

GeoPCA: a new tool for multivariate analysis of dihedral angles based on principal component geodesics

Karen Sargsyan^{1,*}, Jon Wright¹ and Carmay Lim^{1,2,*}

¹Institute of Biomedical Sciences, Academia Sinica, Taipei 115 and ²Department of Chemistry, National Tsing Hua University, Hsinchu 300, Taiwan

Received June 26, 2011; Revised October 27, 2011; Accepted October 28, 2011

ABSTRACT

The GeoPCA package is the first tool developed for multivariate analysis of dihedral angles based on principal component geodesics. Principal component geodesic analysis provides a natural generalization of principal component analysis for data distributed in non-Euclidean space, as in the case of angular data. GeoPCA presents projection of angular data on a sphere composed of the first two principal component geodesics, allowing clustering based on dihedral angles as opposed to Cartesian coordinates. It also provides a measure of the similarity between input structures based on only dihedral angles, in analogy to the root-mean-square deviation of atoms based on Cartesian coordinates. The principal component geodesic approach is shown herein to reproduce clusters of nucleotides observed in an η - θ plot. GeoPCA can be accessed via <http://pca.limlab.ibms.sinica.edu.tw>.

INTRODUCTION

Multivariate statistics is widely applied to biological systems. It is used to unravel hidden trends in large data sets and to analyze the results of molecular dynamics simulations of biomolecules. Among the wide range of available multivariate techniques, principal component analysis (PCA) (1) is one of the most widely used methods. PCA transforms a data set consisting of several correlated variables into a new set of uncorrelated variables called principal components. By a linear orthogonal transformation, the first principal component represents the most variability in the data; the second principal component represents the second most variability in the

data under the constraint that it is orthogonal to the first principal component, and so on. Thus, PCA rotates the axes of data variation, yielding a set of ordered orthogonal axes that represents decreasing proportions of the data variation. Using only the first few principal components, the dimensionality of the transformed data is reduced. For example, the first few principal components have been used to specify a set of representative coordinates of the free energy landscape for biological molecules containing many degrees of freedom (2). They have also been used to yield the dominant modes of structural variation in an ensemble of conformations for a given protein, derived from Nuclear Magnetic Resonance (NMR) and/or X-ray (3); i.e. structures of the free protein solved in different space groups or complexed with different ligands or from simulations (4,5).

In PCA of large biomolecules with many degrees of freedom, it is useful to replace the Cartesian coordinates of the atoms with a smaller set of internal coordinates to reduce the number of variables involved in PCA. A natural choice of internal coordinates would be dihedral angles that change much more than bond lengths and bond angles in structures of a given molecule. However, angular data pose difficulties in PCA and other multivariate statistical analyses due to their circular nature. For example, the arithmetic mean of 10° and 350° is $(10^\circ + 350^\circ)/2 = 180^\circ$ rather than the true mean of 0° . This difficulty remains even if the torsion angles are represented in the interval from -180° to 180° , as the arithmetic mean of -160° and 160° is 0° instead of 180° .

To circumvent the aforementioned difficulties with circular data, angles have been transformed into coordinates using cosine and sine values in PCA (referred to as dPCA in previous work) (2,6). For example, the two backbone dihedral angles ϕ_i and ψ_i of residue i have been replaced by four coordinates $x_{4i-3} = \cos(\phi_i)$,

*To whom correspondence should be addressed. Tel: +886 2 2789-9043; Fax: +886 2 2788 7641; Email: karsar@ibms.sinica.edu.tw
Correspondence may also be addressed to Carmay Lim. Tel: +886 2 2789-9043; Fax: +886 2 2788 7641; Email: carmay@gate.sinica.edu.tw

$x_{4i-2} = \sin(\phi_i)$, $x_{4i-1} = \cos(\psi_i)$ and $x_{4i} = \sin(\psi_i)$. Thus, one disadvantage of the dPCA approach is an increased number of coordinates. Another disadvantage is the neglect of the $\cos^2 + \sin^2 = 1$ correlation (7,8); i.e. the coordinates are not independent since $(x_{4i-3})^2 + (x_{4i-2})^2 = 1$ and $(x_{4i-1})^2 + (x_{4i})^2 = 1$. Furthermore, there is no rigorous mathematical consideration of the applicability of dPCA (to the best of our knowledge). In justifying the transformation of angular data (6), the points are assumed to lie on a sphere in Euclidean space, subject to Euclidean geometry, whereas they should in fact be subjected to non-Euclidean geometry (9). Notably, some properties of non-Euclidean geometry are counterintuitive; e.g. the sum of the angles of a triangle is $>180^\circ$. Furthermore, since all data points are on a sphere, the distance measured by the shortest path between points is an arc rather than a straight line.

