Bio3D-web Summary Report

Lars Skjaerven, Shashank Jariwala & Barry J. Grant

September 06, 2016 (Report template updated: April 12, 2016)

Contents

Overview **1**

1 SEARCH: Structure Search Summary **2**

2 ALIGN: Multiple Sequence Alignment **8**

3 FIT: Structure Superposition **10**

4 PCA: Principal Component Analysis **10**

5 NMA: Normal Mode Analysis **17**

Conventional Usage Example **22**

Citation information **23**

Session and Software Version Information **23**

References **24**

Overview

This report was generated on Tuesday, September 06 2016 at 23:30 (EST) by the *Bio3D principal component analysis and ensemble normal mode analysis web application* (Bio3D-web) version **0.1**. For complete version information of all dependencies please see the session and software version information section below.

All included figures and report values in this document reflect those displayed in the online app with user supplied options (including actual input, graph type, clustering and similarity thresholds

etc.). Further customization of analysis protocols and all resulting figures is possible with Bio3D (Skjaerven et al. 2015) itself. Please see the conventional usage example section together with our collection of tutorials for further details about using Bio3D directly on your own computers.

We also automatically save your data as you proceed through the analysis. To revisit your session, please click the following link:
https://dcmb-grant-shiny.umms.med.umich.edu/pca-app/?SSUID=2016-09-06_46e9498ff410.

1 SEARCH: Structure Search Summary

User input consisted of a single PDB structure code: **1BJU**.

This structure is annotated as: **BETA-TRYPSIN (Bos taurus)** in the RCSB PDB database (Berman et al. 2000). The user selected chain for further analysis was chain id: **A**. Pfam database annotation of this chain can be found in **Table 1** (Finn et al. 2014) along with a simplified structure visualization in **Figure 1** and structural composition log below.

Table 1: Pfam database annotation.

ID	PFAM	Annotation	eValue
1BJU_A	Trypsin (PF00089.22)	Trypsin	0.0

Input PDB composition log A sequence based HMMER (3.1b2 (February 2015); http://hmmer.org/) (S. R. Eddy 2011) search identified **2301** sequences similar hits in the RCSB PDB database. The distribution of alignment bitscores to the input sequence is shown in **Figure 2**. From these hits **30** were selected for further analysis based on bitscore cutoff of **277** (encompassing **608** structures above this cutoff) and inclusion limit of **30** structures. See **Table 2** for selected structures for further analysis.

Figure 1: PDB overview.

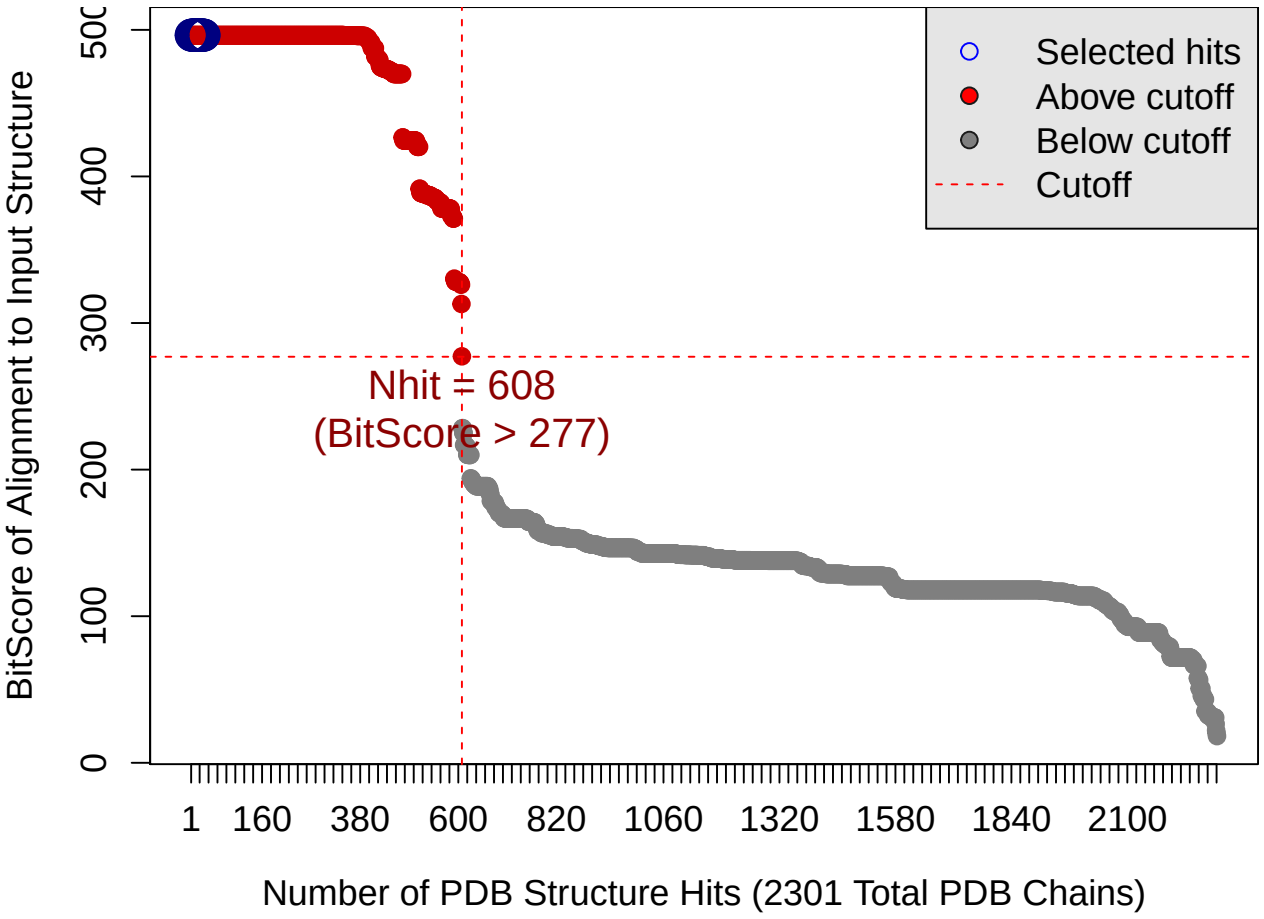

Figure 2: Summary of search results for user input query structure/sequence against the RCSB PDB chain database.

Table 2: Selected sequence similar structures for further analysis (continued below)

