Finding Important Protein Motions in Solution Through Linear Dimensionality Reduction

Krzysztof Borowski*and Forbes Burkowski

Universiy of Waterloo, Waterloo, Canada

ABSTRACT

Automating protein structure analysis is useful to understand their function, especially as structure data grows. Developing methods of elucidating protein motions from the multitude of data currently available, and doing it in a way that allows non-experts to easily perceive protein movement can lead to heightened levels of understanding and hypothesis generation. We apply principal component analysis to examine major modes of motion for protein conformations. We show that a few principal components of the conformation matrix can capture the majority of the motions, and present a visual depiction of one mode of calmodulin.

Keywords: protein flexibility, PCA, SVD, NMR, calmodulin

1 BACKGROUND

As datasets from wet-lab experiments on protein structure grow in size and number, the interpretation of these results becomes difficult. X-ray Crystallography experiments are able produce a large amount of data about crystallized protein structure[1]. For example, there are over 150 conformations of the protein HIV-1 Protease available on the Internet. Due to its pharmaceutical importance (as a drug target in HIV research), it is a valuable endeavor to understand the main modes of movement of this protein. Principal Component Analysis (PCA) has been applied to HIV-1 Protease before, with results showing a similar mobility to movement determined experimentally. It has also been used in conjunction with singular value decomposition (SVD) to that same end with other protein sets[2, 3].

Nuclear Magnetic Resonance (NMR) is another method of attaining structural protein data[4]. Here, the structure of protein is taken from protein in solution, which allows for more variability in the data collected. Consequently NMR can create even more data than X-ray Crystallography. Many solution-NMR experiments have resulted in multiple models, each presenting a different conformation for the protein in question. With many such models within a file, and with many such files, it becomes imperative to develop automated methods of discerning major protein motions in a cost efficient way. PCA is a possible solution to this issue.

Other methods, such as normal mode analysis (NMA), are also applicable to finding protein motions. NMA considers harmonic motions and involves energy minimization. This makes NMA a more biologically relevant type of analysis when compared with PCA. However, NMA suffers from computationally demanding steps, and the results of PCA and NMA can be comparable[5].

In this paper, we implement an automated PCA based strategy for finding principal modes of motion of a protein set[2]. We develop Python code for use with UCSF Chimera[6] to present possible principal modes of motion for a given set of protein conformations, and we apply it to the protein calmodulin.

1.1 Principal Component Analysis

PCA is a dimensionality reduction technique which allows for high dimensional variables to be embedded into a low dimensional space where they are uncorrelated. These reductions in the lower dimension, known as principal components, are linear combinations of the original high dimensional variables. Each principal component is constructed to cover as much variance in the original data as possible; this allows for the mapping from high to low dimensional space to maintain information of the high dimensional system in very few principal components. It is possible to obtain as many principal components as there are original variables, but PCA is intended as a way of finding the smallest number of uncorrelated principal components which cover a large percentage of total variation in the original data. Even though this means some information is lost in the process, the trade-off between representing information through less variables may be beneficial in many applications. One such area is the study of molecular motions.

*Corresponding author. E-mail: kborowsk@uwaterloo.ca

1.2 Singular Value Decomposition

Singular Value Decomposition (SVD) is a common method in linear algebra in which a matrix is factored into a diagonal matrix, and two eigenvector matrices[7]. The SVD method allows the creation of a pseudoinverse for matrices which are close to being singular, but it also allows for extraction of important information from a matrix in a way similar to PCA[3]. It can be an efficient way of obtaining principal components. The SVD of an $m \times n$ matrix A is defined as:

$$A = UDV^T \tag{1}$$

where U and V are orthogonal left and right eigenvector sets of A respectively, and D is a nonnegative diagonal matrix with elements being the singular values of A. For our purposes, the trace of D is the total variance in A, while the square of each singular value represents the variance of the data in A along the corresponding vector in U[2].

