**I. INTRODUCTION**

**Java Essential Dynamics** (JED) is a java library (a package of programs) for analyzing protein trajectories. The trajectories may be derived from MD, FIRST/FRODA, or any other dynamic simulations that output a trajectory as a set of PDB files. The program can handle single chain PDB files with no chain identifier as well as multi chain PDB files that use chain IDs. The user may specify the set of residues to be considered for the analysis, and this set need not be contiguous. A variety of utility tools are provided that use **Principal Component Analysis** (PCA) that are not found in MD-simulation packages or other stand alone PCA software, especially in regards to comparative analysis of multiple trajectories. JED is capable of running on any platform with a suitable Java Runtime Environment (JRE).

**Expected Input to JED:**

Ideally, each PDB structure must follow standard PDB-format. Note that some deviations from standard will often work fine, but JED expects standard format. Moreover, it is required that the structures have been prepared in such a way that there are no gaps in residue labeling. If residues or consecutive regions of residues are missing, these need to be fixed, or the residue labeling has to be altered in order to remove gaps in residue labels. The first residue label must start at 1 or higher. No 0 or negative numbers are allowed for residue labels. All preprocessing of the PDF files must be done with external software before using JED. It is convenient to label PDB files using leading zeros in the name of the files to simplify tracking time progression. For example, if a simulation generates 100,000 frames in the trajectory, it is best to name the PDB files like <file\_name\_000000>, <file\_name\_000001>, … <file\_name\_100001>, which specifies that relative to the starting structure 100,000 frames are generated in successive order. Although this naming scheme is not required, it is highly recommended because it allows the user to track time order easily on operating systems that sort order by literal alphabetic-characters, rather than interpreting 34 is less than 100, for example.

**JED Preprocessing Output:**

As a pre-processing step, JED reads in all PDB files in a specified directory and aligns all the structures in the trajectory to a specified reference structure using a quaternion alignment algorithm. A matrix of the read **PDB coordinates** obtained from all the residues from the input PDB files are created so that they can be used for all subsequent JED runs. A list of all the residues (**residue list**) found in the PDB files (along with the chain IDs when appropriate) is generated. The original and transformed **conformation rmsd** is determined for each member structure in the trajectory relative to the specified reference structure. The **residue rmsd** (also commonly referred to as **RMSF**)is determined from the entire trajectory. An **edited PDB file** is also generated where the B-factors are replaced with the **residue rmsd** values for visualization purposes. This output automatically happens and is non-optional.

**Carbon Alpha Atoms:**

The current implementation of JED only considers Cα atoms. As such, we speak about residues because the information is tied to the Cα atoms, which represents the dynamics of the residue at some coarse grained level of description. For example, the distance between two residues is modeled in JED as the distance between the two Cα atoms associated with the two residues. This choice of working only with Cα atoms allows the labeling of the Cα atoms to be associated with residue labels. For a single chain protein, this is a simple 1 to 1 mapping. For multiple chain proteins, JED also tracks the chain ID.

**Different Types of PCA:**

The core element of essential dynamics is to perform PCA. JED implements two variations of PCA. The first and most common method is based on Cartesian coordinates (**cPCA**). The second method is based on internal coordinates using residue-pair distances (**dPCA**), which is much more computationally expensive and the interpretation is difficult when the number of residue-pairs exceeds 10. Note that dPCA using n residues will yield eigenvectors having n\*(n-1)/2 components, each corresponding to one set of residue-pair distances. Thus, for five residues, one obtains 10 pairs, and this becomes difficulty to interpret. Charles, why can’t I specify 20 residue pairs as an input instruction to JED. I rather focus on residue-pair selection, rather than residue selection. Can this be changed?

Both the cPCA and dPCA can be constructed as a **covariance matrix** (**Q**) and/or **correlation matrix** (**R**). The correlation matrix is a normalized version of the covariance matrix. The results obtained from **Q** and **R** generally differ somewhat due to the inherent statistical biases in each approach. The user is required to specify which approach to use, or both.

**Conditioning of the Q and R Matrices:**

JED handles the removal of outliers prior to the PCA analyses with two approaches. First, the user can specify the **percent** of the structures to be removed based on the conformation rmsd. The most deviant structures are tagged as outliers and subsequently removed prior to the PCA analysis. In this first method, frames that are identified as an outlier are thrown out from the sample. Second, the user can specify a **z-score cutoff** such that when a PCA-entry (either as a Cartesian or internal distance coordinate) has a |deviation| from its mean that exceeds its own z-score cutoff, it is identified as an outlier. For each PCA-entry that is identified as an outlier, it is replaced with its mean. This process is done per frame, and each PCA-entry is treated independently. In this second method, a frame is never thrown away, but only some entries within a frame may be modified. Both methods are intended to reduce the condition numbers of **Q** and **R**. While the first method of conditioning is most commonly employed in the protein field (if at all), the second method of conditioning most commonly used in the field of statistics, and is the preferred method due to its superior effectiveness. It should be noted here that without applying the conditioning, the interpretation of the results of PCA can be completely wrong --- depending on how much statistics one has. It is highly recommended to use the z-score cutoff conditioning method in all applications to avoid misinterpreting the PCA results.

**Visualization of cPCA modes:**

JED computes the **PCA modes** (RMSD and MSD, with and without weight by the corresponding eigenvalue) from the Cartesian eigenvectors so that they may be mapped to the residue set. As noted above, sets of structures can be generated to visually inspect the cPCA modes. Eigenvectors from dPCA cannot be mapped to the residue set in any simple way, so no mapping or visualization is attempted. The user can specify the number of Cartesian modes to visualize. Mode visualization is done by creating a set of 20 PDB files that capture the displacement of the alpha carbons for the given mode. A scale-factor parameter is adjustable to control the amount of displacement in the modes. A Pymol© script is generated to animate the frames.

**Dimension Reduction Level:**

The primary purpose of applying PCA to capture the essential dynamics of a protein is to reduce the large dimension of variables to a much smaller number of variables that captures the greatest variance in protein motion. The **Q** and **R** matrices, once diagonalized, provide a set of eigenvalues and eigenvectors. The eigenvalues for proteins typically fall off fast for the first several modes, out of possibly thousands of modes. The number of dimensions needed to provide a fair assessment of the essential dynamics in a protein is system-dependent. The user can specify any number (say 20, which typically is more than needed) to obtain results for all possible selections ranging from 1 up to the value selected. In this way, the user can see how the added dimensions help glean more information, albeit making it harder to interpret the greater number of dimensions. Eventually, the user must decide based on their purpose/goals what the optimal number of dimensions to use for representing the essential dynamics.

**Displacement Vectors:**

A set of **displacement vectors** (**DV**s) based on the full conformational space is calculated using a specified reference structure. Those **DV**s are then projected onto a set of eigenvector directions to create delta vector projections (**DVP**s), which are similar to principle components (**PC**s). The **PC**s are delta vector projections, but according to the standard definition used in statistics, they are always relative to the mean conformation position as defined in the construction of the **Q** or **R** matrix. In studying the essential dynamics of a protein, it is common to use a reference structure that has a particular physical or biochemical meaning, which is why we call these displacements **DVP**s, and not **PC**s. The DVPs are very useful to have for visualizing protein motions. For example, if the first two eigenvector directions are selected (those eigenvectors associated with the highest and second highest eigenvalues, or variance) the **DV**s can be plotted for each frame to construct the trajectory in conformational space projected onto a two dimensional cross-section. Other eigenvector directions can be specified, allowing the user to investigate how the trajectory projects into the space defined by each eigenvector. The **DVP**s are given using un-normalized and normalized inner products, as well as weighted by the corresponding eigenvalue.

**Post PCA Comparative Subspace Analysis:**

JED performs a subspace analysis (**SSA**) on two different sets of eigenvectors of equal numbers generated from the **Q** or **R** variants of PCA. This allows one to quantitatively determine how different the PCA results are due only to the choice of PCA model. A comparison is also made to a random set of eigenvectors derived from the same vector space. This allows the user to establish a baseline for statistically valid subspace comparisons. I suggest to remove this feature. Additional analysis can be done using the Subspace Analysis classes or the associated driver programs. To perform these kinds of tests, it is best practice to first generate a set of eigenvectors from each trajectory of interest as well as from the pooled trajectories to use as a reference set. Subspace analysis is done by comparing the sets of eigenvectors and determining the **RMSIP**, Principal Angles (**PA**s), cumulative overlap (**CO**), cosine products (**CP**), vectorial angular sum (**VAS**), and the maximum angle between subspaces of the given vector space.

(In this tutorial, code, file paths, and text file content are shown in dark blue 9 point Consolas)

**II. Understanding JED**

**JED Install Instructions:**

The program can be run from compiled source or from the provided executable jar files. It is critical that the environment variable Java **CLASSPATH** be correctly set. Alternatively, you can always specify the complete absolute path for files and or add the -cp option to the java command. Make more extensive instructions for someone like me to be able to install. I think this is too thin. A few step by step instructions (for dummies) on PC, MAC and Linux system would be good.

**Expected Memory Requirements:**

On high performance computer clusters make sure the 64 bit JRE is installed. Memory use is demanding because JED holds the complete covariance matrix (among other data structures) and matrix diagonalization scales as O(N3). Typically 15 to 35 GB of RAM will be needed depending on the size of the protein. For very large proteins consisting of thousands of residues and/or many tens of thousands of frames, make available as much memory per node as possible (> 1TB).

**Two Kinds of JED Drivers:**

There are two driver programs for JED: One (**JED\_Driver**) runs a single job using parameters specified in the input file, and the other (**JED\_Batch\_Driver**) runs a batch of jobs sequentially. The first is suited for running a single job at the command line or when using submit scripts on computer cluster resources. This can be implemented using job arrays so that your jobs run in parallel rather than sequentially. The second is suited for running multiple jobs on a single computer so that a user can submit a batch of jobs, and come back a few hours later with many different jobs finished without having to launch each one separately.

**Input File and Data for JED Driver:**

JED requires an input file for job parameters. The format of this file will be described below. The run command takes only one argument, which is the name of the input file that includes the absolute path to the file. Is this path relative to where the JVM is run, or absolute path? If no argument is specified, then JED assumes that the default input file name is used and the file is located in the same directory from which the JVM was called. The default input file names are:

**JED\_Driver.txt** for JED\_Driver.java (or jar file)

**JED\_Batch\_Driver.txt** for JED\_Batch\_Driver.java (or jar file)

Each job should be assigned to its own directory, which must contain either the PDB files to read or the coordinate file to process, along with the reference PDB file and residue lists for specifying either the Cartesian or Distance subsets.

**JED Command Line format:**

To run **JED** at the command prompt or within a PBS script, you can use one of the following commands:

Java -d64 JED\_Driver “/path/to/your/input/file.txt” (runs the compiled java program)

Java -jar -d64 JED\_Driver.jar “/path/to/your/input/file.txt” (runs the executable jar file)

To run **JED\_Batch\_Driver** at a command prompt or in a PBS script, you can use one of the following commands:

Java -d64 JED\_Batch\_Driver “/path/to/your/input/file.txt”)

Java -jar -d64 JED\_Batch\_Driver.jar “/path/to/your/input/file.txt”

**Organization of Output Files:**

Output files from JED are written to subdirectories within the working directory, structured to organize the multitude of files produced in a meaningful manner. The top level of this directory tree is named "JED\_Results\_**$description**", where **$description** is a user set parameter that succinctly describes the job. Limbs of the tree separate cPCA, dPCA, and mode visualization analyses (**viz**), when present. Each of these in turn contains limbs for **Q** and **R** compartmentalization. Most of the output file names include the **number of residues** in the selected subset for reference plus a description of the file contents.

**Current Limitations:**

Initial input of the protein trajectory must be using PDB files that are expected to conform to the standard format. Only carbon-alpha atomic positions are used to create a **Q** or **R** matrix for essential dynamic analysis. Each PDB file must have the exact same number of residues. JED cannot process a PDB file with missing residues or have alternate conformations within a given frame based on fractional **occupancy** values. Only a single conformation per frame is allowed.

**III. Step by Step Instructions to Run JED**

**IN ALL CASES:**

A preliminary run with NO PCA must be performed to generate the JED formatted coordinate matrix file for all the alpha carbons in the PDB files. This makes subsequent subset analyses much faster to perform. It also serves to guarantee that the specified residues for subset selection are correctly chosen. After this initialization step, the PDB files can be deleted or archived, with the exception of the reference PDB file. Once the coordinate matrix is created, it should be used for all subsequent analyses using different residue subsets and different job parameters.

The name of the coordinate file matrix produced from the PDB files is: "original\_PDB\_coordinates.txt"

The matrix packing is as follows:

**Rows are coordinate variables and columns are frames.**

For N residues, there are 3N rows: N x-coordinates, N y-coordinates, and N-z coordinates, stacked in that order.

**The file to use in all subsequent JED analyses is the original\_PDB\_coordinates matrix.**

Note: The most critical step in preparing to run JED is in the creation of the input file. The input file must have the correct format (shown in examples below) and the entries must be accurate. If either of these conditions is violated, the program will crash, or worse, the results will be corrupt.

**Common Causes for JED to Crash**

If **any** **directory** cannot be found or if **any** **file** cannot be found, JED will crash.

If unexpected format is found in **any** of the input files, JED will crash.

**IV. JED DRIVER:**

**A. The Preliminary Run**

**i. Run Command:**

java -jar -d 64 JED\_Driver.jar “/path/JED\_Driver.txt”

***The PDB files (including the PDB reference file) must be in the working directory.***

***JED\_Driver.txt or user-defined file name must be in the working directory.***

This pre-processing step will read all PDB files in the working directory, but will perform **no PCA.**

The purpose of **no PCA** this is to generate the matrix of coordinates for performing subset analyses efficiently.