ID	BitScore	eValue	Name (PDB Title)	Species	Ligands
1AQ7_A	496.4	2e-148	TRYPSIN	Bos taurus	
1AUJ_A	496.4	2e-148	TRYPSIN	Bos taurus	CA,PPB
1AZ8_A	496.4	2e-148	TRYPSIN	Bos taurus	IN4
1BJU_A	496.4	2e-148	BETA-TRYPSIN	Bos taurus	CA,DMS,GP6,SO4
1BJV_A	496.4	2e-148	BETA-TRYPSIN	Bos taurus	CA,DMS,GP8,SO4
1C1N_A	496.4	2e-148	TRYPSIN	Bos taurus	BAM,CA,SO4,ZN
1C1O_A	496.4	2e-148	TRYPSIN	Bos taurus	CA,MG,SO4
1C1P_A	496.4	2e-148	TRYPSIN	Bos taurus	BAI,CA,DMS,MG,SO4
1C1Q_A	496.4	2e-148	TRYPSIN	Bos taurus	BAI,CA,MG,SO4
1C1R_A	496.4	2e-148	TRYPSIN	Bos taurus	BAI,CA,DMS,MG,SO4,ZN
1C1S_A	496.4	2e-148	TRYPSIN	Bos taurus	BAB,CA,NA,PO4
1C1T_A	496.4	2e-148	TRYPSIN	Bos taurus	BAB,CA,MG,SO4
1C2D_A	496.4	2e-148	TRYPSIN	Bos taurus	BAK,CA,MG,ZN
1C2E_A	496.4	2e-148	TRYPSIN	Bos taurus	BAK,CA,MG,ZN
1C2F_A	496.4	2e-148	TRYPSIN	Bos taurus	BAH,CA,DMS,MG,ZN
1C2G_A	496.4	2e-148	TRYPSIN	Bos taurus	BAH,CA,DMS,MG,ZN
1C2H_A	496.4	2e-148	TRYPSIN	Bos taurus	BAK,CA,MG,ZN
1C2I_A	496.4	2e-148	TRYPSIN	Bos taurus	BAK,CA,CL,DMS,ZN
1C2J_A	496.4	2e-148	TRYPSIN	Bos taurus	BAK,CA,MG,SO4,ZN
1C2K_A	496.4	2e-148	TRYPSIN	Bos taurus	ABI,CA,CL,ZN
1C2L_A	496.4	2e-148	TRYPSIN	Bos taurus	CA,CL,ZN
1C2M_A	496.4	2e-148	TRYPSIN	Bos taurus	CA,MG,SO4,ZN

ID	BitScore	eValue	Name (PDB Title)	Species	Ligands
1C5P_A	496.4	2e-148	PROTEIN (TRYPSIN)	Bos taurus	BAM,CA,MG,SO4
1C5Q_A	496.4	2e-148	PROTEIN (TRYPSIN)	Bos taurus	CA,CL,ESI
1C5R_A	496.4	2e-148	PROTEIN (TRYPSIN)	Bos taurus	CA,ESI,FLC,SO4
1C5S_A	496.4	2e-148	PROTEIN (TRYPSIN)	Bos taurus	CA,ESX,SO4
1C5T_A	496.4	2e-148	PROTEIN (TRYPSIN)	Bos taurus	CA,ESP
1C5U_A	496.4	2e-148	PROTEIN (TRYPSIN)	Bos taurus	CA,ESP,MG,SO4
1C5V_A	496.4	2e-148	PROTEIN (TRYPSIN)	Bos taurus	CA,MG
1C9T_A	496.4	2e-148	TRYPSIN	Bos taurus	

Method	Resolution (A)	Space Group	Citation
X-RAY DIFFRACTION	2.2	P 61	Sandler et al. J.Am.Chem.Soc. (1998)
X-RAY DIFFRACTION	2.1	P 21 21 21	Lee et al. Biochemistry (1997)
X-RAY DIFFRACTION	1.8	P 21 21 21	Alexander et al. To be Published (NA)
X-RAY DIFFRACTION	1.8	P 21 21 21	Presnell et al. Biochemistry (1998)
X-RAY DIFFRACTION	1.8	P 21 21 21	Presnell et al. Biochemistry (1998)
X-RAY DIFFRACTION	1.4	P 31 2 1	Katz et al. Nature (1998)
X-RAY DIFFRACTION	1.4	P 31 2 1	Katz et al. Nature (1998)
X-RAY DIFFRACTION	1.37	P 31 2 1	Katz et al. Nature (1998)
X-RAY DIFFRACTION	1.37	P 31 2 1	Katz et al. Nature (1998)
X-RAY DIFFRACTION	1.37	P 31 2 1	Katz et al. Nature (1998)
X-RAY DIFFRACTION	1.63	P 31 2 1	Katz et al. J.Mol.Biol. (1999)
X-RAY DIFFRACTION	1.37	P 31 2 1	Katz et al. Nature (1998)
X-RAY DIFFRACTION	1.65	P 31 2 1	Katz et al. Nature (1998)
X-RAY DIFFRACTION	1.65	P 31 2 1	Katz et al. Nature (1998)
X-RAY DIFFRACTION	1.7	P 31 2 1	Katz et al. Nature (1998)

Method	Resolution (A)	Space Group	Citation
X-RAY DIFFRACTION	1.65	P 31 2 1	Katz et al. Nature (1998)
X-RAY DIFFRACTION	1.4	P 31 2 1	Katz et al. Nature (1998)
X-RAY DIFFRACTION	1.47	P 31 2 1	Katz et al. Nature (1998)
X-RAY DIFFRACTION	1.4	P 31 2 1	Katz et al. Nature (1998)
X-RAY DIFFRACTION	1.65	P 31 2 1	Katz et al. Nature (1998)
X-RAY DIFFRACTION	1.5	P 31 2 1	Katz et al. Nature (1998)
X-RAY DIFFRACTION	1.4	P 31 2 1	Katz et al. Nature (1998)
X-RAY DIFFRACTION	1.43	P 31 2 1	Katz et al. Chem.Biol. (2000)
X-RAY DIFFRACTION	1.43	P 21 21 21	Katz et al. Chem.Biol. (2000)
X-RAY DIFFRACTION	1.47	P 31 2 1	Katz et al. Chem.Biol. (2000)
X-RAY DIFFRACTION	1.36	P 31 2 1	Katz et al. Chem.Biol. (2000)
X-RAY DIFFRACTION	1.37	P 31 2 1	Katz et al. Chem.Biol. (2000)
X-RAY DIFFRACTION	1.37	P 31 2 1	Katz et al. Chem.Biol. (2000)
X-RAY DIFFRACTION	1.48	P 31 2 1	Katz et al. Chem.Biol. (2000)
X-RAY DIFFRACTION	3.3	P 1 21 1	Rester et al. J.Mol.Biol. (1999)

2 ALIGN: Multiple Sequence Alignment

Multiple sequence alignment of 30 selected hits results in an alignment with **30** rows (sequences) and **223** columns (alignment positions), **223** of which are non-gap positions. See **Figure 3**. for a schematic overview of this alignment where white segments between gray blocks represent gap positions. Note that the full rendered alignment is viewable and downloadable from the ALIGN tab of the web-app.