2 METHODS: APPLYING PRINCIPAL COMPONENT ANALYSIS TO PROTEIN CONFORMATION

We will use SVD to gain the principal components of a set of molecular structures, thereby getting the principal modes of movement[2]. First, we acquire a set of l structures of the same protein. Each atom in the protein model has Cartesian coordinates (x_i, y_i, z_i). Initially, we must complete a rigid least squares fit on all models with respect to one of the structures, which removes translational and rotational degrees of freedom[8].

We represent the entire protein model by creating a conformational vector of the form

$$m_i = (\Delta x_1, \Delta y_1, \Delta z_1, \Delta x_2, \Delta y_2, \Delta z_2, ..., \Delta x_n, \Delta y_n, \Delta z_n) \tag{2}$$

which is a concatenation of the displacement of n atom coordinates from the arithmetic mean along the Cartesian axes[3]. Since the structures are superimposed after the least squares fit, this requires finding the average conformational vector and subtracting it from the conformational vector of the given protein model. Using all the atoms in the model is possible and may be desired for some experiments, but increases calculation time. We limit the conformation vectors to alpha carbons on each residue in the model only, which allows an examination of backbone motion specifically. Others have shown that the principal components are useful when considering atoms in binding sites alone[3].

Once each conformation vector m for all l models are acquired, the matrix A is created by column-wise concatenation of the l conformation vectors:

$$A = [m_1|m_2|...|m_l], A \in \mathbb{R}^{3n \times l} \tag{3}$$

where each m is the conformation vector representation of a model defined above.

After creating the matrix A, we proceed with the SVD of A as presented by Equation (1). The matrix of left eigenvectors U has columns u_i, and it is these columns which are the principal components of A. Each u_i shows the *mode* of motion of the atoms which were used in building the conformational vectors. One can think of u_i as a collection of direction vectors for n atoms, each representing their *mode* of motion at time point i of l[3]. Since this is a PCA approach, each column vector u_i will hold the general directions where most of the variability in atom position occurs. For example, the first three elements of u_1 would define the vector of the first principal component of motion of the first atom under consideration.

The right eigenvectors v_i found in V, are projections of A on the u_i vectors in the U matrix. As the A in our case is a protein conformation matrix, the elements of v_i are the locations of each atom along its principal component u_i. They can also be used to discern preferred protein conformations[3].

3 RESULTS OF APPLICATION TO CALMODULIN DATA

To test the methodology presented above, we applied the SVD based PCA to models of calmodulin created by Zhang et al. (1995)[9]. Calmodulin is a protein involved in a variety of cellular functions, including protein synthesis, gene regulation, cell motility, and cellular secretion[10]. It is composed of two main domains tethered with a loop which allows for fluctuation in protein structure for target binding[11]. The large amount of displacement between conformations of calmodulin makes it an excellent example of the effectiveness of this method. The experimental data from Zhang et al. (1995) has 30 models discerned by NMR.

In Figure 1, we show the amount of variance each principal component exhibits on total variance in the original dataset. The largest singular value shows that the first principal component accounts of 29% of the total variance. To achieve 90%, 11 principal components must be taken into account.

In Figure 2, we visualize the mode described by the first principal component for each alpha carbon in calmodulin. The cylinders indicate the vector of the most important mode along which an atom will move, with the rounded end pointing in the direction of movement.

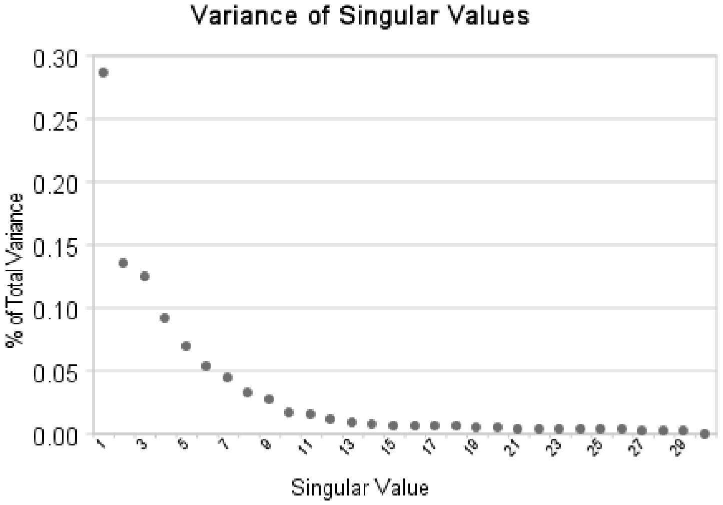

Figure 1.
Amount (%) of total variance of matrix A explained by specific singular values of A.