**ii. Standard NO-PCA Output Files:**

These are written to the **root** of the JED Results directory tree:

/working/directory/JED\_Results\_Description/

**JED LOG** providing a summary of the job parameters and results:

JED\_Log.txt

**PDB READ LOG** listing all the PDB files read, in the order they were read:

PDB\_READ\_Log.txt

**coordinates matrix** from all the alpha carbon coordinates in the PDB files:

original\_PDB\_coordinates.txt

**transformed coordinates matrix**, which aligns all the frames to the reference frame :

ss**\_$num**\_res\_transformed\_PDB\_coordinates.txt

**list of all residues** found in the PDB files for subsequent editing and use:

All\_PDB\_Residues\_JED.txt

All\_PDB\_Residues\_Multi\_JED.txt (for Multi runs)

**original and transformed conformation RMSDs**:

ss\_**$num**\_res\_original\_conformation\_rmsds.txt

ss\_**$num**\_res\_conformation\_rmsds.txt

**residue RMSDs (RMSF)**:

ss\_**$num**\_res\_residue\_rmsd.txt

**edited PDB file** containing all the residues with the RMSF replacing B-factors:

ss\_**$num**\_res\_edidited.PDB

**coordinates z-score matrix** from all the alpha carbon coordinates in the PDB files:

ss\_**$num**\_res\_coordinate\_Z\_scores.txt

**percent** of the frames to remove based on conformation RMSD (OPTIONAL)

**z-cutoff** for adjusting coordinate outliers to their mean values (OPTIONAL)

If these parameters are set to a value other than zero, additional output files are generated with the results. However, the appropriate time to handle outliers is during runs that actually perform PCA and it is recommended that for pre-processing runs, these be set to zero.

**iii. Below is a sample input file for Single Chain PDB files:**

-----------------------------------------------------------------------------------------------------

0.00

0.00

1 0

0 0 0

/working/directory/job/

Description reference\_PDB\_file.pdb

-----------------------------------------------------------------------------------------------------

Notes: This is a tab (space) separated file. There must be NO LEADING blank lines or spaces (no trailing too).

These formatting issues (no leading black lines or trailing blanks) should not be an issue at all. Can you make java read the input files appropriately?

Line 1 Field 1: specifies the **percent** (double) of frames to remove from the data: 0 🡪NONE

Line 2 Field 1: specifies the **z-cutoff** (double) for adjusting outliers: 0 🡪NO ADJUSTMENT

Line 3 Field 1 specifies whether to **read PDB files** (0 or 1) 🡪1 = yes

Field 2: specifies if the PDB files are **Multi** Chain (0 or 1) 🡪 0 = no

Line 4 Field 1: specifies the **number of Cartesian** (integer) modes to process 🡪0 = none

Field 2: specifies the **number of Distance** (integer) modes to process 🡪0 = none

Field 3: specifies the **number of Cartesian modes to Visualize**(integer) 🡪0 = none

Line 5 Field 1: specifies the **working directory** (String)

Line 6 Field 1: specifies the **description** (String) for the requested job

Field 2: specifies the **reference PDB** (String) for the requested job

-----------------------------------------------------------------------------------------------------

**Warning**:

If the number of cPCA modes and number of dPCA modes are set to 0, then JED does a **no PCA** analysis.

The **read** flag must be set to 1 and the **Multi** flag must be set to 0.

All PDB files must have the same number of residues. The matrix of alpha carbon coordinates is determined from the first PDB file read. If other files in the working directory do not match exactly, then the array sizes will not match and the program will crash. IF JED crashes during the reading of PDBs, this is probably the reason.

In subsequent analyses, it is critical that no residues are requested that do not actually exist in the PDB files!

JED maps the specified residue list to an internal list that is aligned to the columns of the coordinates matrix.

**The file to use in all subsequent JED analyses is the original\_PDB\_coordinates matrix.**

This matrix contains all the residues in the PDB files and thus can be used for any subset of those residues. When a subset is chosen, a new correspondence set is generated and a new transformation is done to optimize the alignment of the structures. This removes overall translation and rotation.

**iv. Below is a sample input file for Multi-Chain PDB files:**

**Differences from Single-Chain analysis shown highlighted in amber**

-----------------------------------------------------------------------------------------------------

0.00

0.00

1 **1 2 A B 795 151 0 0**

0 0 0

/working/directory/job/

Description reference\_PDB\_file.pdb

-----------------------------------------------------------------------------------------------------

Notes: This is a space delineated file. There must be NO LEADING blank lines or spaces (no trailing too).

In all modern program Languages, it is easy to incorporate free formatting. Can you eliminate this restriction?

Line 1 specifies the **percent** (double) of frames to remove from the data: 0 🡪NONE

Line 2 specifies the **z-cutoff** (double) for adjusting outliers: 0 🡪NO ADJUSTMENT

Line 3 Field 1 specifies whether to **read PDB files** (0 or 1) 🡪1 = yes

Field 2 specifies if the PDB files are **Multi** Chain (0 or 1) 🡪 1 = yes

Field 3 specifies the number of chains 🡪 (2)

Field 4 specifies the first chain ID 🡪 (A)

Field 5 specifies the second chain ID 🡪 (B)

Field 6 specifies the number of residues in chain A🡪 (795)

Field 7 specifies the number of residues in chain B🡪 (151)

Field 8 specifies the offset of Chain A 🡪 (0)

Field 9 specifies the offset of Chain B 🡪 (0)

Line 4 Field 1 specifies the **number of Cartesian** (integer) modes to process 🡪0 = none

Field 2 specifies the **number of Distance** (integer) modes to process 🡪0 = none

Field 3 specifies the **number of Cartesian modes to Visualize**(integer) 🡪0 = none

Line 5 specifies the **working directory** (String)

Line 6 Field 1 specifies the job **description** (String) for job one

Field 2 specifies the **reference PDB** (String) for job one

-----------------------------------------------------------------------------------------------------

**Warning**:

If the number of cPCA modes and number of dPCA modes are set to 0, then JED does a **no PCA** analysis.

The **read** flag must be set to 1 and the **Multi** flag must be set to 0.

All PDB files must have the same number of residues. The matrix of alpha carbon coordinates is determined from the first PDB file read. If other files in the working directory do not match exactly, then the array sizes will not match and the program will crash. IF JED crashes during the reading of PDBs, this is probably the reason.

Multi Chain PDBs must have unique chain identifiers for every chain. A missing chain identifier will cause JED to crash.

In subsequent analyses, it is critical that no residues are requested that do not actually exist in the PDB files!

JED maps the specified residue list to an internal list that is aligned to the columns of the coordinates matrix.

**The file to use in all subsequent JED analyses is the original\_PDB\_coordinates matrix.**

This matrix contains all the residues in the PDB files and thus can be used for any subset of those residues. When a subset is chosen, a new correspondence set is generated and a new transformation is done to optimize the alignment of the structures. This removes overall translation and rotation.

**B. Debugging Crashes Part I:**

Things that will generally make your life miserable…

**i. Simple mistakes:**

Does the directory exist? If not, CRASH!!!

Does the path to the input file exist? If not, CRASH!!!

Does the input file exist in the proper location? If not, CRASH!!!

Does the input-file start on the first line? If not, CRASH!!!

The directory contains PDB files of different sizes CRASH!!!

The directory does not contain the reference PDB file CRASH!!!

Is that integer really an integer or is it a float or a string? CRASH!!! (what is the difference between 3 and 3?)

Does the directory string end in “/” or “\\”?: If not CRASH!!!

Are the 0 and 1 flags set correctly? If not CRASH!!!

**ii. More subtle mistakes:**

The directory contains PDB files with 2 chains, but no chain IDs Non-sense results!

If the PDB file names are sorted in a different order than how they were generated, then the conformation RMSD results will not reflect what actually occurred in the simulation. Naming the PDB files appropriately by padding the numbers with leading zeros will ensure proper sorting to prevent this problem caused by the operating system.

If the conformation RMSD is very different from what you expect, then you may be using PDB files that contain occupancy information. JED does not use that information. Your results will not be accurate.

If you have pooled data, make sure the combined matrix is accurate.

**C. Performing Cartesian PCA**

**i. run command:**

java -jar -d64 JED\_Driver.jar “/path/JED\_Driver.txt”

***The working directory must contain: The coordinates matrix, the PDB reference file, and the residue list.***

***JED\_Driver. txt or user-defined file name must be in the working directory.***

The purpose of this type of run is to perform Essential Dynamics using cPCA based on **Q** and **R**. The user specifies the subset of interest for the analysis, which may be the entire protein or a sub-region, which can be non-contiguous, by providing a residue list file. This task is simplified since JED has already created a list of all the residues in the protein. The user can simply edit this file. The cPCA results are written to the sub-directory "cPCA" and the Visualizations of the top modes (when selected) are written to the subdirectory "viz". The directory cPCA has sub directories for the **Q** and **R** analysis, as does the viz directory.

**ii. Standard Output Files:**

These are written to the **root** of the JED Results directory tree:

/working/directory/JED\_Results\_Description/

**JED LOG** providing a summary of the job parameters and results:

JED\_Log.txt

**subset transformed coordinates matrix**:

ss\_**$num**\_res\_transformed\_PDB\_coordinates.txt

**original and transformed conformation RMSDs**:

ss\_**$num**\_res\_original\_conformation\_rmsds.txt

ss\_**$num**\_res\_conformation\_rmsds.txt

**residue RMSDs (RMSF)**:

ss\_$num\_res\_residue\_rmsd.txt

**subset edited PDB file** with the RMSF replacing B-factors:

ss\_**$num**\_res\_edidited.PDB

**subset coordinates Z-Score matrix**:

ss\_**$num**\_res\_coordinate\_Z\_scores.txt

**list of frames removed**, based on the **percent** parameter:

ss\_**$num**\_res\_ Removed\_Conformation\_Outliers.txt

**trimmed transformed coordinate matrix**:

ss\_**$num**\_res\_ trimmed\_$percent\_percent\_PDB\_coordinates\_COLS.txt

**list of coordinate variables adjusted**, based on the **z-cutoff** parameter:

ss\_**$num**\_res\_ adjustments\_per\_variable.txt

**adjusted transformed coordinate matrix**:

ss\_**$Z**\_threshold\_**$z**-cutoff\_adjusted\_PDB\_coordinates\_ROWS.txt

**iii. Standard cPCA Output Files:**

These are written to the **/cPCA subdirectory**  of the JED Results directory tree:

/working/directory/JED\_Results\_Description/cPCA/

**centroids (means) of the variables:**

ss\_**$num**\_res\_centroids\_of\_variables.txt

**standard deviations of the variables:**

ss\_**$num**\_res\_std\_devs\_of\_centered\_variables.txt

**displacement vectors:**

ss\_**$num**\_res\_delta\_vectors.txt

**The COV output files are written to the /COV subdirectory of /cPCA**

/working/directory/JED\_Results\_Description/cPCA/COV/

**The CORR output files are written to the /CORR subdirectory of /cPCA**

/working/directory/JED\_Results\_Description/cPCA/CORR/

**covariance matrix:**

ss\_**$num**\_res\_COV\_matrix.txt

**eigenvalues:**

ss\_**$num**\_res\_eigenvalues\_COV.txt

**top eigenvalues:**

ss\_**$num**\_res\_top\_**$num**\_of\_cart\_modes\_eigenvalues\_COV.txt

**top eigenvectors:**

ss\_**$num**\_res\_top\_**$num**\_of\_cart\_modes\_eigenvectors\_COV.txt

**top pca modes and top weighted pca modes:**

ss\_**$num**\_res\_top\_**$num**\_of\_cart\_modes\_pca\_modes\_COV.txt

ss\_**$num**\_res\_top\_**$num**\_of\_cart\_modes\_weighted\_pca\_modes\_COV.txt

**top square pca modes and top weighted square pca modes:**

ss\_**$num**\_res\_top\_**$num**\_of\_cart\_modes\_square\_pca\_modes\_COV.txt

ss\_**$num**\_res\_top\_**$num**\_of\_cart\_modes\_weighted\_square\_pca\_modes\_COV.txt

**top PCs and top weighted PCs:**

ss\_**$num**\_res\_top\_**$num**\_of\_cart\_modes\_PCs\_COV.txt

ss\_**$num**\_res\_top\_**$num**\_of\_cart\_modes\_weighted\_PCs\_COV.txt

**top PCs and top weighted PCs:**

ss\_**$num**\_res\_top\_**$num**\_of\_cart\_modes\_normed\_PCs\_COV.txt

ss\_**$num**\_res\_top\_**$num**\_of\_cart\_modes\_weighted\_normed\_PCs\_COV.txt

**iv. SSA Output Files:**

These are written to the **/SSA subdirectory**  of /cPCA:

/working/directory/JED\_Results\_Description/cPCA/SSA/

**The Fast SSA Iterated Log:**

JED\_FSSA\_Iterated\_log.txt

**The SSA Log:**

JED\_SSA\_dim\_$top\_num\_cart\_modes\_log.txt

**The Random SSA Log**

JED\_Random\_SSA\_log.txt

**There are many more files in the /SSA directory that are flat files of the results reported in the log files:**

**RMSIPs**

**PAs**

**COs**

**Cosine Products**

**Vectorial Sum of Angles**

**v. Below is a sample input file for Single-Chain PDB files:**

------------------------------------------------------------------------------------------

0.01

3.00

0 0

20 0 2 1.0

/working/directory/

Description reference\_PDB\_file.pdb

residues.txt

original\_PDB\_Coordinates.txt 0

------------------------------------------------------------------------------------------

Notes:

Line 1 specifies the **percent** (double) of frames to remove from the data: 0 🡪 1%

Line 2 specifies the **z-cutoff** (double) for adjusting outliers: 0 🡪 values beyond |3.0|

Line 3 Field 1 specifies whether to **read PDB files** (0 or 1) 🡪 0 = no

Field 2 specifies if the PDB files are **Multi** Chain (0 or 1) 🡪 0 = no

Line 4 Field 1 specifies the **number of Cartesian** (integer) modes to process 🡪 20

Field 2 specifies the **number of Distance** (integer) modes to process 🡪 0 = none

Field 3 specifies the **number of Cartesian modes to visualize**(integer) 🡪 2 = top two modes

Field 4 specifies the **scale factor** for the visualizations (double) 🡪 1.0 = default value

Line 5 specifies the **working directory** (String)

Line 6 Field 1 specifies the job **description** (String)

Field 2 specifies the **reference PDB** (String)

Line 7 specifies the **residue list** (String)

Line 8 Field1 specifies the **coordinate matrix** (String)

Field 2 specifies the **reference column** (integer)

------------------------------------------------------------------------------------------

**Warning**:

Both the read PDB file flag and the **Multi** flag must be set to 0 to perform a JED analysis on single chain PDBs.