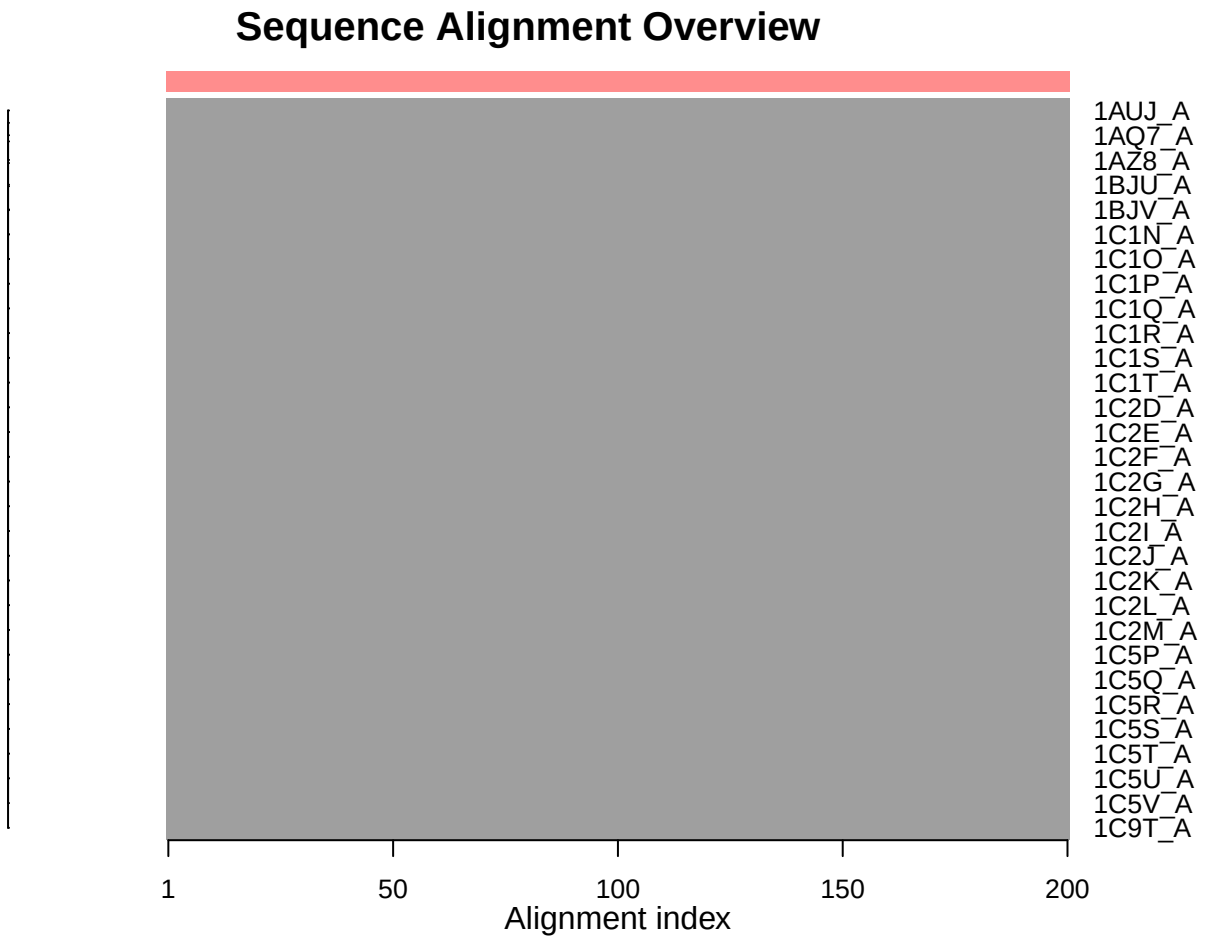

Figure 3: Sequence alignment overview with sequence groups (dendrogram) and gap positions (white segments).

Sequence identity clustering of the aligned sequences with the **ward.D2** method yields a clustering dendrogram that can be partitioned into **1** groups, see **Figure 4**.

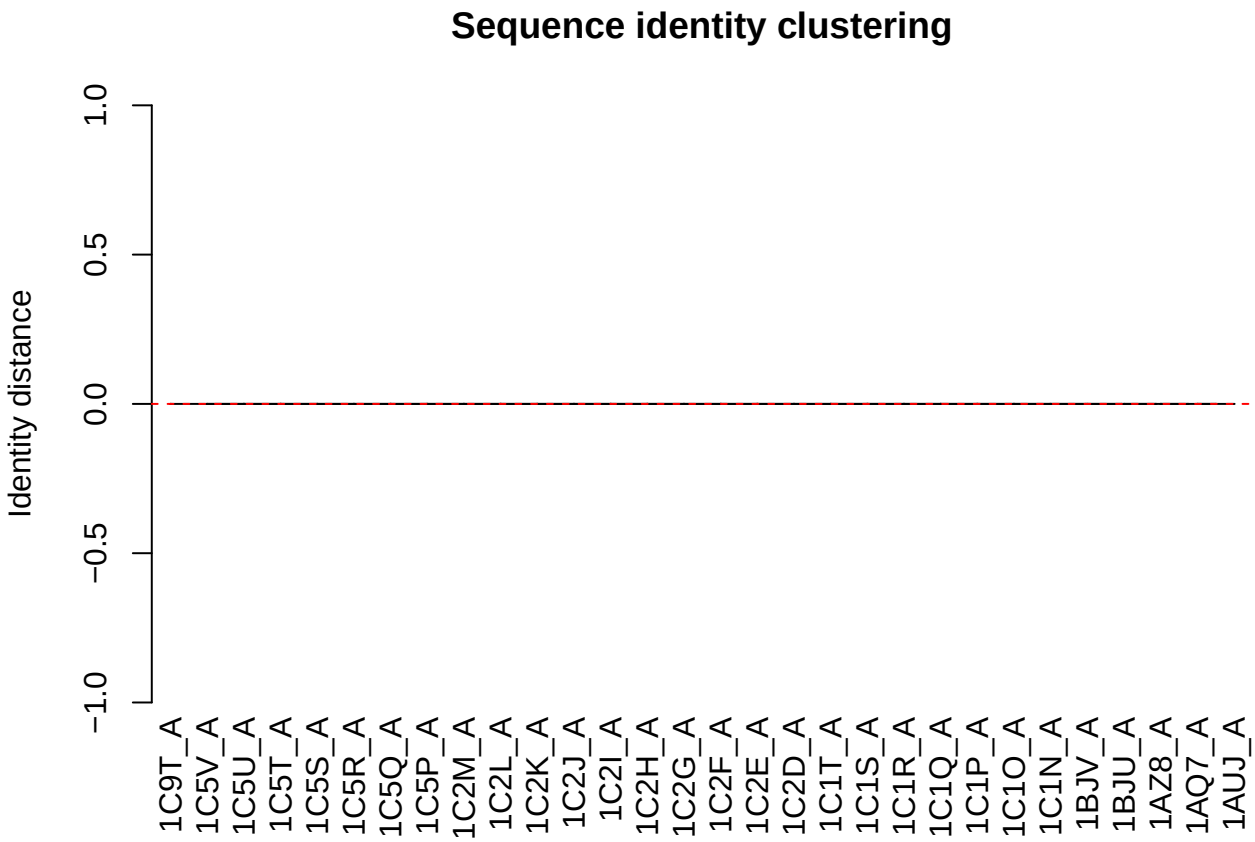

Figure 4: Sequence identity cluster dendrogram.

3 FIT: Structure Superposition

Structures were superposed on their **65** invariant core positions. See **Figure 5**. Superposed coordinate sets and PyMol session files are available from the FIT tab of the web-app.

Figure 5: Superposed PDBs colored from N-terminal (blue) to C-terminal (red) alignment position.

The pair-wise RMSD distribution of superposed structures is shown in **Figure 6**.

A hiearchical cluster analysis of these RMSD values was performed with the **ward.D2** method yielding a dendrogram that was partitioned into **7** major cluster groups. See **Figure 7**.

