JED: A Java Essential Dynamics Program for

Comparative Analysis of Protein Trajectories

Charles C. David1 and Donald J. Jacobs2,3\*

1Department of Bioinformatics and Genomics {cdavid.uncc.edu}, 2Department of Physics and Optical Science and 3Center for Biomedical Engineering and Science {djacobs1.uncc.edu} University of North Carolina at Charlotte, USA

**Abstract.** Principal component analysis (PCA) is commonly used to extract the essential dynamics of a protein described by eigenvectors (modes) of a covariance or correlation matrix constructed from atomic trajectories. The subspace defined by a small set of PCA modes with largest eigenvalues quantifies conformational space explored by a protein. In bioinformatics studies there is a need to compare subspaces among different proteins for similarity assessment, but software to facilitate this comparative-analysis is lacking. We developed the Java Essential Dynamics (JED) package to perform many variants of PCA applied to user-defined regions within a protein. Operationally, JED compares subspaces defined by the top eigenvectors using several metrics to quantify the similarity/overlap of high dimensional vector spaces with features that help to quantify statistical significance.

**Keywords:** Essential Dynamics, PCA, distance PCA, vector space, subspace, RMSIP, principal angle

**1. Introduction**

Essential dynamics is a common application of principal component analysis (PCA) to extract biologically relevant motions from atomic trajectories of proteins. A covariance (Q) or correlation (R) based form of PCA for a selected set of atoms is applied to yield PCA modes (eigenvectors) and their eigenvalues. The displacement vectors (DV) relative to a reference structure are projected onto the top set of modes to produce a set of principal components, which are used to cluster conformations. The subspace defined by a relatively small set of PCA modes with largest eigenvalues quantifies conformational space explored by a protein. Comparisons over a set of different protein trajectories for similarity assessment are needed for bioinformatics studies, but software to facilitate this comparative-analysis is lacking. Using myosin as an illustrative case, we describe the key features in a Java Essential Dynamics (JED) software package open to the public.

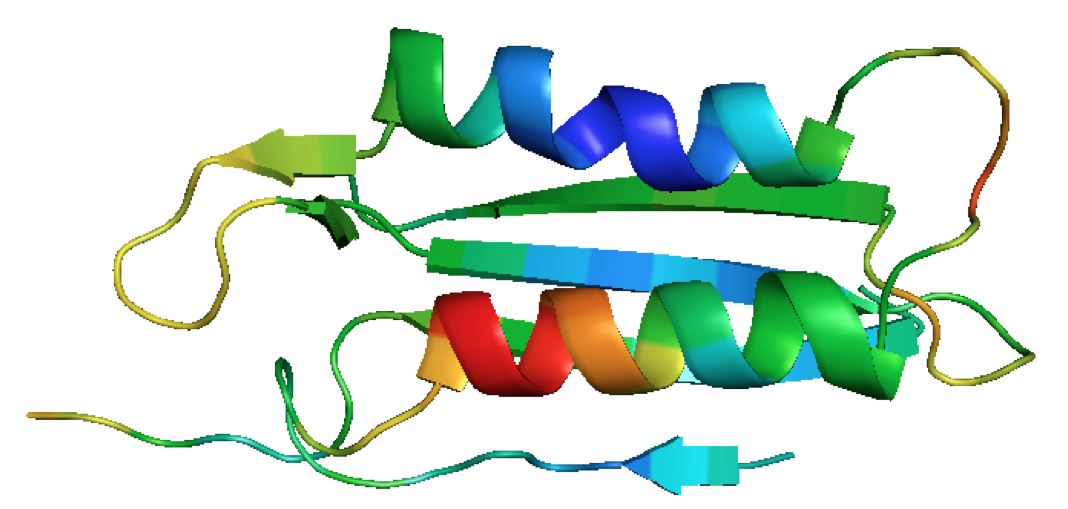
**2. Methods**

The root mean square inner product (RMSIP) [1] and Principal Angles (PA) [2] are two metrics among others that JED employs to compare the similarity/overlap between two subspaces defined by the PCA eigenvectors. RMSIP is a single number that quantifies all inner product pairs of eigenvectors from each space. The set of PA quantifies the optimal alignment between two subspaces based on a singular value decomposition (SVD) [3] of a matrix defined by all inner product pairs between two subspaces of equal dimension, *n*. The PA metric yields a monotonically increasing set of *n* angles, such that values (less, greater) than 30 degrees indicate (good, poor) overlap [4]. In addition, JED allows a user to pool (or augment) trajectory data so that DV projected on the first and second PCA modes becomes an effective way to cluster protein dynamics [4].

