Application of Linear Algebra to Biological Molecules: PCA Analysis of holo-Myoglobin

Christopher Bruner and Dr. Jen-Mei Chang

College of Natural Sciences and Mathematics, CSU Long Beach, 1250 Bellflower Blvd. Long Beach, CA 90840

Introduction:

Principal Component Analysis (PCA) continues to be one of the most valuable tools used in the science; it's derived from linear algebra and it can take a noisy, seemingly random data set and reveal hidden structure that may not be easily apparent. Proteins and other biological molecules possess interesting dynamics, but they occur on the millisecond or less timescale.

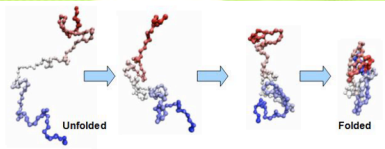

Figure 1: Simulations allow us to see dynamics normally unobservable, but what are the interesting dynamics? That's what PCA wishes to answer using linear algebra. With PCA, we can determine what motions are crucial such as in the folding of a protein.

However, molecular modeling allows us to utilize computers to characterize these motions. Myogolobin (Mb) is a protein involved with the circulatory system, delivering oxygen (like Hemoglobin). E. Papaleo, et al., ran many simulations on holo-Mb (one variant of Mb) and used PCA as their primary analytical tool in hopes of extending the research to other variants of Mb.

Methods:

After collecting as much data on a system as possible, we take all the of vectors and collect them into a set, X. Now PCA asks the question, is there some basis that is a linear combination of the original basis that can more efficiently express the data we have?

Let **X** be an *mxn* matrix, and define a matrix **P** which transforms **X** into another *mxn* matrix **Y** (change of basis):

$$1. \quad \mathbf{PX}=\mathbf{Y}$$

where the rows of **P**, $[p_1,\ldots,p_m]$, represents a set of new basis vectors for expressing **X**. In order to get the best **P** (i.e. best way to re-express **X**), we need know what features **Y** should display.

First, we want to minimize the noise while maximizing signal to get a high signal-to-noise ratio (fig. 2), to produce a best fit line:

$$2. \quad SNR=\sigma^2_{signal}/\sigma^2_{noise}$$

For higher dimensions, we need the covariance matrix to determine the degree of linearity between two variables. Let's define:

$$a = [a_1\ a_2\ \ldots\ a_n]$$

$$b = [b_1\ b_2\ \ldots\ b_n]$$

where a_i and b_i are row vectors. We wish to measure the degree of the linearity between these two data sets such that the covariance between **a** and **b** is:

$$3. \quad \sigma^2_{ab} = [1/(n-1)]ab^T$$

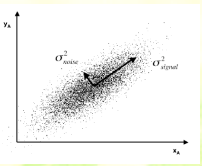

Figure 2: Example Data for 2-D case. 1st PC maximizes variance, 2nd PC captures variance perpendicular to the 1st PC.

(n-1 is the normalization factor). High (small) values represent high (low) redundancy within the data set. For our original data set **X**, we can represent the covariance matrix as $\mathbf{C_X}$ as:

$$4. \quad \mathbf{C_X}=[1/(n-1)]\mathbf{XX^T}$$

$\mathbf{C_x}$ has the following properties:

1. The diagonal terms are the variance of a particular type, where large (small) values correspond to interesting dynamics (noise)
2. The off-diagonal terms are the covariance between measurement types with large (small) values corresponding to high (low) redundancy

We then optimize $\mathbf{C_x}$ into some matrix, $\mathbf{C_Y}$. Again, our goal is to minimize redundancy (covariance) and maximize signal (variance), to do so, we diagonalize $\mathbf{C_Y}$ assuming **P** is an orthonormal matrix.

Now, to find the PCs, we need to define a matrix **P** for **Y**=**PX** such that

$$6. \quad \mathbf{C_Y} = [1/(n-1)]\mathbf{YY^T}$$

The trick is to substitute **PX** into **Y** and $\mathbf{Y^T}$, do some manipulation and the resulting algebra produces

$$7. \quad \mathbf{C_Y}=[1/(n-1)]\mathbf{PAP^T}$$

where **A** is $\mathbf{XX^T}$, a symmetric matrix. For a symmetric matrix, we know that $\mathbf{A}=\mathbf{EDE^T}$, **E** being eigenvectors of **A** and **D** being a diagonal matrix. The major trick to PCA is that we select a matrix **P** such that the rows, p_i, are eigenvectors of $\mathbf{XX^T}$. We then select $\mathbf{P}=\mathbf{E^T}$ and because the inverse of an orthogonal matrix is its transpose, we know that $\mathbf{P^{-1}}=\mathbf{P^T}$. We can then substitute for **A** into $\mathbf{C_Y}$:

$$8. \quad \mathbf{C_Y} = [1/(n-1)]\mathbf{PAP^T} \Rightarrow [1/(n-1)]\mathbf{P(P^TAP)P^T} \Rightarrow$$
$$[1/(n-1)]\mathbf{PP^{-1}APP^{-1}} \Rightarrow [1/(n-1)]\mathbf{A}$$

We have now diagonalized $\mathbf{C_Y}$, thus the rows of **P** (or eigenvectors of $\mathbf{XX^T}$) will become our principal components.