4 PCA: Principal Component Analysis

Principal component analysis (PCA) was used to provide a lower dimensional representation of the superposed structure set that usefully summarizes inter-conformer relationships (Grant et al. 2006). Applying PCA to all **30** structures revealed that **62.77%** of the total coordinate variance can be

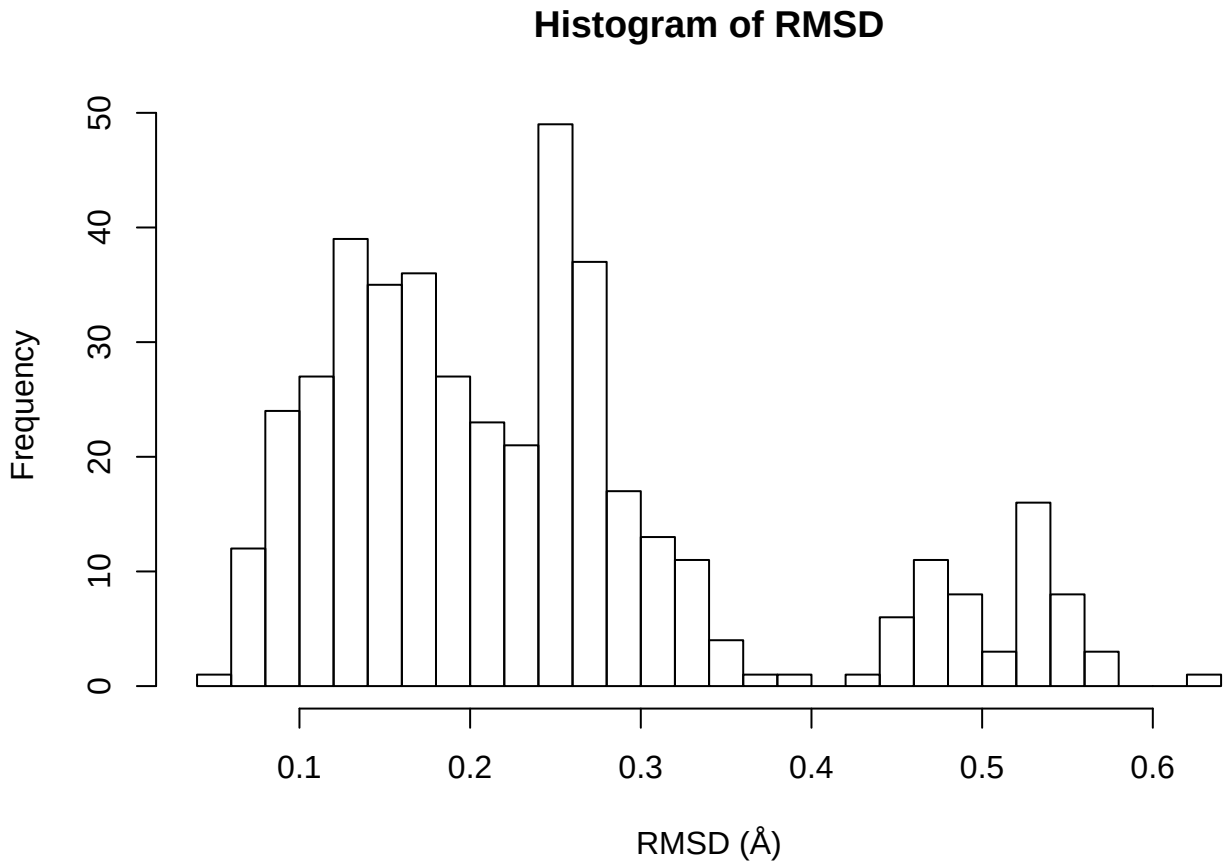

Figure 6: Histogram of pair-wise RMSD values determined from superposed structures.

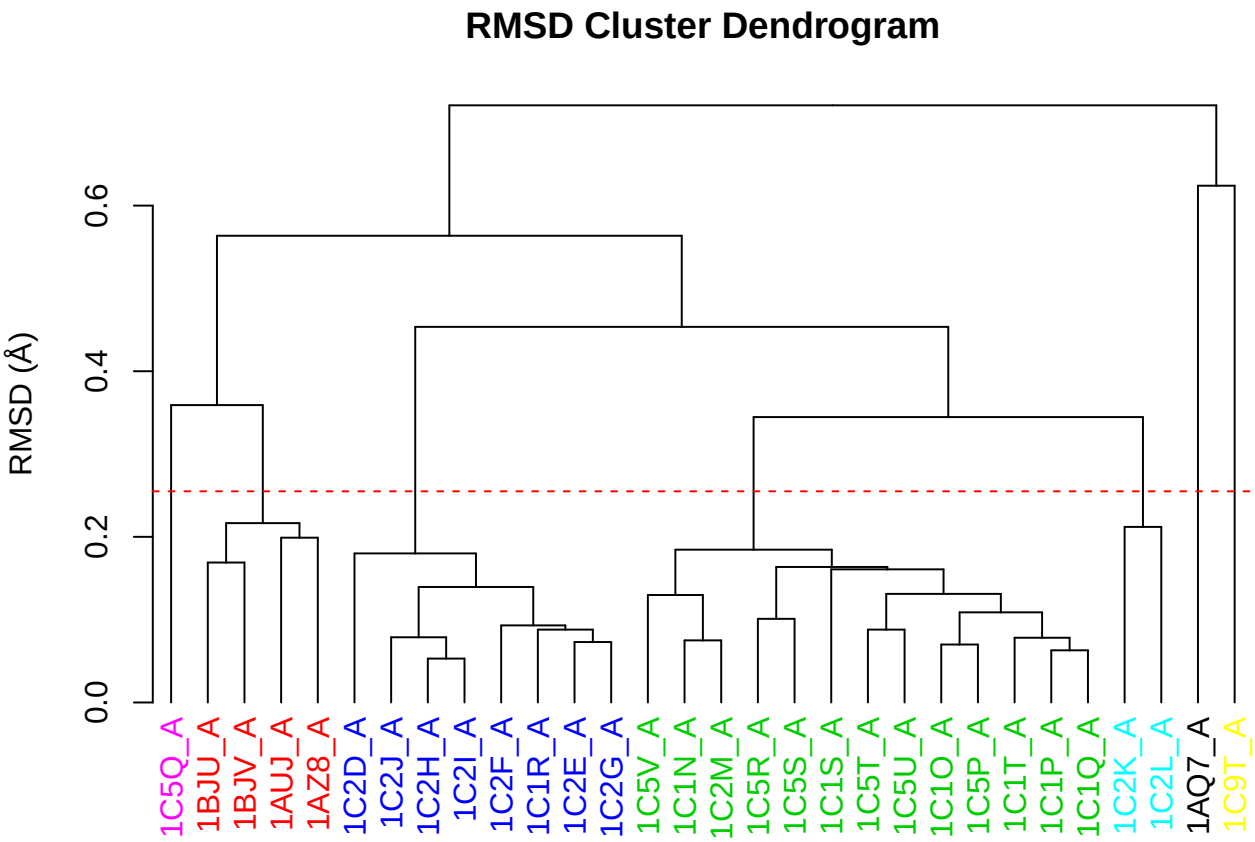

Figure 7: RMSD Cluster Dendogram.

captured in three dimensions (**27.76%** in the first PC, **19.83%** in the second, and **15.18%** in the third; see **Figure 8B** for a so-called scree plot or egienvalue spectrum of these values).

