Journal of
CHEMICAL INFORMATION
AND MODELING

Application Note

pubs.acs.org/jcim

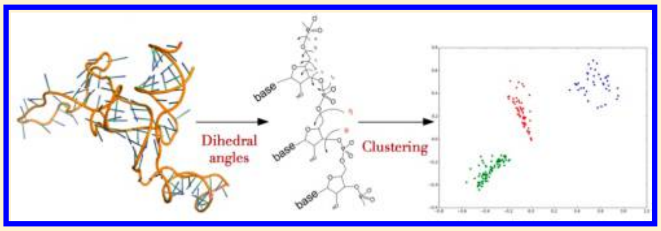
Clustangles: An Open Library for Clustering Angular Data

Karen Sargsyan,*,[†] Yun Hao Hua,[†] and Carmay Lim*,[†,‡]

[†]Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan

[‡]Department of Chemistry, National Tsing Hua University, Hsinchu 300, Taiwan

ABSTRACT: Dihedral angles are good descriptors of the numerous conformations visited by large, flexible systems, but their analysis requires directional statistics. A single package including the various multivariate statistical methods for angular data that accounts for the distinct topology of such data does not exist. Here, we present a lightweight standalone, operating-system independent package called Clustangles to fill this gap. Clustangles will be useful in analyzing the ever-increasing number of structures in the Protein Data Bank and clustering the copious conformations from increasingly long molecular dynamics simulations.



INTRODUCTION

Analysis of the vast numbers of conformations sampled by intrinsically disordered proteins and large, flexible macromolecules from long molecular dynamics (MD) simulations is crucial for understanding the dynamics and function of these biologically important systems. The different conformations differ more in their dihedral angles than their bond lengths or bond angles, hence dihedral angles are good descriptors of different protein, DNA, or RNA conformations. For example, only two torsion angles (ϕ and φ) per residue can describe the protein backbone conformation,[1] whereas two pseudotorsion angles (η and θ) per nucleotide can represent the RNA backbone conformation.[2] Using dihedral angles instead of Cartesian coordinates focuses on the internal degrees of freedom pertinent to the structural dynamics of the system and dramatically reduces the number of variables needed to represent the conformations of large biomolecules.[3] It also eliminates the need to optimally align conformations to remove the translation and overall rotation of the system in order to capture only its internal motion. Hence, compared to Cartesian coordinates, dihedral angles are more attractive in describing large-scale motions of biomolecules and can yield more accurate results: Principal component analysis (PCA) using backbone dihedral angles (termed dPCA) has been found to yield correct free energy landscapes for peptides, small proteins, and RNA, but PCA using Cartesian coordinates leads to artifacts as the rotational fitting procedure cannot correctly and completely remove the overall rotation.[4−7] Clustering dihedral angles using a circular version of k-means has also been shown to be useful in understanding entropic costs of ordering for intrinsically disordered proteins.[8]

Because of the topological differences between circular (angles) and linear data (Cartesian coordinates), multivariate statistical methods such as PCA and k-means for linear data cannot be directly applied to circular data. This is because rules for simple statistical measures (e.g., mean and variance) of linear data do not apply to circular data. For example, the arithmetic mean of 5° and 355° is 180° rather than the true mean of 0° and the sum of angles in a right-angled triangle on a sphere is not equal to 180°. Hence, applying conventional clustering methods to circular data may lead to wrong results. Although several packages such as *circular*[9] and *isocir*[10] in R, *CircStat* in Matlab,[11] as well as *Stata*[12] and *Oriana*[13] under Windows can be used for statistical analysis of circular data, they provide functions for descriptive statistics (e.g., mean/median, variance, trigonometric measurements) and inferential statistics (e.g., circular uniformity, mean/median direction) rather than methods to cluster circular data.