The number of cPCA modes must not be 0.

There must be a Cartesian residue list specified.

The number of dPCA modes must be zero.

There must not be a Distance residue list specified.

If the subset you have chosen contains N residues, then you must NOT request more than 3N modes.

If you request N cPCA modes then you can only visualize up to N top modes.

**If number of cPCA modes > 0, then there must be a Cartesian residue list file!**

**vi. Below is a sample input file for Multi-Chain PDB files:**

**Differences from Single-Chain analysis shown highlighted in amber**

------------------------------------------------------------------------------------------

0.01

3.00

0 1 **2 A B 795 151 0 0**

20 0 2 1.0

/working/directory/

Description reference\_PDB\_file.pdb

residues.txt

original\_PDB\_Coordinates.txt 0

------------------------------------------------------------------------------------------

Notes:

Line 1 specifies the **percent** (double) of frames to remove from the data: 0 🡪 1%

Line 2 specifies the **z-cutoff** (double) for adjusting outliers: 0 🡪 values beyond |3.0|

Line 3 Field 1 specifies whether to **read PDB files** (0 or 1) 🡪 0 = no

Field 2 specifies if the PDB files are **Multi** Chain (0 or 1) 🡪 1 = yes

Field 3 specifies the number of chains 🡪 (2)

Field 4 specifies the first chain ID 🡪 (A)

Field 5 specifies the second chain ID 🡪 (B)

Field 6 specifies the number of residues in chain A🡪 (795)

Field 7 specifies the number of residues in chain B🡪 (151)

Field 8 specifies the offset of Chain A 🡪 (0)

Field 9 specifies the offset of Chain B 🡪 (0)

Line 4 Field 1 specifies the **number of Cartesian** (integer) modes to process 🡪 20

Field 2 specifies the **number of Distance** (integer) modes to process 🡪 0 = none

Field 3 specifies the **number of Cartesian modes to Visualize**(integer) 🡪 2 = top two modes

Field 4 specifies the **scale factor** for the visualizations (double) 🡪 1.0 = default value

Line 5 specifies the **working directory** (String)

Line 6 Field 1 specifies the job **description** (String)

Field 2 specifies the **reference PDB** (String)

Line 7 specifies the **residue list** (String)

Line 8 Field 1 specifies the **coordinate matrix** (String)

Field 2 specifies the **reference column** (integer)

------------------------------------------------------------------------------------------

**Warning**:

The read PDB file flag must be set to 0.

The Multi flag must be set to 1 for JED to perform the analysis on Multi Chain PDBs.

The number of cPCA modes must be **>** 0.

There must be a Cartesian residue list specified.

The number of dPCA modes must be 0.

There must not be a distance residue list specified.

If the subset you have chosen contains N residues, then you must not request more than 3N modes.

If you request ***N*** cPCA modes then you can only visualize up to ***N*** top modes.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**The format of the residue list file is TWO columns for Multi Chain Analysis: Chain ID, residue number.**

**The format of the residue list file is ONE column for Single Chain Analysis: residue number.**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**If number of cPCA modes > 0, then there must be a Cartesian residue list file!**

**D. Performing Distance PCA**

**i. run command:**

java -jar -d64 JED\_Driver.jar “/path/JED\_Driver.txt”

***The working directory must contain: The coordinates matrix, the PDB reference file, and the residue list.***

***JED\_Driver. txt or user-defined file name must be in the working directory.***

The purpose of this type of run is to perform Essential Dynamics using dPCA based on **Q** and **R**. The user specifies the subset of interest for the analysis, which is typically less than 10 residues, by providing a residue list file. This task is simplified since JED has already created a list of all the residues in the protein. The user must simply edit this file. The dPCA results are written to the sub-directory "dPCA". Note that for dPCA no transform is needed since internal distances are used for coordinates and no visualization can be done in JED for the distance modes. The directory dPCA has sub directories for the **Q** and **R** analysis as well as for the subspace analysis (SSA). Choosing more than ten residues for the dPCA analysis makes the interpretation of the results challenging as each component of the distance eigenvectors corresponds to an inter-residue distance pair. Often subsets with three or four residues can be used to investigate experimental findings in critical areas like binding pockets or clefts.

**ii. Standard Output Files:**

These are written to the **root** of the JED Results directory tree:

/working/directory/JED\_Results\_Description/

**JED LOG** providing a summary of the job parameters and results:

JED\_Log.txt

**subset transformed coordinates matrix**:

ss\_**$num**\_res\_transformed\_PDB\_coordinates.txt

**original and transformed conformation RMSDs**:

ss\_**$num**\_res\_original\_conformation\_rmsds.txt

ss\_**$num**\_res\_conformation\_rmsds.txt

**residue RMSDs (RMSF)**:

ss\_**$num**\_res\_residue\_rmsd.txt

**subset edited PDB file** with the RMSF replacing B-factors:

ss\_**$num**\_res\_edidited.PDB

**subset coordinates Z-Score matrix**:

ss\_**$num**\_res\_coordinate\_Z\_scores.txt

**list of frames removed**, based on the **percent** parameter:

ss\_**$num**\_res\_ Removed\_Conformation\_Outliers.txt

**trimmed transformed coordinate matrix**:

ss\_**$num**\_res\_ trimmed\_$percent\_percent\_PDB\_coordinates\_COLS.txt

**list of coordinate variables adjusted**, based on the **z-cutoff** parameter:

ss\_**$num**\_res\_ adjustments\_per\_variable.txt

**adjusted transformed coordinate matrix**:

ss\_**$Z**\_threshold\_**$z**-cutoff\_adjusted\_PDB\_coordinates\_ROWS.txt

**iii. Standard dPCA Output Files:**

These are written to the **/dPCA subdirectory**  of the JED Results directory tree:

/working/directory/JED\_Results\_Description/dPCA/

**subset distance residue stats** from all the alpha carbon coordinates in the **subset**:

ss\_**$num**\_res\_distance\_residue\_stats.txt

**subset distance Z-Score matrix** from all the alpha carbon coordinates in the **subset**:

ss\_**$num**\_res\_distance\_Z\_scores.txt

**subset all-to-all** **distances matrix** from all the alpha carbon coordinates in the **subset**:

ss\_**$num**\_res\_all\_to\_all\_distances.txt

**subset distance variables** adjusted:

ss\_**$num**\_res\_ outliers\_per\_variable.txt

**subset centroids (means) of the variables:**

ss\_**$num**\_res\_centroids\_of\_variables.txt

**subset standard deviations of the variables:**

ss\_**$num**\_res\_std\_devs\_of\_centered\_variables.txt

**subset displacement vectors:**

ss\_**$num**\_res\_delta\_vectors.txt

**The COV output files are written to the /COV subdirectory of /dPCA**

/working/directory/JED\_Results\_Description/cPCA/COV/

**The CORR output files are written to the /CORR subdirectory of /dPCA**

**covariance matrix:**

ss\_**$num**\_res\_distance\_COV\_matrix.txt

**eigenvalues:**

ss\_**$num**\_res\_distance\_eigenvalues\_COV.txt

**top eigenvalues:**

ss\_**$num**\_res\_top\_**$num**\_of\_dist\_modes\_distance\_eigenvalues\_COV.txt

**top eigenvectors:**

ss\_**$num**\_res\_top\_**$num**\_of\_dist\_modes\_distance\_eigenvectors\_COV.txt

**top pca modes and top weighted pca modes:**

ss\_**$num**\_res\_top\_**$num**\_of\_dist\_modes\_distance\_pca\_modes\_COV.txt

ss\_**$num**\_res\_top\_**$num**\_of\_dist\_modes\_weighted\_distance\_pca\_modes\_COV.txt

**top square pca modes and top weighted square pca modes:**

ss\_**$num**\_res\_top\_**$num**\_of\_dist\_modes\_square\_distance\_pca\_modes\_COV.txt

ss\_**$num**\_res\_top\_**$num**\_of\_dist\_modes\_weighted\_square\_distance\_pca\_modes\_COV.txt

**top PCs and top weighted PCs:**

ss\_**$num**\_res\_top\_**$num**\_of\_dist\_modes\_PCs\_COV.txt

ss\_**$num**\_res\_top\_**$num**\_of\_dist\_modes\_weighted\_PCs\_COV.txt

**top normed PCs and top weighted normed PCs:**

ss\_**$num**\_res\_top\_**$num**\_of\_dist\_modes\_normed\_PCs\_COV.txt

ss\_**$num**\_res\_top\_**$num**\_of\_dist\_modes\_weighted\_normed\_PCs\_COV.txt

**iv. Standard SSA Output Files:**

These are written to the **/SSA subdirectory**  of /cPCA:

/working/directory/JED\_Results\_Description/cPCA/SSA/

**The Fast SSA Iterated Log:**

JED\_FSSA\_Iterated\_log.txt

**The SSA Log:**

JED\_SSA\_dim\_$top\_num\_cart\_modes\_log.txt

**The Random SSA Log**

JED\_Random\_SSA\_log.txt

**There are many more files in the /SSA directory that are flat files of the results reported in the log files:**

**RMSIPs**

**PAs**

**COs**

**Cosine Products**

**Vectorial Sum of Angles**

**v. Below is a sample input file for Single-Chain PDB files:**

------------------------------------------------------------------------------------------

0.00

3.00

0 0

0 3 0

/working/directory/

Description reference\_PDB\_file.pdb

residues\_dist.txt

original\_PDB\_Coordinates.txt 0

------------------------------------------------------------------------------------------

Notes:

Line 1 specifies the **percent** (double) of frames to remove from the data: 0 = none

Line 2 specifies the **z-cutoff** (double) for adjusting outliers: 0 🡪 values beyond |3.0|

Line 3 Token 1 specifies whether to **read PDB files** (0 or 1) 🡪 0 = no

Line 3 Token 2 specifies if the PDB files are **Multi** Chain (0 or 1) 🡪 0 = no

Line 4 Token 1 specifies the **number of Cartesian** (integer) modes to process 🡪 0 = none

Line 4 Token 2 specifies the **number of Distance** (integer) modes to process 🡪 3 = top three modes

Line 4 Token 3 specifies the **number of Cartesian modes to Visualize**(integer) 🡪 0 = none

Line 5 specifies the **working directory** (String)

Line 6 Token 1 specifies the job **description** (String)

Line 6 Token 2 specifies the **reference PDB** (String)

Line 7 specifies the **residue list** (String)

Line 8 Token 1 specifies the **coordinate matrix** (String)

Line 8 Token 2 specifies the **reference column** (integer)

------------------------------------------------------------------------------------------

**Warning**:

Both the read PDB file flag and the Multi flag MUST be set to ZERO.

This tells JED to perform the analysis on Single Chain PDBs.

The number of cPCA modes MUST be zero.

There must be no Cartesian residue list specified.

The number of dPCA modes MUST NOT be zero.

There must be a Distance residue list specified.

If the subset you have chosen contains ***N*** residues, then you must NOT request more than ***N(N-1)/2*** modes.

**If number of dPCA modes > 0, then there must be a Distance residue list file!**

**vi. Below is a sample input file for Multi-Chain PDB files:**

**Differences from Single-Chain analysis shown highlighted in amber**

------------------------------------------------------------------------------------------

0.00

3.00

0 1 **2 A B 795 151 0 0**

0 3 0

/working/directory/

Description reference\_PDB\_file.pdb

residues.txt

original\_PDB\_Coordinates.txt 0

------------------------------------------------------------------------------------------

Notes:

Line 1 specifies the **percent** (double) of frames to remove from the data: 0 🡪 none

Line 2 specifies the **z-cutoff** (double) for adjusting outliers: 0 🡪 values beyond |3.0|

Line 3 Token 1 specifies whether to **read PDB files** (0 or 1) 🡪 0 = no

Line 3 Token 2 specifies if the PDB files are **Multi** Chain (0 or 1) 🡪 1 = yes

Line 3 Token 3 specifies the number of chains 🡪 (2);

Line 3 Token 4 specifies the first chain ID 🡪 (A);

Line 3 Token 5 specifies the second chain ID 🡪 (B);

Line 3 Token 6 specifies the number of residues in chain A🡪 (795);

Line 3 Token 7 specifies the number of residues in chain B🡪 (151);

Line 3 Token 8 specifies the offset of Chain A 🡪 (0);

Line 3 Token 9 specifies the offset of Chain B 🡪 (0);

Line 4 Token 1 specifies the **number of Cartesian** (integer) modes to process 🡪 0

Line 4 Token 2 specifies the **number of Distance** (integer) modes to process 🡪 3 = Top 3 modes

Line 4 Token 3 specifies the **number of Cartesian modes to Visualize**(integer) 🡪 0

Line 5 specifies the **working directory** (String)

Line 6 Token 1 specifies the job **description** (String)

Line 6 Token 2 specifies the **reference PDB** (String)

Line 7 specifies the **residue list** (String)

Line 8 Token 1 specifies the **coordinate matrix** (String)

Line 8 Token 2 specifies the **reference column** (integer)

------------------------------------------------------------------------------------------------------------------------------------------------------

**Warning**:

The read PDB file flag MUST be set to ZERO.