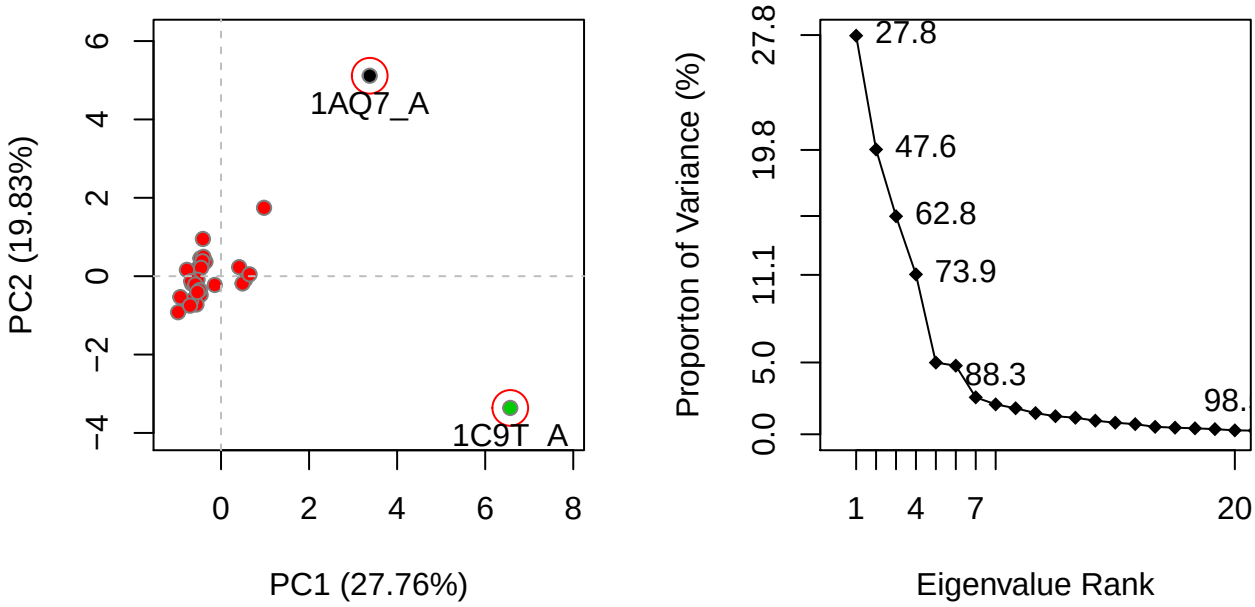

Figure 8: (A) Conformer plot: projection of all structures onto the principal planes defined by the user selected principal components (termed PCs). (B) Eigenvalue spectrum: results obtained from diagonalization of the covariance matrix of superposed coordinates. The magnitude of each eigenvalue is expressed as the percentage of the total variance (mean-square fluctuation) captured by the corresponding eigenvector. Labels beside each point indicate the cumulative sum of the proportion of the total variance accounted for in all preceding eigenvectors.

A projection of the structures onto **PC1** (X-axis) and **PC2** (Y-axis) (that collectively account for **47.59%** of mean square displacements in the original coordinate data) is also shown in **Figure 8A**. Points represent individual structures and are colored by user defined cluster groups from either **PC subspace**, RMSD or sequence identity clustering (user selection in bold) as selected in the PCA tab of the web-app. Both sequence identity and RMSD clustering have been described separately above. PC subspace clustering is described separately further below.

Clustering in the **PC1–2** space with the **ward.D2** method yields a cluster dendrogram that can be partitioned into **3** cluster groups. See **Figure 9**.

Residue contributions to the first 3 PCs are shown in **Figure 10**.

Structural displacements from the mean structure along **PC1** are shown in **Figure 11**. Coordinate trajectory files are available from the PCA tab of web-app.

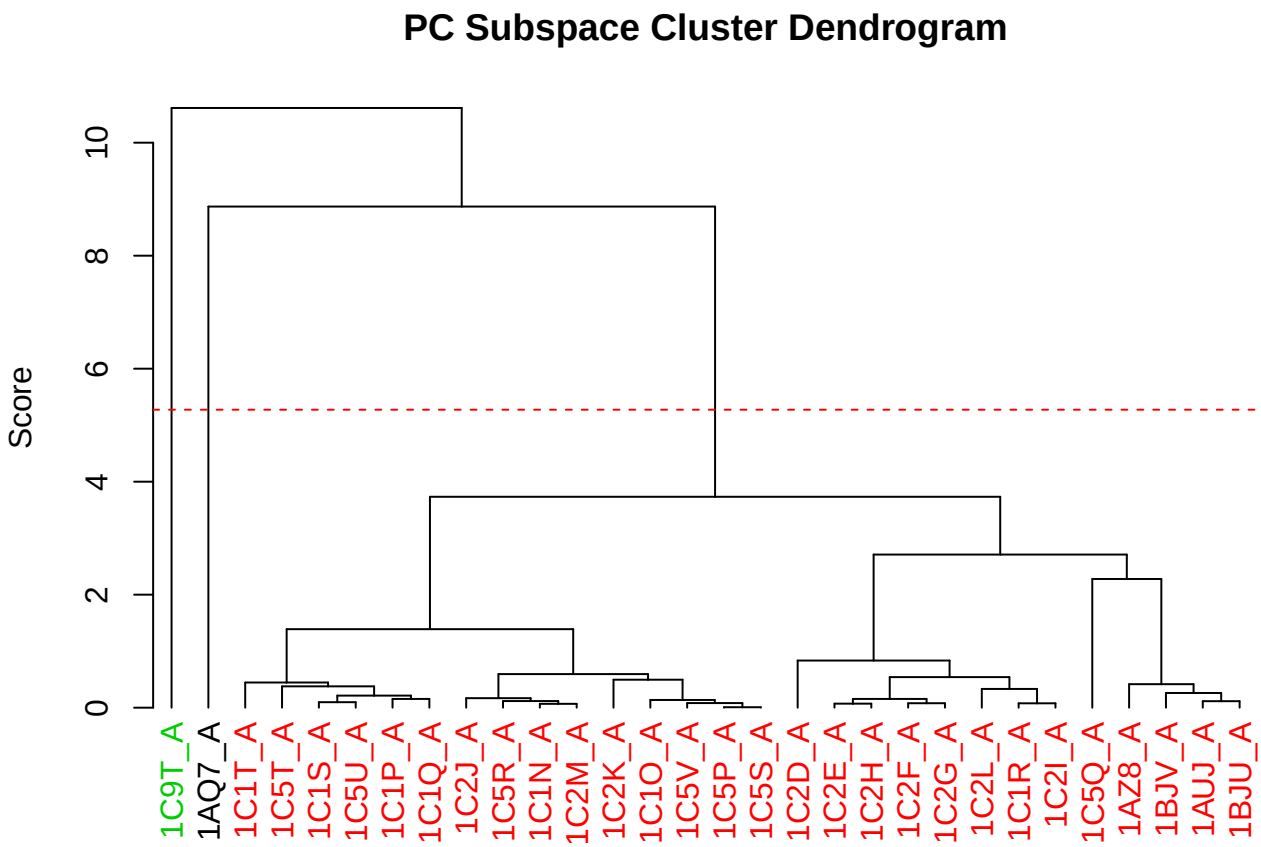

Figure 9: PCA Cluster Dendrogram.

