### 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.

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

6. $C_Y = \frac{1}{(n-1)}Y^TY$

The trick is to substitute $PX$ into $Y$ and $V$, do some manipulation and the resulting algebra produces

7. $C_Y=\frac{1}{(n-1)}P^TP$

where $A$ is $X^T$, a symmetric matrix. For a symmetric matrix, we know that $A=EDE^T$, $E$ 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 $XX^T$. We then select $P=E^T$ and because the inverse of an orthogonal matrix is its transpose, we know that $P^TP=P^TP^T$. We can then substitute for $X$ into $C_Y$:

8. $C_Y = \frac{1}{(n-1)}\sum_{i=1}^{n}P_iP_i^T$(change of basis): $\sigma$ define:

$\sigma_i$ is its transpose, we know that $C_Y$ should display.

*Figure 1.* Example Data for 2-D case. 1st PC maximizes variance, 2nd PC captures the next largest variance, etc., up to $n$ dimensions.

### Methods:

After collecting as much data on a system as possible, we take all the 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 $m \times n$ matrix, and define a matrix $P$ which transforms $X$ into another $m \times n$ matrix $Y$ (change of basis):

1. $PX=Y$

where the rows of $P$, $[p_1 \ldots p_n]$, 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 to determine 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:

### 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 than 70% of the variance
2. The majority of the variance can be described with 3 PCs

However, we note that a given sample is not 3-D but rather multidimensional.

### Conclusions

Element independent simulations were run 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 scientistic investigations.

### Summary:

Molecular modeling is a rapidly maturing field that allows us to view the previously unobservable dynamics of important biomolecules. Because 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 helix $F$

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.


---

Christopher Bruner and Dr. Jen-Mei Chang

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