4 DISCUSSION

Previous work in this area has used the PCA to compare bound and unbound structures of ligand binding protein, and in the majority of cases, has been limited to data generated by X-ray Crystallography studies of proteins and their ligands[2]. Such research allowed for the principal modes of movement during binding to be understood. However, to our knowledge, NMR experiment data on unbound motions of a single protein have not yet been examined with this approach. The results suggest that the method is useful to discover transient motions in protein in solution by using multiple NMR models. By transient movement, we meant movement that is unrelated to binding specifically, but constitutes normal fluctuations in structure while a protein is in solution.

As Figure 1 shows, a majority of the modes of motion can be captured through a number of principal components much smaller than that of the models used in the analysis.

Even with the implied freedom of an unbound structure, the PCA method elucidated modes of movement similar to movement known as biologically relevant for calmodulin[11]. In Figure 2, we see the modes of one of the domains pointing in a direction where the folding movement may initially occur. Consecutive principal components, while not shown, suggest the same directions for the domain, while the 'anchored' domain shows very little mode activity at any point in time (according to the right eigenvectors). All of the atom modes seem to point in similar directions on the domain whose coordinate changes explain the movement of calmodulin. This is not visible in the 'anchored' domain: the principal components do not seem to point in similar directions, nor are they scaled by the right eigenvectors to the extent that the other domain is. This implies that it would be improbable for this domain to experience major movement in any particular direction, relative to the coordinate system of the original data. It is important to note that this is because the principal components are calculated relative to the protein structure coordinates of the original data: the data was pre-fitted around one of the domains, making it appear as though it is immobile to the PCA analysis. This anchoring provides a clearer view of the mobility of the other domain, but should not lead the reader to think the domain is somehow immobile in reality.

Overall, the principal motions of calmodulin in solution are based around the two domains coming closer and moving apart by the folding and extending activity of their tether. Based on the data from Zhang et al. (1995)[9], this 'flopping' action occurs along a plane, which is why most of the modes seen in this work are close to being parallel directions (that is, if we were to translate them all to the axis along which this movement of calmodulin occurs, they would be as close to being on the plane itself).

5 CONCLUSION

Both the range of variability description found within the initial principal components, as well as the visualized modes show that the method can be used to better understand protein movement in solution. We used data on calmodulin in solution, whose 'domains on a tether' structure allows for various conformational possibilities in solution which fluctuate along a limited space of substates; they cannot behave without submitting physical and energy constraints. This is supported by the PCA analysis of the dataset, which suggest that major modes of movement

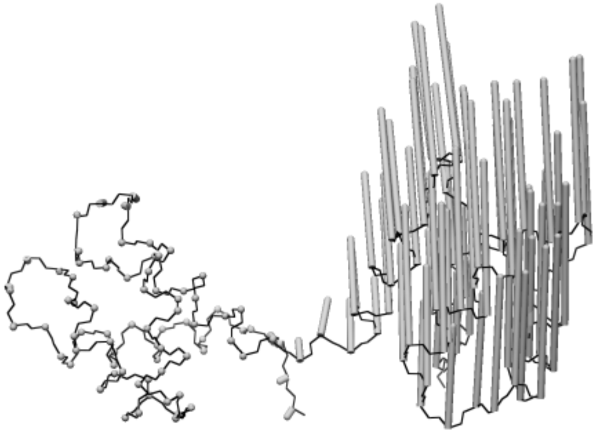

Figure 2.
The backbone of calmodulin with the first principal components, or modes of motion (left eigenvectors of A) indicated by the thin cylinders. The spherical end of the cylinder represents the positive direction of motion. The principal components are scaled by the right eigenvectors of A.

restrict motion of the protein along a specific path. We also show that the majority of said modes can be described with a few principal components found from the original data.