Despite the difficulties with circular data, a few multivariate statistical methods for circular data have been developed. One approach termed dPCA[4,5] adapts PCA for angular data by transforming angles into coordinates using cosine and sine values. As an example, the two backbone dihedral angles ϕ_i and ψ_i of residue i can be replaced by four coordinates $x_{4i-3} = \cos(\phi_i)$, $x_{4i-2} = \sin(\phi_i)$, $x_{4i-1} = \cos(\psi_i)$ and $x_{4i} = \sin(\psi_i)$. Hence, dPCA doubles the number of variables compared to the original set of dihedral angles. It also neglects the $\cos^2 x + \sin^2 x = 1$ correlation of angular data, but whether this neglect would lead to artifacts remains unknown, as applications of dPCA have so far been successful.[14,15] An alternative approach to dPCA that uses non-Euclidean geometry geared for angular data is GeoPCA:[16] Instead of determining a set of ordered orthogonal linear axes that represents decreasing proportions of the data variation, GeoPCA finds a set of ordered orthogonal great circles (principal component geodesics) that minimize the distances from the data points to their projections on the respective great circles (see Figure 1). The distance between any two points is an arc in GeoPCA rather than a straight line. GeoPCA has been applied in conformational analysis of RNA nucleotides[16] and flavin adenine nucleotides.[17]

Received: May 26, 2015
Published: August 4, 2015

DOI: 10.1021/acs.jcim.5b00316

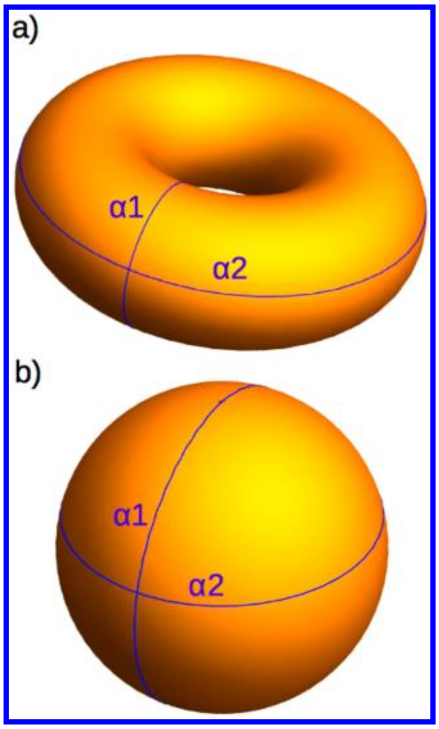


Figure 1. Possible topologies of a space of observed conformations as points with coordinates (α_1,α_2), where α_1,α_2 correspond to rotations along the given circles and are angles we choose to describe a conformation. Each clustering method is modified to account (a) topology corresponding to a torus and (b) topology corresponding to a hypersphere.

Because the multivariate methods for circular data exist as separate pieces of code written in different programming languages, we have developed a package called Clustangles for researchers to analyze and cluster the numerous conformations from MD trajectories using angular data. Our goal is not to outperform the available libraries and/or standalone software for circular statistics and angular clustering, but to provide the most useful methods for analyzing/clustering angular data in a standalone lightweight open package with minimum dependencies. To make the package independent of the operating system version in use, we implement Clustangles in Python. This language was chosen because it is widely used among scientists and is used in a plethora of high-quality open packages, which can be applied along with Clustangles. Furthermore, when tested on MD trajectories, Python could perform analysis and clustering fast enough to be acceptable for all practical purposes. To make our package widely accessible, we provide a simple Python console interface that does not require programming skills for all the methods. The packages and their documentation are freely available at http://clustangles.limlab.ibms.sinica.edu.tw.