The multi flag must be set to ONE;

This tells JED to perform the analysis on MULTI Chain PDBs.

The number of cPCA modes MUST be zero.

There must be no Cartesian residue list specified.

The number of dPCA modes MUST NOT be zero.

There must be a Distance residue list specified.

If the subset you have chosen contains ***N*** residues, then you must NOT request more than ***N(N-1)/2*** modes.

Remember that the format of the residue list file is TWO columns for Multi Chain Analysis: **Chain ID, residue** **number** while for Single Chain Analysis, the file has ONE column: **residue number**.

**If number of dPCA modes > 0, then there must be a Distance residue list file!**

**E. Debugging Crashes Part II:**

Things that will generally make your life miserable…

**i. Dumb mistakes:**

Did you set the Read and Multi flags correctly? If not, CRASH!!!

Did you request cPCA but not specify a Cartesian residue list? If so, CRASH!!!

Did you set the number of modes correctly? If not, CRASH!!!

Are the 0 and 1 flags set correctly? If not CRASH!!!

Are you requesting to read PDBs when you should be specifying a coordinate matrix? If so CRASH!!!

**ii. More subtle mistakes:**

If the PDB file residue numbering starts at 5, then the chain offset is not 0, it is 4 Garbage out!

Did you request more modes than actually exist? For example, if your Cartesian subset contains 12 residues, but you asked for 50 modes, then you are going to crash JED because there are only 36 Cartesian modes in total.

Also, if your Distance subset contains 3 residues and you request 5 modes, then you are going to crash JED because there are only 3 distance modes in total.

If your trajectory has not equilibrated, then you must address the problem or outliers. If you do not, then the Q and R matrices will be highly ill-conditioned and may cause the eigenvalue decomposition to fail. You can check the variables in statistics packages that compute the Kaiser-Myer-Olkin (KMO) statistic as well as the Measure of Sampling Adequacy (MSA) for each coordinate variable to critically assess your data. If it is not well suited for PCA, you can condition the variables by setting the z-cutoff in JED between 2.0 and 3.0 when running your jobs. This type of conditioning is by far not very sophisticated, but it has the effect of lowering the condition numbers of Q and R as well as un-dilating the high and low regions of the eigenspectrum. In particular, it does not alter the ordinality of the eigenspectrum but does correct the distortion that arises from under sampling when trying to estimate the population covariance matrix from the sample covariance matrix.

**F. Performing Cartesian and Distance PCA with Mode Visualization**

***The working directory must contain: The coordinates matrix, the PDB reference file, and both residue lists.***

***It may also contain JED\_Driver.txt***

JED is capable of doing both cPCA and dPCA, using both Q and R, and generating cPCA mode visualizations simultaneously. All outputs are delivered as discussed for the individual components.

The outputs to this type of job include the outputs for both the cPCA and dPCA analyses, as well as all the structures for the top modes chosen for visualization. JED will permute the reference structure for a given subset along the top eigenvectors selected for visualization and output structures (PDBs) that capture one cycle of this motion. The amplitude of the motion is determined by the value of the **$scale\_factor**, whose default value is 1.0, and can be adjusted as necessary. Setting the value too high will cause Visualization software like Pymol to break the ribbon diagrams of the structures. Ultimately, this is controlled by the magnitude of the eigenvector components for any given residue. Setting the scale factor between 1 and 3 is usually safe, but for proteins with highly mobile regions like loops, you may need to choose a smaller scale factor. This is done for both the top Q and R modes. Additionally, Pymol scripts are generated to animate those structures into a movie for better analysis of the physical meanings of the top modes.

These files will be located in the /mode-viz subdirectory of the root of the JED results tree:

/working/directory/JED\_Results\_Description/mode\_viz/

The Q results will be in the subdirectory /COV and the R results will be in the subdirectory /CORR.

One huge advantage of JED is that it is highly configurable and can perform many types of Essential Dynamics analysis concurrently. Combined with the cluster resources or just using the batch feature allows a user to process a great deal of data efficiently.

**i. Below is a sample input file for Single-Chain PDB files:**

------------------------------------------------------------------------------------------

0.00

3.00

0 0

20 3 2 1.0

/working/directory/

Description reference\_PDB\_file.pdb

residues\_cartesian.txt

residues\_dist.txt

original\_PDB\_Coordinates.txt 0

------------------------------------------------------------------------------------------

Notes:

Line 1 specifies the **percent** (double) of frames to remove from the data: 0 = none

Line 2 specifies the **z-cutoff** (double) for adjusting outliers: 0 🡪 values beyond |3.0|

Line 3 Token 1 specifies whether to **read PDB files** (0 or 1) 🡪 0 = no

Line 3 Token 2 specifies if the PDB files are **Multi** Chain (0 or 1) 🡪 0 = no

Line 4 Token 1 specifies the **number of Cartesian** (integer) modes to process 🡪 0 = none

Line 4 Token 2 specifies the **number of Distance** (integer) modes to process 🡪 3 = top three modes

Line 4 Token 3 specifies the **number of Cartesian modes to Visualize**(integer) 🡪 0 = none

Line 5 specifies the **working directory** (String)

Line 6 Token 1 specifies the job **description** (String)

Line 6 Token 2 specifies the **reference PDB** (String)

Line 7 specifies the **Cartesian** **residue list** (String)

Line 7 specifies the **Distance** **residue list** (String)

Line 8 Token 1 specifies the **coordinate matrix** (String)

Line 8 Token 2 specifies the **reference column** (integer)

------------------------------------------------------------------------------------------

**Warning**:

Both the read PDB file flag and the Multi flag MUST be set to ZERO.

The number of cPCA modes MUST be > zero.

The number of dPCA modes MUST be > zero.

There must be a Cartesian residue list specified.

There must be a Distance residue list specified.

They must be in the CORRECT ORDER: Cartesian first, then Distance

For Cartesian subsets containing ***N*** residues, you must NOT request more than **3*N*** modes.

For Distance subsets containing ***N*** residues, you must NOT request more than ***N(N-1)/2*** modes.

**ii. Below is a sample input file for Multi-Chain PDB files:**

**Differences from Single-Chain analysis shown highlighted in amber**

------------------------------------------------------------------------------------------

0.00

3.00

0 1 **2 A B 795 151 0 0**

20 3 2 1.0

/working/directory/

Description reference\_PDB\_file.pdb

residues\_cartesian.txt

residues\_dist.txt

original\_PDB\_Coordinates.txt 0

------------------------------------------------------------------------------------------

Notes:

Line 1 specifies the **percent** (double) of frames to remove from the data: 0 🡪 None

Line 2 specifies the **z-cutoff** (double) for adjusting outliers: 0 🡪 values beyond |3.0|

Line 3 Token 1 specifies whether to **read PDB files** (0 or 1) 🡪 0 = no

Line 3 Token 2 specifies if the PDB files are **Multi** Chain (0 or 1) 🡪 1 = yes

Line 3 Token 3 specifies the number of chains 🡪 (2);

Line 3 Token 4 specifies the first chain ID 🡪 (A);

Line 3 Token 5 specifies the second chain ID 🡪 (B);

Line 3 Token 6 specifies the number of residues in chain A🡪 (795);

Line 3 Token 7 specifies the number of residues in chain B🡪 (151);

Line 3 Token 8 specifies the offset of Chain A 🡪 (0);

Line 3 Token 9 specifies the offset of Chain B 🡪 (0);

Line 4 Token 1 specifies the **number of Cartesian** (integer) modes to process 🡪 20 = top twenty modes

Line 4 Token 2 specifies the **number of Distance** (integer) modes to process 🡪 3 = top three modes

Line 4 Token 3 specifies the **number of Cartesian modes to Visualize**(integer) 🡪 2 = top two modes

Line 5 specifies the **working directory** (String)

Line 6 Token 1 specifies the job **description** (String)

Line 6 Token 2 specifies the **reference PDB** (String)

Line 7 specifies the **Cartesian** **residue list** (String)

Line 7 specifies the **Distance** **residue list** (String)

Line 8 Token 1 specifies the **coordinate matrix** (String)

Line 8 Token 2 specifies the **reference column** (integer)

------------------------------------------------------------------------------------------

**Warning**:

The read PDB file flag MUST be set to ZERO.

The multi flag must be set to ONE; This tells JED to perform the analysis on MULTI Chain PDBs.

The number of cPCA modes MUST be > zero.

The number of dPCA modes MUST be > zero.

There must be a Cartesian residue list specified.

There must be a Distance residue list specified.

They must be in the CORRECT ORDER: Cartesian first, then Distance

For Cartesian subsets containing ***N*** residues, you must NOT request more than **3*N*** modes.

For Distance subsets containing ***N*** residues, you must NOT request more than ***N(N-1)/2*** modes.

Remember that the format of the residue list file is TWO columns for Multi Chain Analysis: **Chain ID, residue** **number** while for Single Chain Analysis, the file has ONE column: **residue number**.

**If number of cPCA modes > 0, then there must be a Cartesian residue list file!**

**If number of dPCA modes > 0, then there must be a Distance residue list file!**

**V. USING JED BATCH DRIVER**

**A. The Preliminary Run**

**i. run command:**

java -jar -d 64 JED\_Batch\_Driver.jar “/path/JED\_Batch\_Driver.txt”

***The PDB files (including the PDB reference file) must be in the working directory.***

***JED\_Driver.txt may be in the working directory.***

This pre-processing step will read all PDB files in the working directory, but will perform NO PCA.

The purpose of this is to generate the matrix of coordinates for performing subset analyses efficiently.

**ii. Standard NO-PCA Output Files:**

These are written to the **root** of the JED Results directory tree:

/working/directory/JED\_Results\_Description/

**JED LOG** providing a summary of the job parameters and results:

JED\_Log.txt

**PDB READ LOG** listing all the PDB files read, in the order they were read:

PDB\_READ\_Log.txt

**coordinates matrix** from all the alpha carbon coordinates in the PDB files:

original\_PDB\_coordinates.txt

**transformed coordinates matrix**, which aligns all the frames to the reference frame :

ss\_$num\_res\_transformed\_PDB\_coordinates.txt

**list of all residues** found in the PDB files for subsequent editing and use:

All\_PDB\_Residues\_JED.txt

All\_PDB\_Residues\_Multi\_JED.txt (for Multi runs)

**original and transformed conformation RMSDs**:

ss\_$num\_res\_original\_conformation\_rmsds.txt

ss\_$num\_res\_conformation\_rmsds.txt

**residue RMSDs (RMSF)**:

ss\_$num\_res\_residue\_rmsd.txt

**edited PDB file** containing all the residues with the RMSF replacing B-factors:

ss\_$num\_res\_edidited.PDB

**coordinates Z-Score matrix** from all the alpha carbon coordinates in the PDB files:

ss\_$num\_res\_coordinate\_Z\_scores.txt

**percent** of the frames to remove based on conformation RMSD (OPTIONAL)

**z-cutoff** for adjusting coordinate outliers to their mean values (OPTIONAL)

If these parameters are set to a value other than zero, additional output files are generated with the results.

However, the appropriate time to handle outliers is during runs that actually perform PCA and it is recommended that for pre-processing runs, these be set to zero.

**iii. Below is a sample input file for Single Chain PDB files:**

-----------------------------------------------------------------------------------------------------

$num\_of\_jobs

0.00

0.00

1 0

0 0 0

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/working/directory/job1/

Description1 reference\_PDB\_file1.pdb

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/working/directory/job2/

Description2 reference\_PDB\_file2.pdb

-----------------------------------------------------------------------------------------------------

Notes: This is a tab (space) separated file. There must be NO LEADING blank lines or spaces (no trailing too).

Line 1 specifies the **$num\_of\_jobs** (integer) for the batch.

Line 2 specifies the **percent** (double) of frames to remove from the data: 0 🡪NONE

Line 3 specifies the **z-cutoff** (double) for adjusting outliers: 0 🡪NO ADJUSTMENT

Line 4 Token 1 specifies whether to **read PDB files** (0 or 1) 🡪1 = yes

Line 4 Token 2 specifies if the PDB files are **Multi** Chain (0 or 1) 🡪 0 = no

Line 5 Token 1 specifies the **number of Cartesian** (integer) modes to process 🡪0 = none

Line 5 Token 2 specifies the **number of Distance** (integer) modes to process 🡪0 = none

Line 5 Token 3 specifies the **number of Cartesian modes to Visualize**(integer) 🡪0 = none

Line 6 is a separator line between the batch parameters and the job parameters

Line 7 specifies the **working directory** (String)

Line 8 Token 1 specifies the job **description** (String) for job one

Line 8 Token 2 specifies the **reference PDB** (String) for job one

Line 9 is a separator line between sets of job parameters

Line 10 Token 1 specifies the job **description** (String) for job two

Line 10 Token 2 specifies the **reference PDB** (String) for job two

-----------------------------------------------------------------------------------------------------

**Warning**:

Both number of modes Cartesian and number of modes Distance MUST be set to ZERO.

This tells JED to perform the NO-PCA analysis.

The Read flag MUST be set to ONE and the MULTI flag MUST be set to ZERO.