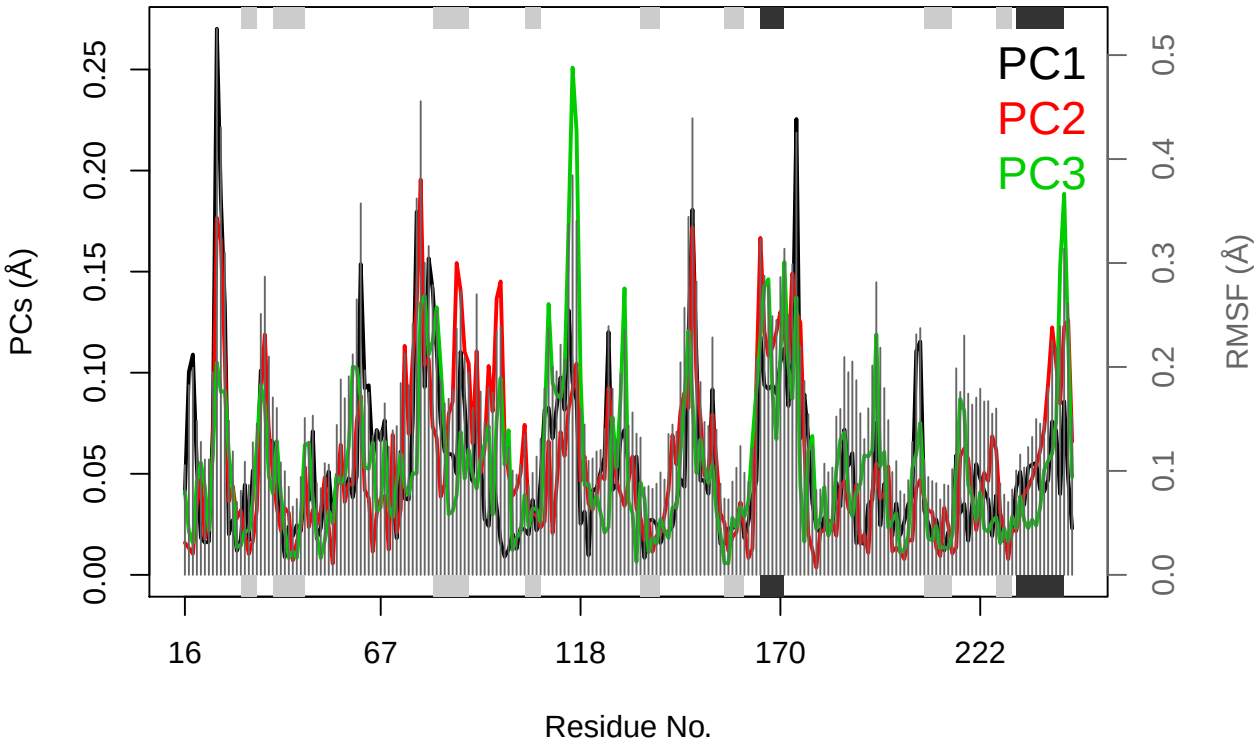

Figure 10: The contribution of each residue to selected principal components (PCs).

Figure 11: PC-1 Displacement Trajectory. Each principal component is represented as equidistant atomic displacements from the mean structure. Displacements are scaled by the standard deviation of the distribution along a given principal component.

5 NMA: Normal Mode Analysis

Normal Mode Analyis (NMA) was performed to predict large-scale motions on selected structures in the ensemble in a way that facilities the interoperation of structural similarity and dissimilarity trends. The following structures were selected for ensemble NMA (eNMA) based on **PC** cutoff of **5**: 1AQ7_A, 1C2J_A, 1C9T_A, 1C5Q_A, 1AUJ_A, 1AZ8_A.

A visualization of non-trivial mode **1** of structure **1C2J_A** is show in **Figure 12**. Visualization of normal modes obtained from the other structures, as well as coordinate trajectory files, are available from the eNMA tab of the web-app.

Figure 12: Non-trivial Mode-1 of 1C2J_A. Visualization is provided as a trajectory.

NMA derived atomic fluctuations are shown for each structure in **Figure 13**. The fluctuation profiles are colored by cluster groups from **PC subspace** based clustering with **3** groups (**Figure 13**).

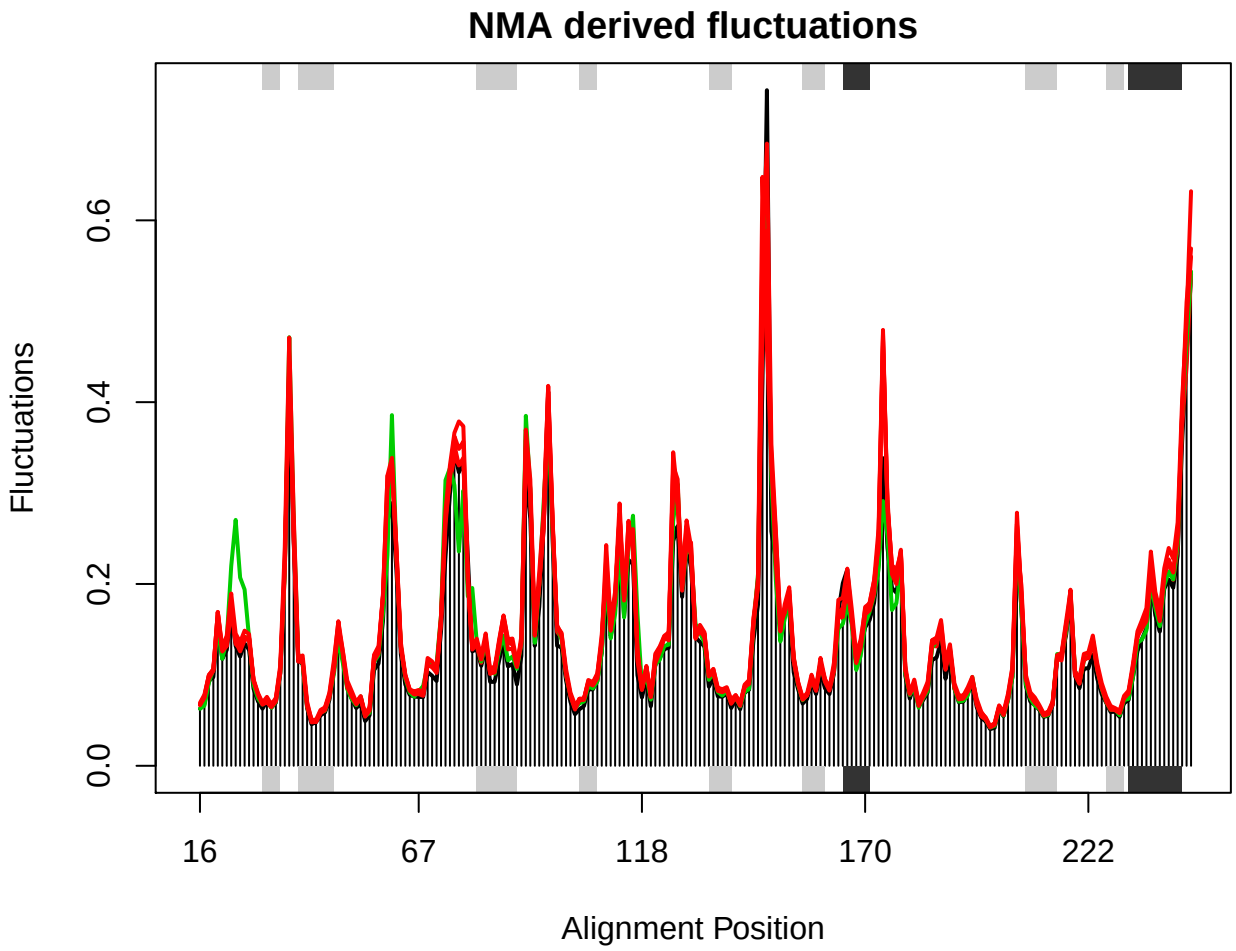

Figure 13: Predicted fluctuations, colored by PC subspace clustering.

A heatmap of Root mean square inner product (RMSIP) values, **Figure 14A**, shows the relationship between the first 10 non-trivial modes of each of the structures selected for NMA. A heatmap of heirarchical cluster analysis on all pair-wise RMSD values from the FIT tab of the web-app is shown in **Figure 14B**.