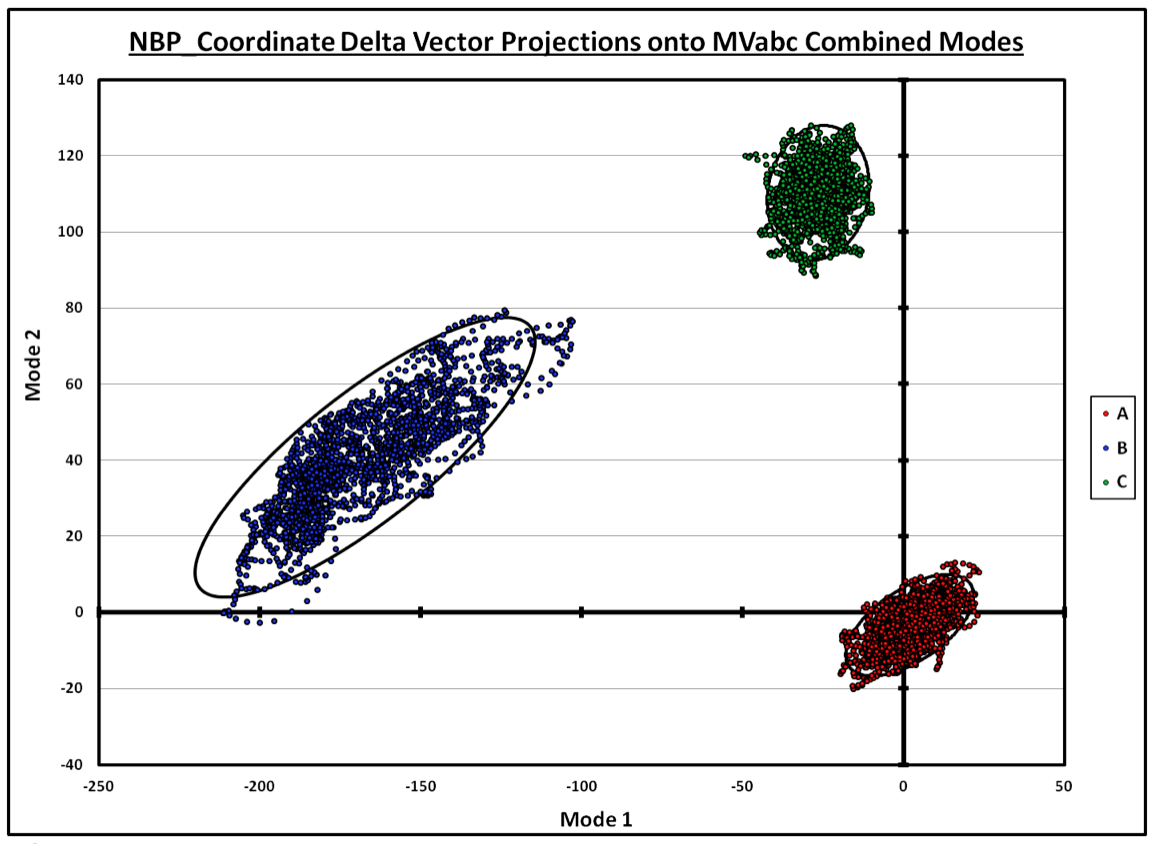
**3. Results**

JED reads trajectory data from sets of PDB files or from a matrix of atomic coordinates. The PDB files may be single or multi-chained. A set of residues can be chosen that need not be contiguous. The analysis is based on alpha carbon atoms from each of the selected residues. Cartesian-based coordinates (cPCA) or internal distance coordinates (dPCA) can be applied with Q and R based PCA. JED outputs results as text files to be processed using graphing software. The output includes the transformed coordinates, conformation and residue RMSDs, PCA modes with their eigenvalues and DV projections onto the top principal components. PDB files are generated along with a Pymol™ script to visualize each of the top modes derived from Q and R cPCA as either a static picture or as a movie.