All PDB files MUST have the same number of residues. The matrix of alpha carbon coordinates is determined from the first PDB file read. If other files in the working directory do not match exactly, then the array sizes will not match and the program will crash. IF JED crashes during the reading of PDBs, this is probably the reason.

In subsequent analyses, it is critical that no residues are requested that do not actually exist in the PDB files!

JED maps the specified residue list to an internal list that is aligned to the columns of the coordinates matrix.

**The file to use in all subsequent JED analyses is the original\_PDB\_coordinates matrix.**

This matrix contains all the residues in the PDB files and thus can be used for any subset of those residues. When a subset is chosen, a new correspondence set is generated and a new transformation is done to optimize the alignment of the structures. This removes overall translation and rotation.

**iv. Below is a sample input file for Multi-Chain PDB files:**

**Differences from Single-Chain analysis shown highlighted in amber**

-----------------------------------------------------------------------------------------------------

$num\_of\_jobs

0.00

0.00

1 **1 2 A B 795 151 0 0**

0 0 0

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/working/directory/job1/

Description1 reference\_PDB\_file1.pdb

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/working/directory/job2/

Description2 reference\_PDB\_file2.pdb

-----------------------------------------------------------------------------------------------------

Notes: This is a space delineated file. There must be NO LEADING blank lines or spaces (no trailing too).

Line 1 specifies the **$num\_of\_jobs** (integer) for the batch.

Line 2 specifies the **percent** (double) of frames to remove from the data: 0 🡪NONE

Line 3 specifies the **z-cutoff** (double) for adjusting outliers: 0 🡪NO ADJUSTMENT

Line 4 Token 1 specifies whether to **read PDB files** (0 or 1) 🡪1 = yes

Line 4 Token 2 specifies if the PDB files are **Multi** Chain (0 or 1) 🡪 1 = yes

Line 4 Token 3 specifies the number of chains 🡪 (2);

Line 4 Token 4 specifies the first chain ID 🡪 (A);

Line 4 Token 5 specifies the second chain ID 🡪 (B);

Line 4 Token 6 specifies the number of residues in chain A🡪 (795);

Line 4 Token 7 specifies the number of residues in chain B🡪 (151);

Line 4 Token 8 specifies the offset of Chain A 🡪 (0);

Line 4 Token 9 specifies the offset of Chain B 🡪 (0);

Line 5 Token 1 specifies the **number of Cartesian** (integer) modes to process 🡪0 = none

Line 5 Token 2 specifies the **number of Distance** (integer) modes to process 🡪0 = none

Line 5 Token 3 specifies the **number of Cartesian modes to Visualize**(integer) 🡪0 = none

Line 6 is a separator line between the batch parameters and the job parameters

Line 7 specifies the **working directory** (String)

Line 8 Token 1 specifies the job **description** (String) for job one

Line 8 Token 2 specifies the **reference PDB** (String) for job one

Line 9 is a separator line between sets of job parameters

Line 10 Token 1 specifies the job **description** (String) for job two

Line 10 Token 2 specifies the **reference PDB** (String) for job two

-----------------------------------------------------------------------------------------------------

**Warning**:

Both number of modes Cartesian and number of modes Distance MUST be set to ZERO.

This tells JED to perform the NO-PCA analysis.

The Read flag MUST be set to ONE and the MULTI flag MUST be set to ONE.

All PDB files MUST have the same number of residues. The matrix of alpha carbon coordinates is determined from the first PDB file read. If other files in the working directory do not match exactly, then the array sizes will not match and the program will crash. IF JED crashes during the reading of PDBs, this is probably the reason.

Additionally, Multi Chain PDBs MUST have unique chain identifiers for every chain present in the file. A missing chain identifier will cause JED to crash.

In subsequent analyses, it is critical that no residues are requested that do not actually exist in the PDB files!

JED maps the specified residue list to an internal list that is aligned to the columns of the coordinates matrix.

**The file to use in all subsequent JED analyses is the original\_PDB\_coordinates matrix.**

This matrix contains all the residues in the PDB files and thus can be used for any subset of those residues. When a subset is chosen, a new correspondence set is generated and a new transformation is done to optimize the alignment of the structures. This removes overall translation and rotation.

**B. Debugging Crashes Part I:**

Things that will generally make your life miserable…

**i. Dumb mistakes:**

Does the directory exist? If not, CRASH!!!

Does the path to the input file exist? If not, CRASH!!!

Does the input file exist in the proper location? If not, CRASH!!!

Does the input-file start on the first line? If not, CRASH!!!

The directory contains PDB files of different sizes CRASH!!!

The directory does not contain the reference PDB file CRASH!!!

Is that integer really an integer or is it a float or a string? CRASH!!!

Does the directory string end in “/” or “\\”?: If not CRASH!!!

Are the 0 and 1 flags set correctly? If not CRASH!!!

**ii. More subtle mistakes:**

The directory contains PDB files with 2 chains, but no chain IDs Non-sense results!

If the PDB file names are sorted in a different order than how they were generated, then the conformation RMSD results will not reflect what actually occurred in the simulation. This can be prevented by naming the PDB files appropriately, specifically, padding the numbers with sufficient leading zeros to ensure proper sorting. Note that this is not just a Java thing, it is a Unix thing too.

If the conformation RMSD is very different from what you expect, then you may be using PDB files that contain occupancy information. JED does not use that information. Your results will not be accurate.

If you have pooled data, make sure the combined matrix is accurate.

**C. Performing Cartesian PCA**

**i. run command:**

java -jar -d 64 JED\_Batch\_Driver.jar “/path/JED\_Batch\_Driver.txt”

***The working directory must contain: The coordinates matrix, the PDB reference file, and the residue list.***

***It may also contain JED\_Driver.txt***

The purpose of this type of run is to perform Essential Dynamics using cPCA based on Q and R.

The user specifies the subset of interest for the analysis, which may be the entire protein or a small, non-contiguous selection of residues, by providing a residue list file. This task is simplified since JED has already created a list of all the residues in the protein. The user must simply edit this file. The cPCA results are written to the sub-directory "cPCA" and the Visualizations of the top modes (when selected) are written to the subdirectory "viz". The directory cPCA has sub directories for the Q and R analysis, as does the viz directory.

**ii. Standard Output Files:**

These are written to the **root** of the JED Results directory tree:

/working/directory/JED\_Results\_Description/

**JED LOG** providing a summary of the job parameters and results:

JED\_Log.txt

**subset transformed coordinates matrix**:

ss\_$num\_res\_transformed\_PDB\_coordinates.txt

**original and transformed conformation RMSDs**:

ss\_$num\_res\_original\_conformation\_rmsds.txt

ss\_$num\_res\_conformation\_rmsds.txt

**residue RMSDs (RMSF)**:

ss\_$num\_res\_residue\_rmsd.txt

**subset edited PDB file** with the RMSF replacing B-factors:

ss\_$num\_res\_edidited.PDB

**subset coordinates Z-Score matrix**:

ss\_$num\_res\_coordinate\_Z\_scores.txt

**list of frames removed**, based on the **percent** parameter:

ss\_$num\_res\_ Removed\_Conformation\_Outliers.txt

**trimmed transformed coordinate matrix**:

ss\_$num\_res\_ trimmed\_$percent\_percent\_PDB\_coordinates\_COLS.txt

**list of coordinate variables adjusted**, based on the **z-cutoff** parameter:

ss\_$num\_res\_ adjustments\_per\_variable.txt

**adjusted transformed coordinate matrix**:

ss\_$Z\_threshold\_$z-cutoff\_adjusted\_PDB\_coordinates\_ROWS.txt

**iii. Standard cPCA Output Files:**

These are written to the **/cPCA subdirectory**  of the JED Results directory tree:

/working/directory/JED\_Results\_Description/cPCA/

**centroids (means) of the variables:**

ss\_$num\_res\_centroids\_of\_variables.txt

**standard deviations of the variables:**

ss\_$num\_res\_std\_devs\_of\_centered\_variables.txt

**displacement vectors:**

ss\_$num\_res\_delta\_vectors.txt

**The COV output files are written to the /COV subdirectory of /cPCA**

/working/directory/JED\_Results\_Description/cPCA/COV/

**The CORR output files are written to the /CORR subdirectory of /cPCA**

/working/directory/JED\_Results\_Description/cPCA/CORR/

**covariance matrix:**

ss\_$num\_res\_COV\_matrix.txt

**eigenvalues:**

ss\_$num\_res\_eigenvalues\_COV.txt

**top eigenvalues:**

ss\_$num\_res\_top\_$num\_of\_cart\_modes\_eigenvalues\_COV.txt

**top eigenvectors:**

ss\_$num\_res\_top\_$num\_of\_cart\_modes\_eigenvectors\_COV.txt

**top pca modes and top weighted pca modes:**

ss\_$num\_res\_top\_$num\_of\_cart\_modes\_pca\_modes\_COV.txt

ss\_$num\_res\_top\_$num\_of\_cart\_modes\_weighted\_pca\_modes\_COV.txt

**top square pca modes and top weighted square pca modes:**

ss\_$num\_res\_top\_$num\_of\_cart\_modes\_square\_pca\_modes\_COV.txt

ss\_$num\_res\_top\_$num\_of\_cart\_modes\_weighted\_square\_pca\_modes\_COV.txt

**top PCs and top weighted PCs:**

ss\_$num\_res\_top\_$num\_of\_cart\_modes\_PCs\_COV.txt

ss\_$num\_res\_top\_$num\_of\_cart\_modes\_weighted\_PCs\_COV.txt

**top PCs and top weighted PCs:**

ss\_$num\_res\_top\_$num\_of\_cart\_modes\_normed\_PCs\_COV.txt

ss\_$num\_res\_top\_$num\_of\_cart\_modes\_weighted\_normed\_PCs\_COV.txt

**iv. Standard SSA Output Files:**

These are written to the **/SSA subdirectory**  of /cPCA:

/working/directory/JED\_Results\_Description/cPCA/SSA/

**The Fast SSA Iterated Log:**

JED\_FSSA\_Iterated\_log.txt

**The SSA Log:**

JED\_SSA\_dim\_$top\_num\_cart\_modes\_log.txt

**The Random SSA Log**

JED\_Random\_SSA\_log.txt

**There are many more files in the /SSA directory that are flat files of the results reported in the log files:**

**RMSIPs**

**PAs**

**COs**

**Cosine Products**

**Vectorial Sum of Angles**

**v. Below is a sample input file for Single-Chain PDB files:**

------------------------------------------------------------------------------------------

$num\_of\_jobs

0.01

3.00

0 0

20 0 2 1.0

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/working/directory1/

Description1 reference\_PDB\_file1.pdb

residues1.txt

original\_PDB\_Coordinates.txt 0

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/working/directory2/

Description2 reference\_PDB\_file2.pdb

residues2.txt

original\_PDB\_Coordinates.txt 0

------------------------------------------------------------------------------------------

Notes:

Line 1 specifies the **$num\_of\_jobs** (integer) for the batch

Line 2 specifies the **percent** (double) of frames to remove from the data: 0 🡪 1%

Line 3 specifies the **z-cutoff** (double) for adjusting outliers: 0 🡪 values beyond |3.0|

Line 4 Token 1 specifies whether to **read PDB files** (0 or 1) 🡪 0 = no

Line 4 Token 2 specifies if the PDB files are **Multi** Chain (0 or 1) 🡪 0 = no

Line 5 Token 1 specifies the **number of Cartesian** (integer) modes to process 🡪 20

Line 5 Token 2 specifies the **number of Distance** (integer) modes to process 🡪 0 = none

Line 5 Token 3 specifies the **number of Cartesian modes to Visualize**(integer) 🡪 2 = top two modes

Line 5 Token 4 specifies the **scale factor** for the visualizations (double) 🡪 1.0 = default value

Line 6 is a separator line between the batch parameters and the job parameters

Line 7 specifies the **working directory** (String) for job one

Line 8 Token 1 specifies the job **description** (String) for job one

Line 8 Token 2 specifies the **reference PDB** (String) for job one

Line 9 specifies the **residue list** (String) for job one

Line 10 specifies the **coordinate matrix** (String) for job one

Line 11 is a separator line between sets of job parameters

Line 12 specifies the **working directory** (String) for job two

Line 13 Token 1 specifies the job **description** (String) for job two

Line 13 Token 2 specifies the **reference PDB** (String) for job two

Line 14 specifies the **residue list** (String) for job two

Line 15 specifies the **coordinate matrix** (String) for job two

------------------------------------------------------------------------------------------

**Warning**:

Both the Read PDB file flag and the Multi flag MUST be set to ZERO.

This tells JED to perform the analysis on Single Chain PDBs.

The number of cPCA modes MUST not be zero.

There must be a Cartesian residue list specified.

The number of dPCA modes MUST be zero.

There must NOT be a Distance residue list specified.

If the subset you have chosen contains N residues, then you must NOT request more than 3N modes.

If you request N cPCA modes then you can only visualize up to N top modes.