INPUT

The input to Clustangles is dihedral angles describing each conformation from MD trajectories. These angles can be extracted using the MDtraj python package,[18] as illustrated in the Clustangles documentation. Each data set is represented internally as a 2D matrix (numpy object, from numpy Python package), where columns correspond to angles and rows to frames or observations. Clustangles recognizes several file formats such as .csv (comma separated values), output of AMBER[19] traj tool for dihedral angles, as well as dihedral angle outputs from GROMOS[20] and CHARMM.[21] To read a data set from a .csv file, create an object with an empty data set, then read data from the file as

angle = Angles()
angle.read_csv(input-filename, units = “degrees”)

The default input angles are in degrees, but the program can convert angles in degrees to radians using the command *angle.to_radians*() or the default could be changed by the option, units = “radians”. The *angle* object contains useful attributes and methods itself and is passed as a parameter to other methods. For convenience sake, it is possible to merge data sets from different objects in row or column format, as described in the documentation. It is also possible to extract selected columns from a given data set for further analysis as

angle_new = angle.select(0, 3, 6)

This will create a new *angle_new* object containing data from the first, fourth, and seventh columns extracted from the *angle* object. The list of columns to choose could be arbitrarily long. In all examples herein, we assume that the Clustangles package is imported in Python in such a way that the command names are not appended with the package name.

OUTPUT

Clustangles provides as output circular statistical measures, principal components from dPCA, or principal component geodesics (see below). It contains ways to write all relevant information in a .csv file or directly store pieces of information as attributes. One way is to use *angle.write(output-filename)*, which writes data set to a file for the objects. This attribute exists for all objects. To minimize dependencies on external software packages that require installation, we deliberately did not provide accessories for plotting figures in Clustangles. However, figures can easily be plotted by combining outputs of the methods and utilities from other python packages or nonpython standalone software. Examples of how to plot graphs are given in the documentation.

CIRCULAR STATISTICS

Clustangles provides useful circular statistical measures such as mean direction, circular variance, circular standard deviation, and circular correlation. We refer the user to previous works[22,23] for the definitions of these quantities, but a short description of the meaning of these measures could be obtained by invoking the command *help(method_name)*. As an example, for *large_mww_test*, the command *help(large_mww_test)* gives the following:

“Large-sample Mardia−Watson−Wheeler could be applied to several independent samples to establish whether these samples are drawn from a common distribution. It requires sample length to be greater than 10 elements. The null hypothesis is: The distribution is common.”

Values of descriptors for a particular column may be obtained; for example, the command mean_direction (angle, 0) provides a circular mean direction (in units of *angle.unit*) for the first column of the data set in the object *angle*.

For quantities requiring two columns (e.g., correlation), the following format should be used:

circular_correlation (angle_first, 0, angle_second, 2), where the first column is taken from the data set in *angle_first* and the third column from *angle_second*.

Clustangles also provides methods for hypothesis testing of circular statistics results. These include Rayleigh tests for uniformity of circular data with specified and nonspecified mean direction, tests for a common mean direction, tests for a common median direction, and tests for a common distribution. Several statistical tests may accept more than two samples, hence the user provides an *Angles()* object containing a data set, where columns are samples. As an example, large_mww_test(angle) provides not only the *p*-value for the Mardia−Watson−Wheeler test (see above) on samples given by the *angle* object, but some of its interpretation as well. An example output might be

"The *p*-value is 0.1202. Because the *p*-value is >0.05, the null hypothesis that two or more samples are drawn from a common distribution holds."

PRINCIPAL COMPONENT ANALYSIS

Clustangles incorporates two different versions of PCA for angular data: dPCA and GeoPCA.

dPCA. dPCA converts angular data into Cartesian coordinates by means of sine and cosine transformations (see the Introduction). We implement both covariance and correlation matrix approaches, as it is not always immediately clear which one is preferable and the correlation matrix tends to inflate the contributions of variables with small variance and reduce the influence of variables with large dimensions.[24] As in the case of atomic coordinates, the success of dPCA in dimension reduction such as determining the essential motions of biomolecules depends on sufficient conformational sampling to construct the covariance/correlation matrix. Our implementation of dPCA includes two measures for sampling significance; viz., the Kaiser−Meyer−Olkin (KMO) score and the associated measure of sampling adequacy (MSA), which should ideally be greater than half.[24] Besides adequate sampling, the number of principal components to efficiently describe the motion of the macromolecule can be estimated from (i) the Cattell criterion[25] or (ii) a specific amount of variance. Hence, Clustangles provides the "kink" in the scree plot together with the number of principal components to describe *x*% of the total variance, where *x* is user-specified.