Comparative analyses for the myosin V protein illustrate key features in JED. The dataset consists of X-ray crystal structures for myosin V in different nucleotide states with two protein domains (labeled as chain A and B) that attach to the lever arm. All-atom trajectories were generated for Myosin V [5] in its Apo form, ATP and ADP bound states with chains A and B removed, chain B removed and both chains A and B removed. The question of interest is how dynamics changes in the nucleotide-binding pocket (NBP) for these different structures. The NBP consists of residues xxx to yyy in consecutive order and from zzz to aaa in consecutive order. Thus the dimension of the subspace being analyzed is bbb, corresponding to the non-contiguous region shown in **Fig. 1**.



**Figure 1:** The nucleotide-binding pocket colored according to regions showing the most to the least mobility indicated by colors: red, orange, yellow, green, cyan and dark blue.

Differences in the NBP dynamics in the apo forms for three nucleotide-binding states are observed in the DV scatter plot on the first two PCA modes when trajectories are pooled as shown in **Fig. 2**. The native state dynamics within the NBP is markedly modified by the nucleotide binding state, related to subtle structural differences and not because of the ****presence of a substrate. This latter effect is quantified using the RMSIP and PA metrics.

**Figure 2:** The essential dynamics in the nucleotide-binding pocket is well separated by pooling trajectories for the apo forms of the three nucleotide states where A, B and C correspond to the rigor, ATP and ADP nucleotide states.

JED was used to quantify essential motion similarity within the NBP when the substrate ATP or ADP is present or when the chains A and/or B are present. Fifteen different pairs of comparisons were considered as defined in **Table 1**.

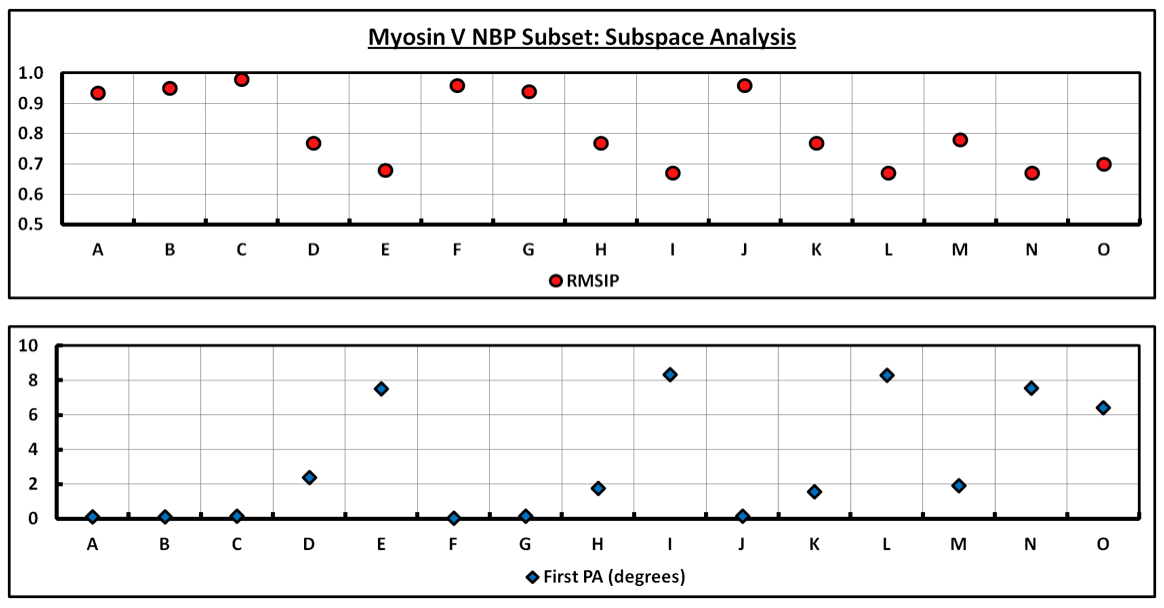
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
| 1-2 | 1-3 | 1-4 | 1-5 | 1-6 | 2-3 | 2-4 | 2-5 | 2-6 | 3-4 | 3-5 | 3-6 | 4-5 | 4-6 | 5-6 |

