis. *Proceedings of the National Academy of Sciences of the United States of America*, 108(39):16159—16163, 2011.

[4] Wei-Chen Chen, George Ostrouchov, Drew Schmidt, Pragneshkumar Patel, and Hao Yu. *A Quick Guide for the pbdMPI Package*, 2012. R Vignette, URL http://cran.r-project.org/package=pbdMPI.

[5] Y.H. Chen, K. Reilly, and Y.C. Chang. Evolutionarily conserved allosteric network in the cys loop family of ligand-gated ion channels revealed by statistical covariance analyses. *Journal of Biological Chemistry*, 281(26):18184—18192, 2006.

[6] C.N. Chi et al. Reassessing a sparse energetic network within a single protein domain. *Proceedings of the National Academy of Sciences of the United States of America*, 105(12):4679—84, 2008.

[7] R.R. Coifman and S. Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis: Special issue on Diffusion Maps and Wavelets*, 21:5—30, 2006.

[8] R. Dennis Cook. Fisher lecture: Dimension reduction in regression. *Statistical Science*, 22(1):1—26, 2007.

[9] B.K. Dey. Optimal non-linear dimension reduction scheme for classical molecular dynamics. *Journal of Mathematical Chemistry*, 49(9):2032—2052, 2011.

[10] A.L. Ferguson et al. Integrating diffusion maps with umbrella sampling: Application to alanine dipeptide. *Journal of Chemical Physics*, 134(13), 2011.

[11] V.J. Hilser. An ensemble view of allostery. *Science*, 327(5966):653—654, 2010.

[12] B. Jana et al. Dynamic coupling between the lid and nmp domain motions in the catalytic conversion of atp and amp to adp by adenylate kinase. *Journal of Chemical Physics*, 134(3):035101, 2011.

[13] R. Kadirvelraj et al. Role of packing defects in the evolution of allostery and induced fit in human udp-glucose dehydrogenase. *Biochemistry*, 50(25):5780—9, 2011.

[14] H.X. Kondo et al. Free-energy landscapes of protein domain movements upon ligand binding. *Journal of Physical Chemistry B*, 115(23):7629—7636, 2011.

[15] M. Kurylowicz, C.H. Yu, and R. Pomes. Systematic study of anharmonic features in a principal component analysis of gramicidin a. *Biophysical Journal*, 98(3):386—395, 2010.

[16] J. Lee et al. Surface sites for engineering allosteric control in proteins. *Science*, 322(5900):438—442, 2008.

[17] S.W. Lockless and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295—299, 1999.

[18] C. McClendon et al. Quantifying correlations between allosteric sites in thermodynamic ensembles. *Journal of Chemical Theory and Computation*, 5:2486—2502, 2009.

[19] K. Okazaki and S. Takada. Dynamic energy landscape view of coupled binding and protein conformational change: Induced-fit versus population-shift mechanisms. *Proceedings of the National Academy of Sciences of the United States of America*, 105(32):11182—7, 2008.

[20] G. Ostrouchov, W.-C. Chen, D. Schmidt, and P. Patel. Programming with big data in R, 2012. `http://r-pbd.org`.

[21] A. Panjkovich and X. Daura. Assessing the structural conservation of protein pockets to study functional and allosteric sites: Implications for drug discovery. *BMC Structural Biology*, 10:9, 2010.

[22] S.N. Pieniazek, M.M. Hingorani, and D.L. Beveridge. Dynamical allosterism in the mechanism of action of dna mismatch repair protein muts. *Biophysical Journal*, 101(7):1730—1739, 2011.

[23] Andrew M. Raim. Introduction to distributed computing with pbdR at the UMBC High Performance Computing Facility. Technical Report HPCF–2013–2, UMBC High Performance Computing Facility, University of Maryland, Baltimore County, 2013.

[24] A. Ramanathan et al. Discovering conformational sub-states relevant to protein function. *Public Library of Science One*, 6(1):e15827, 2011.

[25] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(22):2323—2326, 2000.

[26] B. Scholkopf, A. Smola, and K. Muller. Kernel principal component analysis. *Neural Computation*, 10(5):1299—1319, 1998.

[27] H. Stamati, C. Clementi, and L.E. Kavraki. Application of nonlinear dimensionality reduction to characterize the conformational landscape of small peptides. *Proteins: Structure, Function and Bioinformatics*, 78(2):223—35, 2010.

[28] J.F. Swain and L.M. Gierasch. The changing landscape of protein allostery. *Current Opinion in Structural Biology*, 16(1):102—8, 2006.

[29] J.B. Tenenbaum, D.S. V., and J.C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319—2323, 2000.

[30] Luke Tierney, A. J. Rossini, Na Li, and H. Sevcikova. *snow: Simple Network of Workstations*, 2011. R package version 0.3-8.

[31] S. Wang et al. Molecular dynamics analysis reveals structural insights into mechanism of nicotine n-demethylation catalyzed by tobacco cytochrome p450 mono-oxygenase. *Plos One*, 6(8):e23342, 2011.

[32] W.J. Zheng, B.R. Brooks, and D. Thirumalai. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. *Proceedings of the National Academy of Sciences of the United States of America*, 103(20):7664—7669, 2006.

[33] P.I. Zhuravlev and G.A. Papoian. Protein functional landscapes, dynamics, allostery: A tortuous path towards a universal theoretical framework. *Quarterly Reviews of Biophysics*, 43(3):295—332, 2010.

[34] J. Zimmermann et al. Molecular description of flexibility in an antibody combining site. *The Journal of Physical Chemistry B*, 114(21):7359—70, 2010.