Circular correlation and covariance matrices computed using known formulas for the circular correlation coefficient and the circular mean, respectively, can be used in PCA (10). PCA has been applied to torsion angles of a set of RNA trinucleotides using five different representations; viz., (i) angles between 0° and 360° , (ii) angles between -180° and 180° , (iii) angles represented by cosine and sine values (see dPCA above), (iv) circular correlation matrix and (v) circular covariance matrix (11). The results were compared with those from PCA applied to Cartesian coordinates of the same data set of RNA trinucleotides. The outcome of the PCA results was found to depend on the choice of interval for representing the angles [$0^\circ, 360^\circ$] or [$-180^\circ, 180^\circ$]. Thus, for each torsion angle, its variance has to be analyzed *a priori* to determine if it should be represented by a ($0^\circ, 360^\circ$) or ($-180^\circ, 180^\circ$) interval. The interval that yields the larger total variance of the first principal component was assumed to be more accurate. Moreover, using a linear orthogonal transformation in PCA, the non-Euclidean nature of the circular data was not taken into account.

Various manifold (locally Euclidean space) learning and non-linear dimensionality reduction approaches may be considered as alternatives to linear PCA for angular data. These include self-organizing maps (12), principal curves (13), kernel PCA (14), isomap (15), diffusion maps (16) and principal geodesics (17). Most of them apply machine learning such as neural networks. For most of these methods, there is no simple interpretation of the results unlike linear principal components. Furthermore, these methods have not been used in lieu of linear PCA for dihedral angles (to the best of our knowledge).

Our aim is to develop a tool applying a generalization of PCA for angular data. Among the various manifold learning and non-linear dimensionality reduction approaches, geodesic PCA was chosen because (i) it is a straightforward generalization of PCA for manifolds that are generally only locally Euclidean and (ii) the mathematics underlying principal component geodesic has been described (17). Instead of determining a set of ordered orthogonal linear axes, which represents decreasing proportions of the data variation, we find a set of ordered orthogonal great circles (principal component geodesics)

that minimizes the distances from the data points to their projections on the respective great circles. The distance between any two data points is an arc rather than a straight line, as in linear PCA.

Below, we first present the essence of the principal component geodesic approach and the properties of principal geodesic components; we refer the reader to previous works for proofs of the necessary theorems (17). We then validate the principal component geodesic approach by using it to cluster a set of RNA conformations that had been classified as follows. Just as the protein backbone conformation can be described by two torsion angles (ϕ and φ), the RNA backbone conformation can be described by two pseudotorsion angles ($\eta = C4'_{i-1}-P_i-C4'_i-P_{i+1}$ and $\theta = P_i-C4'_i-P_{i+1}-C4'_{i+1}$) and the sugar pucker, instead of the seven conventional torsion angles, α , β , γ , δ , ϵ , ζ and χ (Figure 1). A plot of the θ versus η angles of all nucleotides in a database containing 52 RNA structures revealed distinct clusters of nucleotides (18). Within a given cluster, the nucleotides share similar η and θ values as well as structural features such as A-platforms and GNRA tetraloops. These clusters of nucleotides have been statistically validated and refined using a larger data set containing 73 RNA structures (19). This work shows that the principal component geodesic approach provides a means of distinguishing clusters of nucleotides using seven conventional torsion angles per nucleotide. Its application is

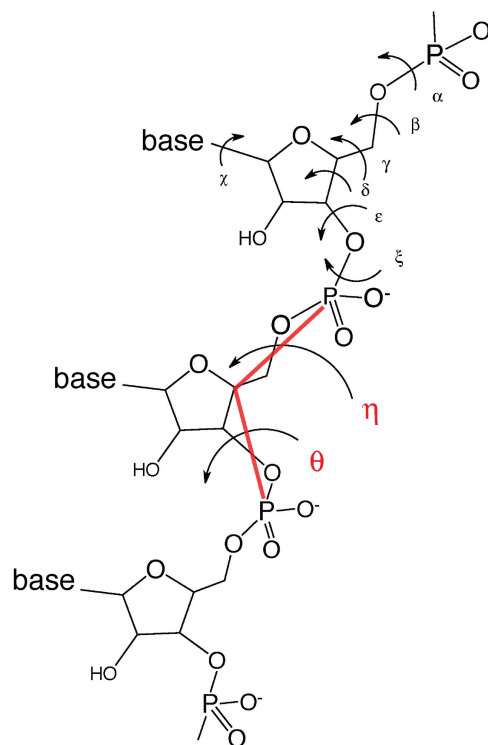


Figure 1. Structure of RNA nucleotides showing the seven conventional torsion angles, α , β , γ , δ , ϵ , ζ , χ and two pseudotorsion angles, η , θ , which are defined as $C4'_{i-1}-P_i-C4'_i-P_{i+1}$ and $P_i-C4'_i-P_{i+1}-C4'_{i+1}$, respectively. The bold red lines are pseudobonds connecting P and C4' along the backbone.