Clustangles provides advice on how to interpret results of dPCA by the advice() command. For example, dp = dpca(angle, "corr") first derives all necessary information for performing dPCA on a data set stored in the object *angle* and stores it in object *dp*. If the option "corr" is chosen, then correlation matrix is used, otherwise dpca uses covariance matrix by default.

The dp.advice(60) command provides advice on the number of principal components and quality of sampling, as well as how many principal components to choose to describe at least 60% of the total variance. For a data set consisting of 820 columns and 10,000 rows of random angles, the dPCA output might yield:

"The KMO value, 0.454747, is smaller than 0.5, so there might be problems with sampling. The MSA is [0.45349001 0.46180901 0.45607339 ..., 0.45691787 0.44243296 0.45704818]. According to the elbow method applied to the scree plot, choose 1,639 principal components to account 100% variance, and 704 for 60% of the total variance."

Such a large number of principal components to account for 60% of the total variance is not surprising because the input angles were random.

After choosing a suitable number of principal components, say first three, the user might project a data set from object *b* onto the first three principal components as

projected = dp.project(3, b)

GeoPCA. Unlike dPCA, GeoPCA represents angular data as points on a unit hypersphere and determines principle component geodesics.[16] We separate mapping of data points onto a unit hypersphere from computation of the principal component geodesics because different mappings might be of interest. The default mapping is available as a function *unitsphere()*:

gp = geopca(angle.unitsphere())

No advice on how to interpret results of GeoPCA is given, as the method implemented currently provides only projection onto the first two principal component geodesics. The output of *geopca()* is an object containing all relevant information about these two principal component geodesics. One may project an angular data set from object *b* onto the two principal component geodesics using *.project(b)*.

CLUSTERING

After performing dimensionality reduction, clustering is needed to separate conformations into distinct groups. Clustangles provides clustering techniques for Cartesian coordinates, which are needed following dPCA. It also provides two circular analogs of hierarchical and *k*-means clustering using angular data. One analog treats a sequence of angles for a particular conformation $(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,...,\boldsymbol{\alpha}_n)$ as a point on a high-dimensional torus, where each $\boldsymbol{\alpha}_k$ is a rotation along the corresponding great circle, as shown in Figure 1a. The distance between two points $\boldsymbol{\alpha},\boldsymbol{\beta}$ is given by

$$D(\alpha, \beta) = \sqrt{\sum_{i=1}^{n} ((\sin(\alpha_i) - \sin(\beta_i))^2 + (\cos(\alpha_i) - \cos(\beta_i))^2)} \tag{1}$$

The other circular analog of hierarchical and *k*-means clustering treats $(\boldsymbol{\alpha}_1,\boldsymbol{\alpha}_2,...,\boldsymbol{\alpha}_n)$ as a point on a high-dimensional unit sphere (hypersphere, Figure 1b). The distance between two points $\boldsymbol{\alpha},\boldsymbol{\beta}$ is given by[16]

$$D(\alpha, \beta) = \arccos\langle\alpha, \beta\rangle \tag{2}$$

As for GeoPCA, we decouple mapping onto the unit sphere and clustering on the unit sphere.