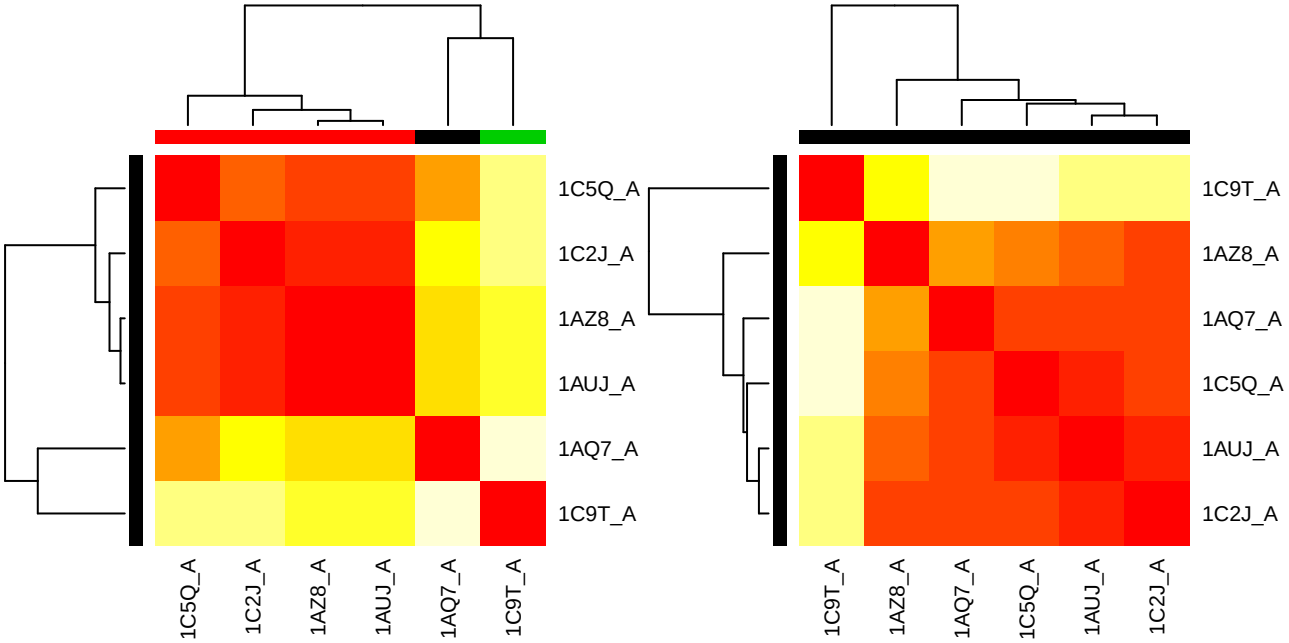

Figure 14: (A) Heatmap of cluster analysis performed on RMSIP values of first 10 modes of selected structures. (B) Heatmap of PC subspace cluster analysis.

The normal modes of structure **1C2J__A** are compared to the principal components of the filtered structure ensemble (see **Figure 15**). The comparison over the 10 first modes yields an RMSIP of **0.29**. The highest overlap is given by **PC-2** and **NM-5** with a value of **0.1**. The overlap is a similarity score where orthogonal vectors give a score of 0 and identical vectors give a score of 1.

Mode **19** of PDB ID **1C5Q__A** shows the highest individual overlap with a value of **0.12**. PDB ID **1AQ7__A** shows the highest cumulative overlap after 5 modes with a value of **0.21**. Overlap analysis determines the agreement between the normal mode vectors and a given conformational change (i.e. the difference between the PDB structures being compared). See **Figure 16**.

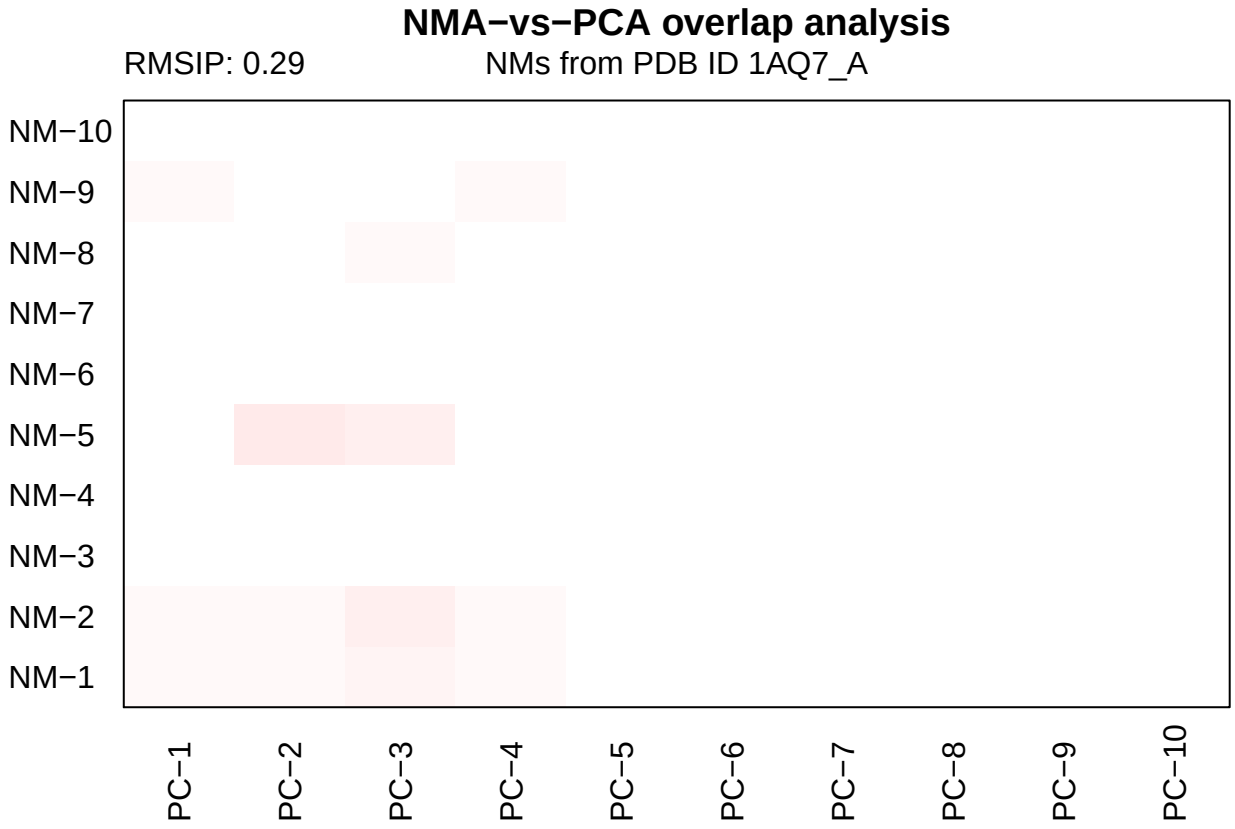

Figure 15: Comparison of calculated normal modes of **1C2J__A** with principal components obtained from the ensemble from PCA tab.