not limited to dihedral angles or nucleotides, but can be applied to analyze angular data of large, complex macromolecules.

METHODS

Embedding of an m -Sphere in $(m+1)$ -dimensional Euclidean space

Nash theorem (20) postulates that every Riemannian manifold M can be isometrically embedded in a Euclidean space of sufficiently higher dimension. Thus, an m -dimensional unit sphere can be embedded in a $(m+1)$ -dimensional Euclidean space. It is defined by

$$\phi(x) := \langle x, x \rangle - 1 = 0, \quad (1)$$

where x are points in the $(m+1)$ -dimensional Euclidean space and $\langle \bullet, \bullet \rangle$ denote the inner product. The inner product, $\langle x, x \rangle$, in an m -dimensional sphere is equal to the scalar or dot product in the $(m+1)$ -dimensional Euclidean space. Thus, although the geometry on an m -sphere is non-Euclidean, it can be described in terms of $(m+1)$ -dimensional Euclidean space since the m -sphere has been embedded in $(m+1)$ -dimensional Euclidean space. The tangent space of the unit sphere S at x is defined as the set of all tangent unit vectors v (Figure 2a) that satisfy

$$2\langle x, v \rangle = 0, \langle v, v \rangle = 1 \quad (2)$$

Input data

Let $P = (p_1, p_2, \dots, p_n)$ denote a set of torsion angle measurements describing a molecule of interest. Each p_i represents the i th conformation of the molecule. If P contains n conformations, there will be n observations for each torsion angle in the molecule. Let a_i^k ($k = 1, \dots, m$) denote the value of k th torsion angle of p_i . Each $p_i = (a_i^1, a_i^2, \dots, a_i^m)$ can be treated as a point on the m -dimensional unit sphere, representing the i th conformation. For our test data set (see below), the nucleotides all have the same C3'-endo sugar pucker conformation. Hence, the input data consist of the seven conventional torsion angles in Figure 1 describing each nucleotide.

Geodesics

Instead of using straight-line axes as principal components, curves are used as principal components, so-called principal component geodesics. A geodesic is a curve on the m -dimensional sphere, which locally minimizes the distance between points on the surface. It is a straight line in the plane and a great circle (like the earth's equator) on a sphere. Just as the distance between two points, a and b , in Euclidean space can be represented by a straight line of the form $b + (t-1)(b-a)$, $0 \leq t \leq 1$, geodesics on spheres are great circles given by

$$\gamma_{x,v}(t) = a \cos t + b \sin t \quad (3a)$$

where $0 \leq t \leq 2\pi$, $a = x$ and $b = v$, and Equations (1) and (2) are satisfied; i.e.

$$\gamma_{x,v}(t) = x \cos t + v \sin t \quad (3b)$$

Spherical distance to a geodesic

The embedding of an m -dimensional sphere into a $(m+1)$ -dimensional Euclidean space induces a simple expression for a metric on the sphere. Since the inner product $\langle a, b \rangle$ is the standard scalar product of the $(m+1)$ -dimensional Euclidean space, the spherical distance $d(a, b)$ between any two points a and b (Figure 2a) on the m -dimensional unit sphere is given by:

$$d(a, b) = \arccos \langle a, b \rangle \quad (4)$$

The projection of point a onto the geodesic $\gamma_{x,v}$ is the point a' (Figure 2b) given by

$$d = \frac{\langle x, a \rangle x + \langle v, a \rangle v}{\sqrt{\langle x, a \rangle^2 + \langle v, a \rangle^2}}. \quad (5)$$