In the future, larger datasets should be used in order to find more modes. The limitation of structures derived from NMR experiments is that they are all found at the same experimental conditions. Combining multiple NMR experiment files to create input to PCA analysis may elucidate new modes of motion not found with a dataset from only one experiment. As well, this work only examined backbone movement; others have completed the PCA on all atoms, including those of residue sidechains[2]. Including sidechains in the analysis may shed light on more detailed movement of a protein in solution.

We used UCSF Chimera[6] to visualize the modes of calmodulin in a static fashion, but programs such as Chimera can be used to animate proteins undergoing changes along their principal components. Automated animation of structures about their main modes of movement would allow for quick analysis of probable protein mobility by non-experts.

Finally, PCA is a linear method of dimensionality reduction. Non-linear methods may improve results in dimensionality reduction applications. Analyzing main modes of motion may result in different outcomes if non-linear methods are used, such as locally linear embedding[12], Laplacian eigenmaps[13], or even kernel methods[14].

References

[1] M. S. Smyth and J. H. J. Martin, “X-ray crystallography,” *Molecular Pathology* **53**(1), p. 8, 2000.

[2] M. L. Teodoro, G. N. P. Jr, and L. E. Kavraki, “A dimensionality reduction approach to modeling protein flexibility,” in *Proceedings of the sixth annual international conference on Computational biology*, pp. 299–308, 2002.

[3] T. D. Romo, J. B. Clarage, D. C. Sorensen, and G. N. P. Jr, “Automatic identification of discrete substates in proteins: singular value decomposition analysis of time-averaged crystallographic refinements,” *Proteins: Structure, Function, and Bioinformatics* **22**(4), pp. 311–321, 1995.

[4] K. Wuthrich, “NMR of proteins and nucleic acids,” *The George Fisher Baker non-resident lectureship in chemistry at Cornell Unversity (USA)* **1**, 1986.

[5] S. Hayward and B. L. Groot, “Normal modes and essential dynamics,” in *Molecular Modeling of Proteins*, A. Kukol, ed., pp. 89–106, Humana Press, 2008.

[6] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, “UCSF chimera–a visualization system for exploratory research and analysis,” *Journal of computational chemistry* **25**(13), pp. 1605–1612, 2004.

[7] G. H. Golub and C. F. V. Loan, *Matrix computations*, Johns Hopkins Univ Pr, 1996.

[8] W. Kabsch, “A discussion of the solution for the best rotation to relate two sets of vectors,” *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* **34**(5), pp. 827–828, 1978.

[9] M. Zhang, T. Tanaka, and M. Ikura, “Calcium-induced conformational transition revealed by the solution structure of apo calmodulin,” *Nature Structural Biology* **2**, pp. 758–767, Sept. 1995.

[10] W. Y. Cheung, “Calmodulin plays a pivotal role in cellular regulation,” *Science (New York, NY)* **207**(4426), p. 19, 1980.

[11] Y. S. Babu, C. E. Bugg, and W. J. Cook, “Structure of calmodulin refined at 2.2 a resolution,” *Journal of molecular biology* **204**(1), p. 191, 1988.

[12] S. T. Roweis and L. K. Saul, “Nonlinear dimensionality reduction by locally linear embedding,” *Science* **290**(5500), p. 2323, 2000.

[13] M. Belkin and P. Niyogi, “Laplacian eigenmaps for dimensionality reduction and data representation,” *Neural computation* **15**(6), pp. 1373–1396, 2003.

[14] K. Q. Weinberger, F. Sha, and L. K. Saul, “Learning a kernel matrix for nonlinear dimensionality reduction,” in *Proceedings of the twenty-first international conference on Machine learning*, 2004.