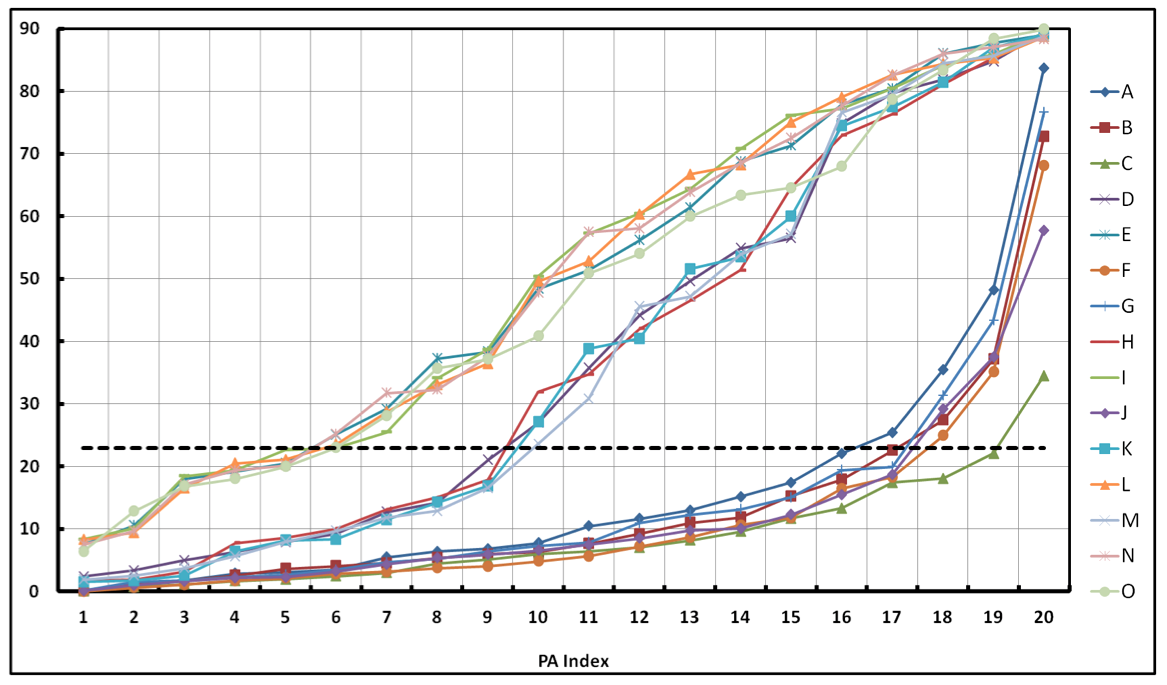
**Table 1:** Letters A to O define pairs of structures each assigned an identification number (ID #). The ID #s are defined as: 1 🡪 apo form with chain A attached, 2 🡪 apo form with chains A and B attached, 3 🡪 holo form with chains A and B attached, 4 🡪 holo form with chain A attached, 5 🡪 ADP substrate with no chain attached, and 6 🡪 ATP substrate with no chain attached.

The RMSIP values and the first principal angles are shown in **Fig. 3** for each of the 15 different pairs (A to O). Both metrics track each other, showing that the NBP has similar dynamics oftentimes, but can shift considerably for certain comparisons.

**4. Conclusions**

JED encourages best practices for essential dynamics analysis that includes comparative studies over multiple protein trajectories over the entire protein or user-defined regions of interest, and to interrogate the significance of the results.

**Figure 3:** Large RMSIP corresponding to a small first principal angle indicates similar essential dynamics, whereas small RMSIP and large first principal angle indicates different dynamics.



**Figure 4:** Large RMSIP corresponding to a small first principal angle indicates similar essential dynamics, whereas small RMSIP and large first principal angle indicates different dynamics.

**5. References**

[1] good source for RMSIP

[2] good source for PA

[3] protein motion paper.

[4] book chapter.

[5] 2012 myosin paper in Biophysical Journal.