Besides circular analogs of hierarchical and *k*-means clustering, Clustangles also provides two soft expectation−maximization (EM)-based clustering techniques for points on a unit sphere. Soft clustering not only separates points into clusters, but also provides the probabilities for points to belong to each cluster. The two EM-based clustering techniques are based on clustering on a unit hypershere using von Mises−Fisher[26] or Watson[27] distributions. The von Mises−Fisher distribution may be viewed as an analog of the Gaussian distribution for a unit hypersphere and inherently models only circular or tight clusters. The Watson distribution is considered to be more capable of modeling clusters with nonstandard shape than the von Mises−Fisher distribution.[28]

CONCLUDING REMARKS

The methods implemented in Clustangles package contain tests for individual components yielding a unit test coverage over

90%. They have also been tested for CPU performance on MD trajectories, each of which contains an average of 10,000 configurations. The dihedral angles of each MD conformation were extracted using MDtraj and used as input to Clustangles. dPCA on a data set containing 8,200,000 angles for a total of 10,000 configurations took 15 s on a 2.3 GHz Intel Core i5 machine. Other methods also demonstrate comparably fast execution times. Although it is possible to add parts of the code in C to increase performance, the execution times provided by the Python code would fit the need of most users. Our main policy for Clustangles is to keep the package lightweight, independent of the operating system with minimal dependence on external packages.

AUTHOR INFORMATION

Corresponding Authors

*K. Sargsyan. E-mail: karsar@ibms.sinica.edu.tw.
*C. Lim. E-mail: carmay@gate.sinica.edu.tw.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank Cédric Grauffel and Jon Wright for providing MD trajectories, Jon Wright for helping with the web server, and Karine Mazmanian for her assistance with the TOC figure. Supported by funds from Academia Sinica and MOST (Grant # 98-2113-M-001-011), Taiwan.

REFERENCES

(1) Ramachandran, G. N.; Sasisekharan, V. Conformation of Polypeptides and Proteins. *Adv. Protein Chem.* **1968**, *23*, 283−438.

(2) Duarte, C. M.; Wadley, L. M.; Pyle, A. M. RNA Structure Comparison, Motif Search and Discovery using a Reduced Representation of RNA Conformational Space. *Nucleic Acids Res.* **2003**, *31*, 4755−4761.

(3) Kuppuraj, G.; Sargsyan, K.; Hua, Y.-H.; Merrill, A. R.; Lim, C. Linking Distinct Conformations of Nicotinamide Adenine Dinucleotide with Protein Fold/Function. *J. Phys. Chem. B* **2011**, *115*, 7932−7939.

(4) Mu, Y.; Nguyen, P. H.; Stock, G. Energy Landscape of a Small Peptide Revealed by Dihedral Angle Principal Component Analysis. *Proteins: Struct., Funct., Bioinf.* **2005**, *58*, 45−52.

(5) Altis, A.; Nguyen, P. H.; Hegger, R.; Stock, G. Dihedral Angle Principal Component Analysis of Molecular Dynamics Simulations. *J. Chem. Phys.* **2007**, *126*, 244111-1−244111-10.

(6) Riccardi, L.; Nguyen, P. H.; Stock, G. Free-energy Landscape of RNA Hairpins Constructed via Dihedral Angle Principal Component Analysis. *J. Phys. Chem. B* **2009**, *113*, 16660−8.

(7) Sittel, F.; Jain, A.; Stock, G. Principal Component Analysis of Molecular Dynamics: On the Use of Cartesian vs. Internal Coordinates. *J. Chem. Phys.* **2014**, *141*, 014111.

(8) Cukier, R. I. Dihedral Angle Entropy Measures for Intrinsically Disordered Proteins. *J. Phys. Chem. B* **2015**, *119*, 3621−34.

(9) Agostinelli, C.; Lund, U. *R Package 'Circular': Circular Statistics*, 0.4−7; California, U.S.A., 2013.

(10) Barragan, S.; Fernandez, M.; Rueda, C.; Peddada, S. isocir: An R Package for Constrained Inference Using Isotonic Regression for Circular Data, with an Application to Cell Biology. *J. Stat. Softw.* **2013**, *54*, 1−17.