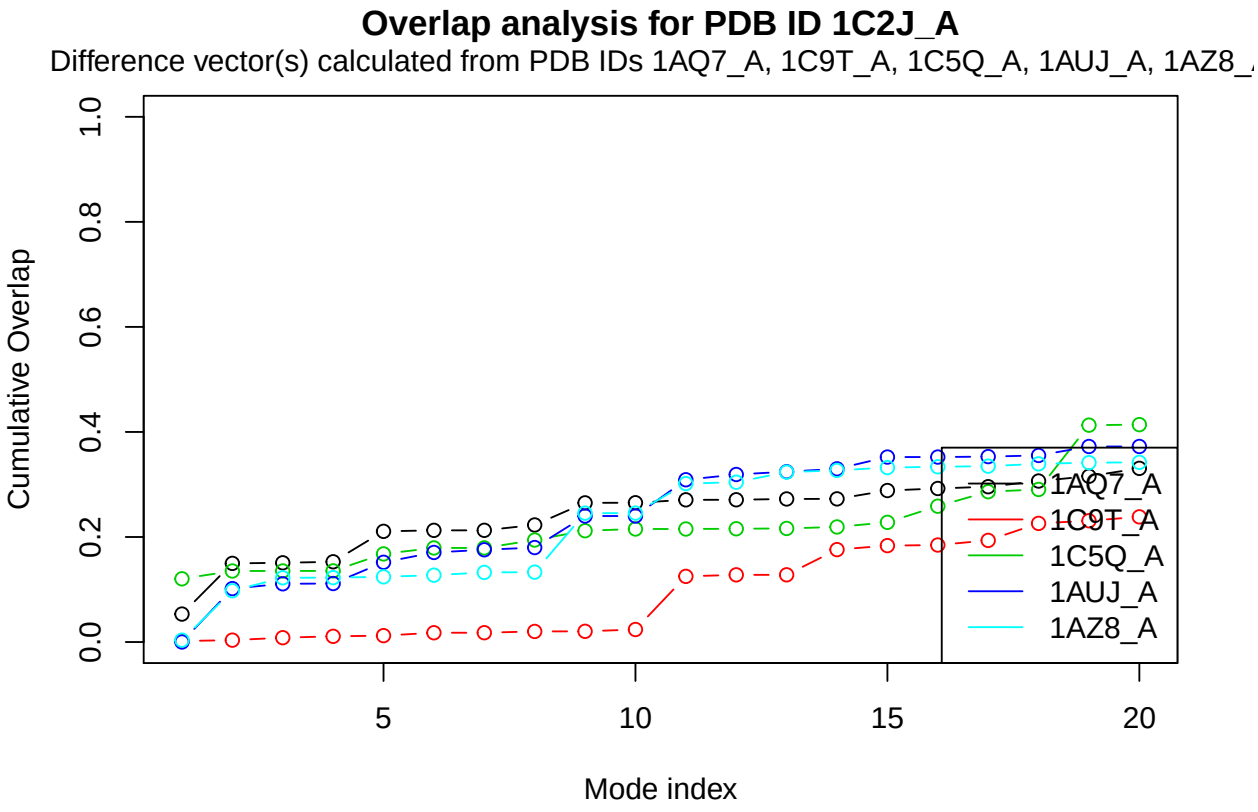

Figure 16: Overlap analysis. The plot shows the cumulative overlap values calculated by the dot product between individual modes of **1C2J__A** and the conformational difference vector between the reference structure **1C2J__A** and PDB IDs **1AQ7__A, 1C9T__A, 1C5Q__A, 1AUJ__A, 1AZ8__A**.

Conventional Usage Example

To read your selected input structure (with PDB ID: PDBCODE) into Bio3D directly you can use the following command sequence:

```
# NOTE: Replace PDBCODE with your chosen 4 character PDB ID
library(bio3d)
pdb <- read.pdb("PDBCODE")
summary(pdb)
```

To search the online RCSB PDB database with the sequence of your query structure you could use the following commands:

```
# Use hmmer or blast
blast <- blast.pdb(pdbseq(pdb))
hits <- plot(blast)
```

To download and align the identified structures you can use the following commands:

```
# Use the optional 'path' input argument to set a specific a download location
files <- get.pdb(hits$pdb.id, split=TRUE)
pdbs <- pdbaln(files)
```

For rigid core identification and structural superposition use:

```
core <- core.find(pdbs)

# Use the optional 'outpath' argument to write superimposed PDBs to disk
xyz <- pdbfit(pdbs, core)
```

Investigate pairwise structural deviations and perform cluster analysis with:

```
rd <- rmsd(xyz)
hc <- hclust(as.dist(rd))
hclustplot(hc, k=2)
```

Perform principal component analysis with:

```
pc <- pca(xyz)
plot(pc)
mktrj(pc)
```

To perform normal mode analysis use:

```
modes <- nma(pdbs)
plot(modes)
mktrj(modes)
```

Citation information

This Bio3D web-app should be referenced with the URL http://thegrantlab.org/bio3d/webapps and the following citation: Skærven, L., Jariwala, S., Yao, X.-Q., & Grant, B. J. (2016). Online interactive analysis of protein structure ensembles with Bio3D-web. Bioinformatics. doi: 10.1093/bioinformatics/btw482.

Session and Software Version Information

This report was auto-magically generated by Bio3D along with the additional R packages noted below.

You can install and run **Bio3D-web** locally by following these instructions.

```
## R version 3.3.1 (2016-06-21)
## Platform: x86_64-redhat-linux-gnu (64-bit)
## Running under: Red Hat Enterprise Linux Server release 6.8 (Santiago)
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
```

```
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] knitr_1.12       RMySQL_0.10.3    DBI_0.3.1        RCurl_1.95-4.6
##  [5] bitops_1.0-6     rmarkdown_0.9.5  abind_1.4-3      maptools_0.8-36
##  [9] sp_1.1-0         reshape2_1.4.1   rgl_0.93.963     lattice_0.20-33
## [13] bio3d_2.2-2.9000 shinyBS_0.61     threejs_0.2.1    rCharts_0.4.5
## [17] shinyRGL_0.1.1   DT_0.1.40        shiny_0.13.2
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.5      highr_0.5.1      formatR_1.2.1    plyr_1.8.3
##  [5] base64enc_0.1-2 shinyjs_0.5.2    tools_3.3.1      digest_0.6.9
##  [9] evaluate_0.8     jsonlite_0.9.21 yaml_2.1.13      parallel_3.3.1
## [13] stringr_1.0.0    htmlwidgets_0.6 grid_3.3.1       R6_2.1.1
## [17] foreign_0.8-66  RJSONIO_1.3-0    pander_0.5.2     magrittr_1.5
## [21] whisker_0.3-2    htmltools_0.3.5 mime_0.4         xtable_1.8-0
## [25] httpuv_1.3.3     stringi_1.0-1    miniUI_0.1.1     Cairo_1.5-6
```

References

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. 2000. "The Protein Data Bank." *Nucleic Acids Research* 28 (1): 235–42. doi:10.1093/nar/28.1.235.

Eddy, Sean R. 2011. "Accelerated Profile HMM Searches." *PLoS Comput Biol* 7 (10): e1002195. doi:10.1371/journal.pcbi.1002195.

Finn, Robert D., Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Andreas Heger, et al. 2014. "Pfam: The Protein Families Database." *Nucleic Acids Research* 42 (D1): D222–30. doi:10.1093/nar/gkt1223.

Grant, B.J., A.P.D.C Rodrigues, K.M. Elsawy, A.J. Mccammon, and L.S.D. Caves. 2006. "Bio3D: An R Package for the Comparative Analysis of Protein Structures." *Bioinformatics* 22: 2695–96. doi:10.1093/bioinformatics/btl461.

Skjaerven, L., X.Q. Yao, G. Scarabelli, and B.J. Grant. 2015. "Integrating Protein Structural Dynamics and Evolutionary Analysis with Bio3D." *BMC Bioinformatics* 15: 399. doi:10.1186/s12859-014-0399-6.