For this particular study, each alpha carbon on each amino acid acted as a point which fluctuated in 3-D space using concatenated and single trajectories. With this, E. Papaleo, et al., were able to extract the PCs from the holo-Mb which are utilized further.

Results:

Scientists have to assume that the most interesting dynamics occur along the motions with the largest amplitude (i.e. most principal). E. Papaleo, et al., found that:

1. 15 PCs are required to cover more then 70% of the variance
2. The majority of the variance can be described with 3 PCs

These PCs let us see the protein folding trajectories and coupled with FEL analysis, we can see a 3-D probability distribution (fig. 3): the higher the peak, the more likely that holo-Mb will be found in that particular conformation. However, we note that a given sample is not 3-D but rather multidimensional; fig. 3 does not show the full conformational space of holo-Mb.

From the FEL, after ensemble averaging, the two most probable conformations, are thus extracted and shown (A and B, fig. 4). Further analysis revealed that most of the changes between the two conformations centered on helices F and H and the region between helices C and D. This is important as helice F has experimentally been found to be responsible for reversible

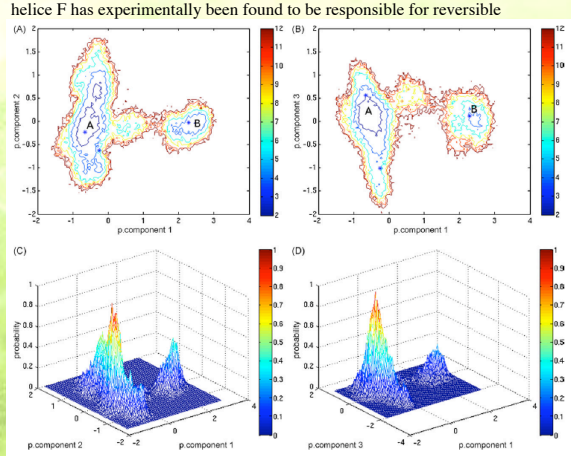

Figure 3: PC's graphed against FEL (A,B) and PC's graphed against probability (C,D). Asterisks (*) indicate localization of average structure of various ensembles (A and B) derived from the cluster analysis. Notice that the areas with lowest free energy (blue) correspond to areas of highest probability as predicted by thermodynamics. Free energy is calculated in kJ/mol.

Conclusions

Elevent independent simulations were ran on holo-Mb in order to sample the local conformational space. Using PCA to reduce the actual trajectories into lower dimensions and FEL, they were able to see two major conformations. These major conformations show shifts around helices **F** and **H**.

Experimental results show holo-Mb is highly constrained, we see two main conformations. These promising results have lead Papaleo, et al., to the possibility of extending this particular analysis to other variants of Mb. These other variants of Mb have much greater conformational flexibility, meaning greater variance. It becomes quite clear that applied linear algebra, such as PCA, supplements invaluable to scientific investigations.

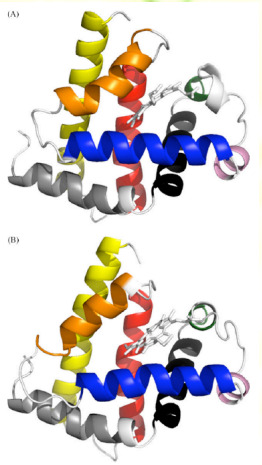

Figure 4: Average structure for each cluster ensemble: A (A) and B (B). The helices are colored as so: Helix A (gray), B (black), C (green), D (pink), E (blue), F (orange), G (red), and H (yellow).

Summary:

Molecular modeling is a rapidly maturing field that allows us to view the previously unobservable dynamics of important biomolecules. Because of the immense amount of data produced by a given simulation, it becomes extremely important to break it all down in order to see the wider picture. PCA allows us to take the various movement vectors of these biomolecules, and allows you to decompose the motions so we can see the important/principal dynamics that govern the system.

Coupled with other forms of analysis and experiment, we can better understand the nature of biomolecules such as holo-Mb at a much greater depth. Here, Papaleo, et al., made some important discoveries and correlations:

1. The major conformational shifts occur between helices **F** and **H** along with the region between helices **C** and **D**
2. Experimental results have shown reversible oxygen/carbon monoxide binding occurs through conformational shifts around helice **F**

This shows that these methods can be used to further study other interesting biologically relevant molecules.

Acknowledgements and References:

I thank Dr. Chang and Dr. Sorin for guidance and advice on this project.

Kavaraki, L. E. Connexions [Online] 2007. http://cnx.org/content/m11461/latest/ (accessed April 1, 2009).

Papaleo, E.; Mereghetti, P.; Fantucci, P.; Grandori, R., Gioia, L. D.; *J. Mol. Graph. Model.* (2009), doi: 10.1016/j.jmgm.2009.01.006

Shlens, J. [Online] 2009. http://www.snl.salk.edu/~shlens/pub/notes/pca.pdf (accessed Feb 9, 2009).