**vi. Below is a sample input file for Multi-Chain PDB files:**

**Differences from Single-Chain analysis shown highlighted in amber**

------------------------------------------------------------------------------------------

$num\_of\_jobs

0.01

3.00

0 1 **2 A B 795 151 0 0**

20 0 2 1.0

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/working/directory1/

Description1 reference\_PDB\_file1.pdb

residues1.txt

original\_PDB\_Coordinates.txt 0

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/working/directory2/

Description2 reference\_PDB\_file2.pdb

residues2.txt

original\_PDB\_Coordinates.txt 0

------------------------------------------------------------------------------------------

Notes:

Line 1 specifies the **$num\_of\_jobs** (integer) for the batch

Line 2 specifies the **percent** (double) of frames to remove from the data: 0 🡪 1%

Line 3 specifies the **z-cutoff** (double) for adjusting outliers: 0 🡪 values beyond |3.0|

Line 4 Token 1 specifies whether to **read PDB files** (0 or 1) 🡪 0 = no

Line 4 Token 2 specifies if the PDB files are **Multi** Chain (0 or 1) 🡪 1 = yes

Line 4 Token 3 specifies the number of chains 🡪 (2);

Line 4 Token 4 specifies the first chain ID 🡪 (A);

Line 4 Token 5 specifies the second chain ID 🡪 (B);

Line 4 Token 6 specifies the number of residues in chain A🡪 (795);

Line 4 Token 7 specifies the number of residues in chain B🡪 (151);

Line 4 Token 8 specifies the offset of Chain A 🡪 (0);

Line 4 Token 9 specifies the offset of Chain B 🡪 (0);

Line 5 Token 1 specifies the **number of Cartesian** (integer) modes to process 🡪 20

Line 5 Token 2 specifies the **number of Distance** (integer) modes to process 🡪 0 = none

Line 5 Token 3 specifies the **number of Cartesian modes to Visualize**(integer) 🡪 2 = top two modes

Line 5 Token 4 specifies the **scale factor** for the visualizations (double) 🡪 1.0 = default value

Line 6 is a separator line between the batch parameters and the job parameters

Line 7 specifies the **working directory** (String) for job one

Line 8 Token 1 specifies the job **description** (String) for job one

Line 8 Token 2 specifies the **reference PDB** (String) for job one

Line 9 specifies the **residue list** (String) for job one

Line 10 specifies the **coordinate matrix** (String) for job one

Line 11 is a separator line between sets of job parameters

Line 12 specifies the **working directory** (String) for job two

Line 13 Token 1 specifies the job **description** (String) for job two

Line 13 Token 2 specifies the **reference PDB** (String) for job two

Line 14 specifies the **residue list** (String) for job two

Line 15 specifies the **coordinate matrix** (String) for job two

------------------------------------------------------------------------------------------

**Warning**:

The read PDB file flag MUST be set to ZERO.

The Multi flag MUST be set to ONE

This tells JED to perform the analysis on Multi Chain PDBs.

The number of cPCA modes MUST be **>** zero.

There must be a Cartesian residue list specified.

The number of dPCA modes MUST be zero.

There must NOT be a Distance residue list specified.

If the subset you have chosen contains N residues, then you must NOT request more than 3N modes.

If you request ***N*** cPCA modes then you can only visualize up to ***N*** top modes.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**The format of the residue list file is TWO columns for Multi Chain Analysis: Chain ID, residue number.**

**The format of the residue list file is ONE column for Single Chain Analysis: residue number.**

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

**If number of cPCA modes > 0, then there must be a Cartesian residue list file!**

**D. Performing Distance PCA**

**i. run command:**

java -jar -d 64 JED\_Batch\_Driver.jar “/path/JED\_Batch\_Driver.txt”

***The working directory must contain: The coordinates matrix, the PDB reference file, and the residue list.***

***It may also contain JED\_Driver.txt***

The purpose of this type of run is to perform Essential Dynamics using dPCA based on Q and R.

The user specifies the subset of interest for the analysis, which is typically less than 10 residues, by providing a residue list file. This task is simplified since JED has already created a list of all the residues in the protein. The user must simply edit this file. The dPCA results are written to the sub-directory "dPCA". Note that for dPCA no transform is needed since internal distances are used for coordinates and no visualization can be done in JED for the distance modes. The directory dPCA has sub directories for the Q and R analysis as well as for the subspace analysis (SSA).

Note that choosing more than ten residues for the dPCA analysis makes the interpretation of the results quite challenging as each component of the distance eigenvectors corresponds to an inter-residue distance pair. Often subsets with three or four residues can be used to investigate experimental findings in critical areas like binding pockets or clefts.

**ii. Standard Output Files:**

These are written to the **root** of the JED Results directory tree:

/working/directory/JED\_Results\_Description/

**JED LOG** providing a summary of the job parameters and results:

JED\_Log.txt

**subset transformed coordinates matrix**:

ss\_$num\_res\_transformed\_PDB\_coordinates.txt

**original and transformed conformation RMSDs**:

ss\_$num\_res\_original\_conformation\_rmsds.txt

ss\_$num\_res\_conformation\_rmsds.txt

**residue RMSDs (RMSF)**:

ss\_$num\_res\_residue\_rmsd.txt

**subset edited PDB file** with the RMSF replacing B-factors:

ss\_$num\_res\_edidited.PDB

**subset coordinates Z-Score matrix**:

ss\_$num\_res\_coordinate\_Z\_scores.txt

**list of frames removed**, based on the **percent** parameter:

ss\_$num\_res\_ Removed\_Conformation\_Outliers.txt

**trimmed transformed coordinate matrix**:

ss\_$num\_res\_ trimmed\_$percent\_percent\_PDB\_coordinates\_COLS.txt

**list of coordinate variables adjusted**, based on the **z-cutoff** parameter:

ss\_$num\_res\_ adjustments\_per\_variable.txt

**adjusted transformed coordinate matrix**:

ss\_$Z\_threshold\_$z-cutoff\_adjusted\_PDB\_coordinates\_ROWS.txt

**iii. Standard dPCA Output Files:**

These are written to the **/dPCA subdirectory**  of the JED Results directory tree:

/working/directory/JED\_Results\_Description/dPCA/

**subset distance residue stats** from all the alpha carbon coordinates in the **subset**:

ss\_$num\_res\_distance\_residue\_stats.txt

**subset distance Z-Score matrix** from all the alpha carbon coordinates in the **subset**:

ss\_$num\_res\_distance\_Z\_scores.txt

**subset all-to-all** **distances matrix** from all the alpha carbon coordinates in the **subset**:

ss\_$num\_res\_all\_to\_all\_distances.txt

**subset distance variables** adjusted:

ss\_$num\_res\_ outliers\_per\_variable.txt

**subset centroids (means) of the variables:**

ss\_$num\_res\_centroids\_of\_variables.txt

**subset standard deviations of the variables:**

ss\_$num\_res\_std\_devs\_of\_centered\_variables.txt

**subset displacement vectors:**

ss\_$num\_res\_delta\_vectors.txt

**The COV output files are written to the /COV subdirectory of /dPCA**

/working/directory/JED\_Results\_Description/cPCA/COV/

**The CORR output files are written to the /CORR subdirectory of /dPCA**

**covariance matrix:**

ss\_$num\_res\_distance\_COV\_matrix.txt

**eigenvalues:**

ss\_$num\_res\_distance\_eigenvalues\_COV.txt

**top eigenvalues:**

ss\_$num\_res\_top\_$num\_of\_dist\_modes\_distance\_eigenvalues\_COV.txt

**top eigenvectors:**

ss\_$num\_res\_top\_$num\_of\_dist\_modes\_distance\_eigenvectors\_COV.txt

**top pca modes and top weighted pca modes:**

ss\_$num\_res\_top\_$num\_of\_dist\_modes\_distance\_pca\_modes\_COV.txt

ss\_$num\_res\_top\_$num\_of\_dist\_modes\_weighted\_distance\_pca\_modes\_COV.txt

**top square pca modes and top weighted square pca modes:**

ss\_$num\_res\_top\_$num\_of\_dist\_modes\_square\_distance\_pca\_modes\_COV.txt

ss\_$num\_res\_top\_$num\_of\_dist\_modes\_weighted\_square\_distance\_pca\_modes\_COV.txt

**top PCs and top weighted PCs:**

ss\_$num\_res\_top\_$num\_of\_dist\_modes\_PCs\_COV.txt

ss\_$num\_res\_top\_$num\_of\_dist\_modes\_weighted\_PCs\_COV.txt

**top normed PCs and top weighted normed PCs:**

ss\_$num\_res\_top\_$num\_of\_dist\_modes\_normed\_PCs\_COV.txt

ss\_$num\_res\_top\_$num\_of\_dist\_modes\_weighted\_normed\_PCs\_COV.txt

**iv. Standard SSA Output Files:**

These are written to the **/SSA subdirectory**  of /cPCA:

/working/directory/JED\_Results\_Description/cPCA/SSA/

**The Fast SSA Iterated Log:**

JED\_FSSA\_Iterated\_log.txt

**The SSA Log:**

JED\_SSA\_dim\_$top\_num\_cart\_modes\_log.txt

**The Random SSA Log**

JED\_Random\_SSA\_log.txt

**There are many more files in the /SSA directory that are flat files of the results reported in the log files:**

**RMSIPs**

**PAs**

**COs**

**Cosine Products**

**Vectorial Sum of Angles**

**v. Below is a sample input file for Single-Chain PDB files:**

------------------------------------------------------------------------------------------

$num\_of\_jobs

0.00

3.00

0 0

0 3 0

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/working/directory1/

Description1 reference\_PDB\_file1.pdb

residues\_dist1.txt

original\_PDB\_Coordinates.txt 0

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/working/directory2/

Description2 reference\_PDB\_file2.pdb

residues\_dist2.txt

original\_PDB\_Coordinates.txt 0

------------------------------------------------------------------------------------------

Notes:

Line 1 specifies the **$num\_of\_jobs** (integer) for the batch

Line 2 specifies the **percent** (double) of frames to remove from the data: 0 = none

Line 3 specifies the **z-cutoff** (double) for adjusting outliers: 0 🡪 values beyond |3.0|

Line 4 Token 1 specifies whether to **read PDB files** (0 or 1) 🡪 0 = no

Line 4 Token 2 specifies if the PDB files are **Multi** Chain (0 or 1) 🡪 0 = no

Line 5 Token 1 specifies the **number of Cartesian** (integer) modes to process 🡪 0 = none

Line 5 Token 2 specifies the **number of Distance** (integer) modes to process 🡪 3 = top three modes

Line 5 Token 3 specifies the **number of Cartesian modes to Visualize**(integer) 🡪 0 = none

Line 6 is a separator line between the batch parameters and the job parameters

Line 7 specifies the **working directory** (String) for job one

Line 8 Token 1 specifies the job **description** (String) for job one

Line 8 Token 2 specifies the **reference PDB** (String) for job one

Line 9 specifies the **residue list** (String) for job one

Line 10 Token 1 specifies the **coordinate matrix** (String) for job one

Line 10 Token 2 specifies the **reference column** in the coordinate matrix (String) for job one

Line 11 is a separator line between sets of job parameters

Line 12 specifies the **working directory** (String) for job two

Line 13 Token 1 specifies the job **description** (String) for job two

Line 13 Token 2 specifies the **reference PDB** (String) for job two

Line 14 specifies the **residue list** (String) for job two

Line 15 Token 1 specifies the **coordinate matrix** (String) for job two

Line 15 Token 2 specifies the **reference column** in the coordinate matrix (String) for job two

------------------------------------------------------------------------------------------------------------------------------------------------------

**Warning**:

Both the read PDB file flag and the Multi flag MUST be set to ZERO.

This tells JED to perform the analysis on Single Chain PDBs.

The number of cPCA modes MUST be zero.

There must be no Cartesian residue list specified.

The number of dPCA modes MUST NOT be zero.

There must be a Distance residue list specified.

If the subset you have chosen contains ***N*** residues, then you must NOT request more than ***N(N-1)/2*** modes.

**vi. Below is a sample input file for Multi-Chain PDB files:**

**Differences from Single-Chain analysis shown highlighted in amber**

------------------------------------------------------------------------------------------

$num\_of\_jobs

0.00

3.00

0 1 **2 A B 795 151 0 0**

0 3 0

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/working/directory1/

Description1 reference\_PDB\_file1.pdb

residues1.txt

original\_PDB\_Coordinates.txt 0

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/working/directory2/

Description2 reference\_PDB\_file2.pdb

residues2.txt

original\_PDB\_Coordinates.txt 0

------------------------------------------------------------------------------------------

Notes:

Line 1 specifies the **$num\_of\_jobs** (integer) for the batch

Line 2 specifies the **percent** (double) of frames to remove from the data: 0 🡪 none

Line 3 specifies the **z-cutoff** (double) for adjusting outliers: 0 🡪 values beyond |3.0|

Line 4 Token 1 specifies whether to **read PDB files** (0 or 1) 🡪 0 = no

Line 4 Token 2 specifies if the PDB files are **Multi** Chain (0 or 1) 🡪 1 = yes

Line 4 Token 3 specifies the number of chains 🡪 (2);

Line 4 Token 4 specifies the first chain ID 🡪 (A);

Line 4 Token 5 specifies the second chain ID 🡪 (B);

Line 4 Token 6 specifies the number of residues in chain A🡪 (795);

Line 4 Token 7 specifies the number of residues in chain B🡪 (151);

Line 4 Token 8 specifies the offset of Chain A 🡪 (0);

Line 4 Token 9 specifies the offset of Chain B 🡪 (0);

Line 5 Token 1 specifies the **number of Cartesian** (integer) modes to process 🡪 0

Line 5 Token 2 specifies the **number of Distance** (integer) modes to process 🡪 3 = Top 3 modes

Line 5 Token 3 specifies the **number of Cartesian modes to Visualize**(integer) 🡪 0

Line 6 is a separator line between the batch parameters and the job parameters

Line 7 specifies the **working directory** (String) for job one

Line 8 Token 1 specifies the job **description** (String) for job one

Line 8 Token 2 specifies the **reference PDB** (String) for job one

Line 9 specifies the **residue list** (String) for job one

Line 10 Token 1 specifies the **coordinate matrix** (String) for job one

Line 10 Token 2 specifies the **reference column** in the coordinate matrix (String) for job one

Line 11 is a separator line between sets of job parameters

Line 12 specifies the **working directory** (String) for job two

Line 13 Token 1 specifies the job **description** (String) for job two

Line 13 Token 2 specifies the **reference PDB** (String) for job two

Line 14 specifies the **residue list** (String) for job two

Line 15 Token 1 specifies the **coordinate matrix** (String) for job two

Line 15 Token 2 specifies the **reference column** in the coordinate matrix (String) for job two

------------------------------------------------------------------------------------------------------------------------------------------------------

**Warning**:

The read PDB file flag MUST be set to ZERO.