(11) Berens, P. CircStat: A MATLAB Toolbox for Circular Statistics. *J. Statistical Software* **2009**, *31*, 1−21.

(12) Cox, N. *CIRCSTAT: Stata Modules to Calculate Circular Statistics*; Boston College: Chestnut Hill, MA , 2004.

(13) *Oriana - Circular Statistics for Windows*, version 3.0; Kovach Computer Services: Pentraeth, Wales, U. K., 2009.

(14) Hinsen, K. Comment on: "Energy Landscape of a Small Peptide Revealed by Dihedral Angle Principal Component Analysis. *Proteins: Struct., Funct., Bioinf.* **2006**, *64*, 795−797.

(15) Mu, Y.; Nguyen, P.; Stock, G. Reply to the Comment on "Energy Landscape of a Small Peptide Revealed by Dihedral Angle Principal Component Analysis. *Proteins: Struct., Funct., Bioinf.* **2006**, *64*, 798−799.

(16) Sargsyan, K.; Wright, J.; Lim, C. GeoPCA: A New Tool for Multivariate Analysis of Dihedral Angles Based on Principal Component Geodesics. *Nucleic Acids Res.* **2012**, *40*, e25.

(17) Kuppuraj, G.; Kruise, D.; Yura, K. Conformational Behavior of Flavin Adenine Dinucleotide: Conserved Stereochemistry in Bound and Free states. *J. Phys. Chem. B* **2014**, *118*, 13486−97.

(18) McGibbon, R. T.; Beauchamp, K. A.; Schwantes, C. R.; Wang, L.-P.; Hernandez, C. X.; Herrigan, M. P.; Lane, T. J.; Swails, J. M.; Pande, V. S. MDTraj: A Modern, Open Library for the Analysis of Molecular Dynamics Trajectories. *bioRxiv* **2014**, DOI: 10.1101/008896.

(19) Pearlman, D. A.; Case, D. W.; Caldwell, J. W.; Ross, W. R.; Cheatham, T. E., I.; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, A Computer Program for Applying Molecular Mechanics, Normal Mode Analysis, Molecular Dynamics and Free Energy Calculations to Elucidate the Structures and Energies of Molecules. *Comput. Phys. Commun.* **1995**, *91*, 1−41.

(20) Scott, W. P. R.; Henberger, P. H.; Tironi, I. G. The GROMOS Biomolecular Simulation Program Package. *J. Phys. Chem. A* **1999**, *103*, 3596−3607.

(21) Brooks, B. R.; Brooks, C. L., III; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30*, 1545−1615.

(22) Pewsey, A.; Neuhäuser, M.; Ruxton, G. D. *Circular Statistics in R*; Oxford University Press: Oxford, U. K., 2013.

(23) Mardia, K. V.; Jupp, P. *Directional Statistics*, 2nd ed.; John Wiley & Sons Ltd: New York, NY, 2000.

(24) David, C. C.; Jacobs, D. J. Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins. *Methods Mol. Biol.* **2014**, *1084*, 193−226.

(25) Cattell, R. B. The Scree Test for the Number of Factors. *Multivariate Behav. Res.* **1966**, *1*, 245−276.

(26) Banerjee, A.; Dhillon, I. S.; Ghosh, J.; Sra, S. Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *J. Mach. Learn. Res.* **2005**, *6*, 1345−1382.

(27) Sra, S.; Karp, D. The Multivariate Watson Distribution: Maximum-likelihood Estimation and Other Aspects. *J. Multivariate Anal.* **2013**, *114*, 256−269.

(28) Bijral, A. S.; Breitenbach, M.; Grudic, G. In *Proceedings of the Eleventh International Conference on Artificial Intelligence & Statistics (AISTATS-07)*, Meila, M.; Shen, X., Eds.; *J. Mach. Learn. Res.*: **2007**; Vol. 2, pp 35−42.