The multi flag must be set to ONE;

This tells JED to perform the analysis on MULTI Chain PDBs.

The number of cPCA modes MUST be zero.

There must be no Cartesian residue list specified.

The number of dPCA modes MUST NOT be zero.

There must be a Distance residue list specified.

If the subset you have chosen contains ***N*** residues, then you must NOT request more than ***N(N-1)/2*** modes.

Remember that the format of the residue list file is TWO columns for Multi Chain Analysis: **Chain ID, residue** **number** while for Single Chain Analysis, the file has ONE column: **residue number**.

**If number of dPCA modes > 0, then there must be a Distance residue list file!**

**E. Debugging Crashes Part II:**

Things that will generally make your life miserable…

**i. Dumb mistakes:**

Did you set the Read and Multi flags correctly? If not, CRASH!!!

Did you request cPCA but not specify a Cartesian residue list? If so, CRASH!!!

Did you set the number of modes correctly? If not, CRASH!!!

Are the 0 and 1 flags set correctly? If not CRASH!!!

Are you requesting to read PDBs when you should be specifying a coordinate matrix? If so CRASH!!!

**ii. More subtle mistakes:**

If the PDB file residue numbering starts at 5, then the chain offset is not 0, it is 4 Garbage out!

Did you request more modes than actually exist? For example, if your Cartesian subset contains 12 residues, but you asked for 50 modes, then you are going to crash JED because there are only 36 Cartesian modes in total.

Also, if your Distance subset contains 3 residues and you request 5 modes, then you are going to crash JED because there are only 3 distance modes in total.

If your trajectory has not equilibrated, then you must address the problem or outliers. If you do not, then the Q and R matrices will be highly ill-conditioned and may cause the eigenvalue decomposition to fail. You can check the variables in statistics packages that compute the Kaiser-Myer-Olkin (KMO) statistic as well as the Measure of Sampling Adequacy (MSA) for each coordinate variable to critically assess your data. If it is not well suited for PCA, you can condition the variables by setting the z-cutoff in JED between 2.0 and 3.0 when running your jobs. This type of conditioning is by far not very sophisticated, but it has the effect of lowering the condition numbers of Q and R as well as un-dilating the high and low regions of the eigenspectrum. In particular, it does not alter the ordinality of the eigenspectrum but does correct the distortion that arises from under sampling when trying to estimate the population covariance matrix from the sample covariance matrix.

**F. Performing Cartesian and Distance PCA with Mode Visualization**

***The working directory must contain: The coordinates matrix, the PDB reference file, and both residue lists.***

***It may also contain JED\_Batch\_Driver.txt***

JED is capable of doing both cPCA and dPCA, using both Q and R, and generating cPCA mode visualizations simultaneously. All outputs are delivered as discussed for the individual components.

The outputs to this type of job include the outputs for both the cPCA and dPCA analyses, as well as all the structures for the top modes chosen for visualization. JED will permute the reference structure for a given subset along the top eigenvectors selected for visualization and output structures (PDBs) that capture one cycle of this motion. The amplitude of the motion is determined by the value of the **$scale\_factor**, whose default value is 1.0, and can be adjusted as necessary. Setting the value too high will cause Visualization software like Pymol to break the ribbon diagrams of the structures. Ultimately, this is controlled by the magnitude of the eigenvector components for any given residue. Setting the scale factor between 1 and 3 is usually safe, but for proteins with highly mobile regions like loops, you may need to choose a smaller scale factor. This is done for both the top Q and R modes. Additionally, Pymol scripts are generated to animate those structures into a movie for better analysis of the physical meanings of the top modes.

These files will be located in the /VIZ subdirectory of the root of the JED results tree:

/working/directory/JED\_Results\_Description/VIZ/

The Q results will be in the subdirectory /COV and the R results will be in the subdirectory /CORR.

One huge advantage of JED is that it is highly configurable and can perform many types of Essential Dynamics analysis concurrently. Combined with the cluster resources or just using the batch feature allows a user to process a great deal of data efficiently.

**i. Below is a sample input file for Single-Chain PDB files:**

------------------------------------------------------------------------------------------

$num\_of\_jobs

0.00

3.00

0 0

20 3 2 1.0

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/working/directory1/

Description1 reference\_PDB\_file1.pdb

residues\_cartesian1.txt

residues\_dist1.txt

original\_PDB\_Coordinates.txt 0

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/working/directory2/

Description2 reference\_PDB\_file2.pdb

residues\_cartesian2.txt

residues\_dist2.txt

original\_PDB\_Coordinates.txt 0

------------------------------------------------------------------------------------------

Notes:

Line 1 specifies the **$num\_of\_jobs** (integer) for the batch

Line 2 specifies the **percent** (double) of frames to remove from the data: 0 = none

Line 3 specifies the **z-cutoff** (double) for adjusting outliers: 0 🡪 values beyond |3.0|

Line 4 Token 1 specifies whether to **read PDB files** (0 or 1) 🡪 0 = no

Line 4 Token 2 specifies if the PDB files are **Multi** Chain (0 or 1) 🡪 0 = no

Line 5 Token 1 specifies the **number of Cartesian** (integer) modes to process 🡪 0 = none

Line 5 Token 2 specifies the **number of Distance** (integer) modes to process 🡪 3 = top three modes

Line 5 Token 3 specifies the **number of Cartesian modes to Visualize**(integer) 🡪 0 = none

Line 6 is a separator line between the batch parameters and the job parameters

Line 7 specifies the **working directory** (String) for job one

Line 8 Token 1 specifies the job **description** (String) for job one

Line 8 Token 2 specifies the **reference PDB** (String) for job one

Line 9 specifies the **Cartesian** **residue list** (String) for job one

Line 10 specifies the **Distance** **residue list** (String) for job one

Line 11 Token 1 specifies the **coordinate matrix** (String) for job one

Line 11 Token 2 specifies the **reference column** in the coordinates matrix (integer) for job one

Line 12 is a separator line between sets of job parameters

Line 13 specifies the **working directory** (String) for job two

Line 14 Token 1 specifies the job **description** (String) for job two

Line 14 Token 2 specifies the **reference PDB** (String) for job two

Line 15 specifies the **Cartesian** **residue list** (String) for job two

Line 16 specifies the **Distance** **residue list** (String) for job two

Line 17 Token 1 specifies the **coordinate matrix** (String) for job two

Line 17 Token 2 specifies the **reference column** in the coordinates matrix (integer) for job two

---------------------------------------------------------------------------------------------------------------------------------------------------------

**Warning**:

Both the read PDB file flag and the Multi flag MUST be set to ZERO.

The number of cPCA modes MUST be > zero.

The number of dPCA modes MUST be > zero.

There must be a Cartesian residue list specified.

There must be a Distance residue list specified.

For Cartesian subsets containing ***N*** residues, you must NOT request more than **3*N*** modes.

For Distance subsets containing ***N*** residues, you must NOT request more than ***N(N-1)/2*** modes.

**ii. Below is a sample input file for Multi-Chain PDB files:**

**Differences from Single-Chain analysis shown highlighted in amber**

------------------------------------------------------------------------------------------

$num\_of\_jobs

0.00

3.00

0 1 **2 A B 795 151 0 0**

20 3 2 1.0

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/working/directory1/

Description1 reference\_PDB\_file1.pdb

residues\_cartesian1.txt

residues\_dist1.txt

original\_PDB\_Coordinates.txt 0

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

/working/directory2/

Description2 reference\_PDB\_file2.pdb

residues\_cartesian2.txt

residues\_dist2.txt

original\_PDB\_Coordinates.txt 0

------------------------------------------------------------------------------------------

Notes:

Line 1 specifies the **$num\_of\_jobs** (integer) for the batch

Line 2 specifies the **percent** (double) of frames to remove from the data: 0 🡪 None

Line 3 specifies the **z-cutoff** (double) for adjusting outliers: 0 🡪 values beyond |3.0|

Line 4 Token 1 specifies whether to **read PDB files** (0 or 1) 🡪 0 = no

Line 4 Token 2 specifies if the PDB files are **Multi** Chain (0 or 1) 🡪 1 = yes

Line 4 Token 3 specifies the number of chains 🡪 (2);

Line 4 Token 4 specifies the first chain ID 🡪 (A);

Line 4 Token 5 specifies the second chain ID 🡪 (B);

Line 4 Token 6 specifies the number of residues in chain A🡪 (795);

Line 4 Token 7 specifies the number of residues in chain B🡪 (151);

Line 4 Token 8 specifies the offset of Chain A 🡪 (0);

Line 4 Token 9 specifies the offset of Chain B 🡪 (0);

Line 5 Token 1 specifies the **number of Cartesian** (integer) modes to process 🡪 20

Line 5 Token 2 specifies the **number of Distance** (integer) modes to process 🡪 0 = none

Line 5 Token 3 specifies the **number of Cartesian modes to Visualize**(integer) 🡪 2 = top two modes

Line 5 Token 4 specifies the **scale factor** for the visualizations (double) 🡪 1.0 = default value

Line 6 is a separator line between the batch parameters and the job parameters

Line 7 specifies the **working directory** (String) for job one

Line 8 Token 1 specifies the job **description** (String) for job one

Line 8 Token 2 specifies the **reference PDB** (String) for job one

Line 9 specifies the **Cartesian** **residue list** (String) for job one

Line 10 specifies the **Distance** **residue list** (String) for job one

Line 11 Token 1 specifies the **coordinate matrix** (String) for job one

Line 11 Token 2 specifies the **reference column** in the coordinates matrix (integer) for job one

Line 12 is a separator line between sets of job parameters

Line 13 specifies the **working directory** (String) for job two

Line 14 Token 1 specifies the job **description** (String) for job two

Line 14 Token 2 specifies the **reference PDB** (String) for job two

Line 15 specifies the **Cartesian** **residue list** (String) for job two

Line 16 specifies the **Distance** **residue list** (String) for job two

Line 17 Token 1 specifies the **coordinate matrix** (String) for job two

Line 17 Token 2 specifies the **reference column** in the coordinates matrix (integer) for job two

-----------------------------------------------------------------------------------------------------------------------------------------------------

**Warning**:

The read PDB file flag MUST be set to ZERO.

The multi flag must be set to ONE; This tells JED to perform the analysis on MULTI Chain PDBs.

The number of cPCA modes MUST be > zero.

The number of dPCA modes MUST be > zero.

There must be a Cartesian residue list specified.

There must be a Distance residue list specified.

For Cartesian subsets containing ***N*** residues, you must NOT request more than **3*N*** modes.

For Distance subsets containing ***N*** residues, you must NOT request more than ***N(N-1)/2*** modes.

Remember that the format of the residue list file is TWO columns for Multi Chain Analysis: **Chain ID, residue** **number** while for Single Chain Analysis, the file has ONE column: **residue number**.

**If number of cPCA modes > 0, then there must be a Cartesian residue list file!**

**If number of dPCA modes > 0, then there must be a Distance residue list file!**

**VI. Additional Types of Analysis**

**A. Pooling Data:**

It is often useful to pool trajectory statistics. This can be done in JED by combining coordinate files and then performing the usual analysis. To combine the coordinate files, there is a utility program called **JED\_Pool\_Data.java** that will combine multiple matrices into one. Of courses, the number of rows in the coordinate files must match. The matrices to combine are specified by an input file called **pool.txt** that the user must construct correctly.

**Below is a sample pool.txt file:**

------------------------------------------------------------------------------------------

1

3

/output/directory/

/path/to/first/matrix/matrix\_1.txt

/path/to/first/matrix/matrix\_2.txt

/path/to/first/matrix/matrix\_3.txt

------------------------------------------------------------------------------------------

Notes:

Line #1 specifies the number of jobs (integer) 🡪1

**Then for each job you must specify the following:**

Line #2 specifies the number of matrices to combine (integer) 🡪3

Line #3 specifies the output directory (string)

Line #4 specifies the path to the first matrix (String)

Line #5 specifies the path to the second matrix (String)

Line #6 specifies the path to the third matrix (String)

------------------------------------------------------------------------------------------

**B. Subspace Analysis:**

Once JED Driver has been run on multiple trajectories as well as pooled trajectories, an analysis can be done to compare how similar the essential subspaces derived from those trajectories are to each other. JED contains a program called **Subspace\_Analysis.java** along with 3 driver programs that perform those functions. The core program takes as input two matrices of eigenvectors derived from PCA (or NMA, ANM, etc.). The matrices must have the same number of rows and columns, meaning the vectors being compared come from the same vector space and that the subspaces have the same dimensions. For example, in an analysis of lysozyme you might choose to process 20 cPCA modes while examining 10 different experimental conditions plus pooled data. As long as all the subsets in the analysis are the same then all the 20 dimensional subspaces can be compared.

Like most of the JED programs, the subspace analysis program driver reads an input file called **SSA.txt** to obtain runtime information. This file must be constructed properly by the user to perform the analysis. The three driver programs are **SSA\_Driver.java**, **FSSA\_Driver.java**, and **FSSA\_Iterated\_Driver.java** and are different in how much analysis is requested. The SSA\_Driver gives full outputs for non-iterated subspace comparison including both log files and individual flat files. The FSSA\_Driver is a light-weight version with only RMSIP and PA output in the log files. The Iterated version performs a recursive variation of the above where all equidimensional subspaces are compared up to the size that was provided, for example, from 1 to 20 by step-size 1 for a 20 column input file.

ALL three drivers use the same input file, only the outputs are different.

**Below is a sample subspace\_analysis.txt file:**

The format for the file shown below is:

LINE 1: Number\_of\_Jobs (integer)

LINE 2: Output\_Directory (string ending in "/" or "\\")

LINE 3: Batch\_Decription (string)

FOR EACH JOB:

Description (string)

Directory1 (string ending in "/" or "\\") Name1 (string)

Directory2 (string ending in "/" or "\\") Name2 (string)

------------------------------------------------------------------------------------------

4

/output/directory/

Single\_Combo\_SSA

All-vs-A

/Users/physicslabs/ all\_combo\_SS\_75\_top\_20\_eigenvectors.txt

/Users/physicslabs/ 1a6n\_combo\_SS\_75\_top\_20\_eigenvectors.txt

All-vs-B

/Users/physicslabs/ all\_combo\_SS\_75\_top\_20\_eigenvectors.txt

/Users/physicslabs/ 1wit\_combo\_SS\_75\_top\_20\_eigenvectors.txt

All-vs-A+B

/Users/physicslabs/ all\_combo\_SS\_75\_top\_20\_eigenvectors.txt

/Users/physicslabs/ 1ubq\_combo\_SS\_75\_top\_20\_eigenvectors.txt

All-vs-A\_B

/Users/physicslabs/ all\_combo\_SS\_75\_top\_20\_eigenvectors.txt

/Users/physicslabs/ 1ypi\_combo\_SS\_75\_top\_20\_eigenvectors.txt

------------------------------------------------------------------------------------------

It is possible to specify hundreds of comparisons when using the driver input file, but please construct the input file carefully... Just one typo in a list of one thousand comparisons 🡪 CRASH.

**APPENDIX**

**A. Sample JED Log file:**

Java Essential Dynamics

Job Description: Single\_cPCA\_dPCA\_Viz

Working directory: C:\\Users\\Charles\\workspace\\JED\_1.0\\JED\_Test\\Single\\

Output directory: C:\\Users\\Charles\\workspace\\JED\_1.0\\JED\_Test\\Single\\JED\_Results\_Single\_cPCA\_dPCA\_Viz/

Performing cPCA: 20 modes.

Performing dPCA: 3 modes.

Performing mode visualization on top 2 cPCA modes

Alpha carbon coordinates were obtained from file: C:\\Users\\Charles\\workspace\\JED\_1.0\\JED\_Test\\Single\\original\_PDB\_coordinates.txt

The number of residues read: 151

The number of conformations read: 2001

The dimension of the coordinates matrix is: 453 by 2001

The transformed PDB coordinates were obtained by quaternion least-squares alignment.

PDB reference file is 1A6N.pdb

Reference conformation (column) in the coordinate matrix: 0

Residue list for Cartesian subset: residues.txt

Number of residues in Cartesian subset: 81

The transformed data was trimmed by removing 10.0 percent of the samples based on conformation RMSD.

The coordinates were 'conditioned' by adjusting outliers with Z-scores beyond 3.0 to their mean value.

Condition number of the covariance matrix (Q): 52,101,318,291,300,056

LOG of the Q Condition Number: 17

Trace of Q: 69

Condition number of the correlation matrix (R): 35,505,888,101,013,780

LOG of the R Condition Number: 17

Trace of R: 243

PDB file with B-factors replaced by residue RMSDs: Single\_cPCA\_dPCA\_Viz\_SS\_81.pdb

Number of residues in distance subset: 12

The coordinates were 'conditioned' by adjusting outliers with Z-scores beyond |3.0| to their mean value.

Condition number of the distance Covariance matrix (Q\_dist): 80,427,675

LOG of Q\_dist: 8

Trace of Q\_dist: 2

Condition number of the distance Correlation matrix (R\_dist): 183,847

LOG of R\_dist: 5

Trace of R\_dist: 66

MEANs and STD\_DEVs for the residue distances:

Res1 Res2 Mean Std\_Dev

1 2 3.780 0.004

1 3 6.388 0.245

1 4 9.460 0.425

1 5 11.239 0.242

1 6 9.326 0.284

1 7 9.790 0.511

1 8 13.292 0.427

1 9 13.982 0.305

1 10 13.117 0.431

1 11 15.209 0.647

1 12 17.872 0.479

2 3 3.808 0.005

2 4 6.632 0.067

2 5 7.994 0.080

2 6 5.818 0.115

2 7 6.416 0.145

2 8 9.854 0.134

2 9 10.503 0.138

2 10 9.922 0.150

2 11 12.086 0.152

2 12 14.519 0.149

3 4 3.783 0.009

3 5 5.308 0.010

3 6 5.113 0.011

3 7 6.309 0.011

3 8 8.648 0.012

3 9 10.012 0.012

3 10 10.692 0.012

3 11 12.367 0.012

3 12 14.171 0.014

4 5 3.804 0.004

4 6 5.409 0.005

4 7 5.070 0.003

4 8 6.172 0.005

4 9 8.766 0.006

4 10 9.846 0.006

4 11 10.448 0.006

4 12 12.007 0.008

5 6 3.784 0.003

5 7 5.458 0.004

5 8 5.064 0.003

5 9 6.444 0.004

5 10 8.797 0.005

5 11 10.033 0.006

5 12 10.526 0.006

6 7 3.809 0.006

6 8 5.357 0.006

6 9 5.167 0.004

6 10 6.341 0.006

6 11 8.741 0.007

6 12 9.919 0.007

7 8 3.823 0.005

7 9 5.407 0.006

7 10 4.992 0.003

7 11 6.075 0.004

7 12 8.364 0.007

8 9 3.815 0.004

8 10 5.431 0.004

8 11 5.248 0.004

8 12 5.868 0.005

9 10 3.806 0.005

9 11 5.620 0.005

9 12 5.210 0.004

10 11 3.777 0.004

10 12 5.545 0.005

11 12 3.787 0.006

Sets of structures were generated to animate each of the top 2 cPCA modes using both Q and R PCA methods.

MODE AMPLITUDE: 1.000

Analysis completed: 2013-12-16 02:25:55

**B. Sample PDB READ Log file:**

1A6N.pdb

1A6N\_froda\_00000001.pdb

1A6N\_froda\_00000002.pdb

1A6N\_froda\_00000003.pdb

1A6N\_froda\_00000004.pdb

1A6N\_froda\_00000005.pdb

1A6N\_froda\_00000006.pdb

1A6N\_froda\_00000007.pdb

1A6N\_froda\_00000008.pdb

1A6N\_froda\_00000009.pdb

1A6N\_froda\_00000010.pdb

1A6N\_froda\_00000011.pdb

1A6N\_froda\_00000012.pdb

1A6N\_froda\_00000013.pdb

1A6N\_froda\_00000014.pdb

1A6N\_froda\_00000015.pdb

1A6N\_froda\_00000016.pdb

1A6N\_froda\_00000017.pdb

1A6N\_froda\_00000018.pdb

1A6N\_froda\_00000019.pdb

1A6N\_froda\_00000020.pdb

1A6N\_froda\_00000021.pdb

1A6N\_froda\_00000022.pdb

1A6N\_froda\_00000023.pdb

1A6N\_froda\_00000024.pdb

1A6N\_froda\_00000025.pdb

**C. Sample Single Chain PDB Residue List file:**

1

2

3

4

5

6

7

8

9

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

**D. Sample Multi Chain PDB Residue List file:**

A 1

A 2

A 3

A 4

A 5

A 6

A 7

A 8

A 9

A 10

A 11

A 12

A 13

A 14

A 15

A 16

A 17

A 18

A 19

A 20

A 21

A 22

A 23

A 24

A 25

B 1

B 2

B 3

B 4

B 5

B 6

B 7

B 8

B 9

B 10

B 11

B 12

B 13

B 14

B 15

B 16

B 17

B 18

B 19

B 20

**E. Sample SSA Log File:**

Top\_COV\_Eigenvectors

Rows: 45

Cols: 3

Top\_CORR\_Eigenvectors

Rows: 45

Cols: 3

Projections file written to:

C:\\Users\\Charles\\workspace\\JED\_1.0\\JED\_Test\\Multi\\JED\_results\_Multi\_cPCA\_dPCA\_Viz/dPCA/SSA/Multi\_cPCA\_dPCA\_Viz\_projections\_dim\_3.txt

Cumulative overlaps 1 --> 2 file written to:

C:\\Users\\Charles\\workspace\\JED\_1.0\\JED\_Test\\Multi\\JED\_results\_Multi\_cPCA\_dPCA\_Viz/dPCA/SSA/Multi\_cPCA\_dPCA\_Viz\_CO\_1\_2\_dim\_3.txt

Cumulative overlaps 2 --> 1 file written to:

C:\\Users\\Charles\\workspace\\JED\_1.0\\JED\_Test\\Multi\\JED\_results\_Multi\_cPCA\_dPCA\_Viz/dPCA/SSA/Multi\_cPCA\_dPCA\_Viz\_CO\_2\_1\_dim\_3.txt

Principle Angles file written to:

C:\\Users\\Charles\\workspace\\JED\_1.0\\JED\_Test\\Multi\\JED\_results\_Multi\_cPCA\_dPCA\_Viz/dPCA/SSA/Multi\_cPCA\_dPCA\_Viz\_PA\_dim\_3.txt

Cosine Products file written to:

C:\\Users\\Charles\\workspace\\JED\_1.0\\JED\_Test\\Multi\\JED\_results\_Multi\_cPCA\_dPCA\_Viz/dPCA/SSA/Multi\_cPCA\_dPCA\_Viz\_Cosine\_Products\_dim\_3.txt

Vectorial sums of angles file written to:

C:\\Users\\Charles\\workspace\\JED\_1.0\\JED\_Test\\Multi\\JED\_results\_Multi\_cPCA\_dPCA\_Viz/dPCA/SSA/Multi\_cPCA\_dPCA\_Viz\_Vector\_sums\_of\_Angles\_dim\_3.txt

The absolute projections of each vector in subspace 1 with each vector in subspace 2 are:

0.76 0.52 0.07

0.31 0.69 0.47

0.06 0.37 0.70

The cumulative overlaps CO\_3 for each vector in subspace 1 with all the vectors in subspace 2 are:

Vector 1 0.919

Vector 2 0.890

Vector 3 0.797

The cumulative overlaps CO\_3 for each vector in subspace 2 with all the vectors in subspace 1 are:

Vector 1 0.820

Vector 2 0.936

Vector 3 0.850

The RMSIP score is 0.870

The principle angles (in degrees) are:

PA1 6

PA2 21

PA3 31

The cosine products (in degrees) are:

CP1 6

CP2 30

CP3 51

The vectorial sums of angles (in degrees) are:

VS1 6

VS2 22

VS3 38

Maximum possible angle between two subspaces of this dimension is 90 degrees

Analysis completed: 2013-12-16 02:15:05

**E. Sample FSSA Iterated Log File:**

Principle Angle Spectra file written to:

C:\\Users\\Charles\\workspace\\JED\_1.0\\JED\_Test\\Multi\\JED\_results\_Multi\_cPCA\_dPCA\_Viz/dPCA/SSA/Multi\_cPCA\_dPCA\_Viz\_PA\_Spectra.txt

RMSIPs file written to:

C:\\Users\\Charles\\workspace\\JED\_1.0\\JED\_Test\\Multi\\JED\_results\_Multi\_cPCA\_dPCA\_Viz/dPCA/SSA/Multi\_cPCA\_dPCA\_Viz\_RMSIPs.txt

RMSIPs:

Dim 1 0.758

Dim 2 0.839

Dim 3 0.870

The PA spectra for the range of subspaces are:

29 0 0

14 32 0

6 21 31

Analysis completed: 2013-12-16 02:15:05

**E. Sample Random SSA Log File:**

Average RMSIPs file written to:

C:\\Users\\Charles\\workspace\\JED\_1.0\\JED\_Test\\Multi\\JED\_results\_Multi\_cPCA\_dPCA\_Viz/dPCA/SSA/Multi\_cPCA\_dPCA\_Viz\_avg\_random\_RMSIPs.txt

Average PAs file written to:

C:\\Users\\Charles\\workspace\\JED\_1.0\\JED\_Test\\Multi\\JED\_results\_Multi\_cPCA\_dPCA\_Viz/dPCA/SSA/Multi\_cPCA\_dPCA\_Viz\_avg\_random\_PAs.txt

Average COs file written to:

C:\\Users\\Charles\\workspace\\JED\_1.0\\JED\_Test\\Multi\\JED\_results\_Multi\_cPCA\_dPCA\_Viz/dPCA/SSA/Multi\_cPCA\_dPCA\_Viz\_avg\_random\_COs.txt

RMSIP Std Devs file written to:

C:\\Users\\Charles\\workspace\\JED\_1.0\\JED\_Test\\Multi\\JED\_results\_Multi\_cPCA\_dPCA\_Viz/dPCA/SSA/Multi\_cPCA\_dPCA\_Viz\_random\_RMSIP\_std\_devs.txt

The dimension of the vector space is 45

SS\_DIM avg\_RMSIP std\_dev

1 0.883 0.094

2 0.647 0.069

3 0.557 0.064

The avg random PA spectra for the range of subspaces are:

27 0 0

26 71 0

25 61 75

The avg random CO scores for the range of subspaces are:

0.794 0.000 0.000

0.803 0.176 0.000

0.810 0.242 0.156

Analysis completed: 2013-12-16 02:15:05