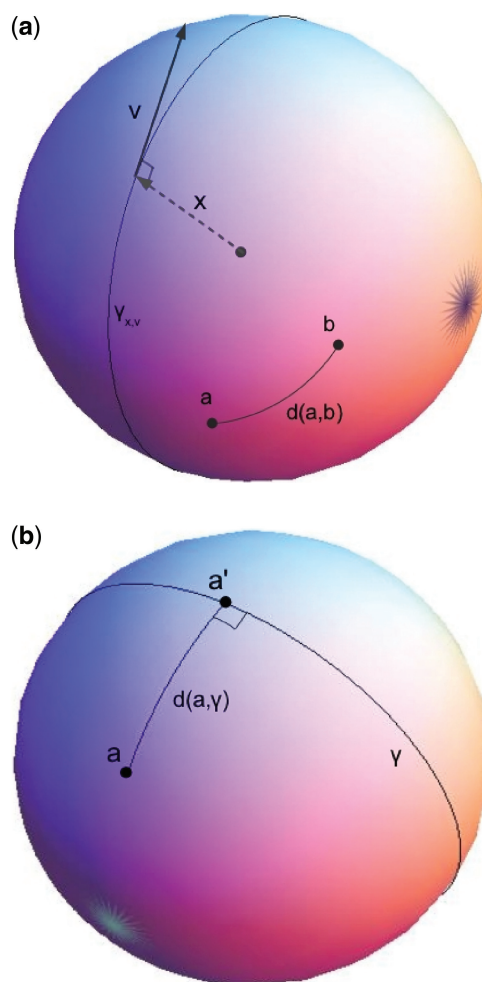


Figure 2. (a) The x, v vectors, the $\gamma_{x,v}$ geodesic and the spherical distance $d(a, b)$ between two points a and b are illustrated on a 2D unit sphere embedded in 3D Euclidean space. Although point x is on the sphere, its Euclidean radius-vector is in the direction from the sphere center to x (dashed line) and is therefore *not* on the sphere. The x vector is orthogonal to the Euclidean tangent vector v , which is in the direction of a path on the sphere and is on the sphere itself. Both x and v are vectors in $(m+1)$ -dimensional Euclidean space and are thus subjected to Euclidean geometry. (b) The spherical distance between point a and its projection onto the geodesic $\gamma_{x,v}$, point a' , are illustrated on a 2D unit sphere embedded in 3D Euclidean space.

Thus, the spherical distance between point a and its projection onto the geodesic $\gamma_{x,v}$, point a' , is given by

$$d(a, \gamma_{x,v}) = \arccos \sqrt{\langle a, x \rangle^2 + \langle a, v \rangle^2} \quad (6)$$

Principal component geodesics

Consider the following distance function that describes the mean distance between data points p_i and their projections onto the geodesic $\gamma_{x,v}$.

$$F(x, v) = \sum_{i=1}^n d(p_i, \gamma_{x,v})^2 \quad (7)$$

where $d(p_i, \gamma_{x,v})$ is given by Equation (6). Finding a first principal component geodesic that accounts for most of the data variability is equivalent to minimizing $F(x, v)$ under the constraints given by Equations (1) and (2). Given the first principal component geodesic $\gamma_{x,v}^{(1)}$, the second principal component geodesic, $\gamma_{x,v}^{(2)}$, can be found as a geodesic that intersects $\gamma_{x,v}^{(1)}$ and is orthogonal to $\gamma_{x,v}^{(1)}$ by minimizing $F(x, v)$ with the respective constraints.

To obtain the other principal component geodesics, we define a principal component geodesic mean, \hat{p} , as the point that minimizes the mean of $d(p_i, z)^2$ over all common points of $\gamma_{x,v}^{(1)}$ and $\gamma_{x,v}^{(2)}$. A principal component geodesic of higher order s ($3 \leq s \leq n$) minimizes the function $F(x, v)$, passes through the principal component geodesic mean and is orthogonal to all geodesics of order $\leq s-1$.

Geodesic variance

The variance explained by the s th principal component geodesic, obtained by projection of the data points p_i on the s th principal component geodesic, is given by:

$$V^{(s)} = \frac{\sum_{i=1}^n d(p_i^{(s)}, \hat{p})^2}{n}, \quad (8)$$

where $p_i^{(s)}$, the projection of p_i on the s th principal component geodesic, is obtained using Equation (5), and n is the number of conformations of a given molecule (see above). As in conventional PCA, the first principal component geodesic represents the most variability in the data and has the smallest variance. However, if its variance was comparable with the variance of a randomly chosen geodesic on the sphere, then the principal component geodesic analysis would not help to reduce the dimensionality of the given input data.

Output data

The above approach has been implemented in a program called GeoPCA. In the current version of GeoPCA, Lagrange multipliers are used to minimize $F(x, v)$ under constraints, as described by Huckemann and Ziezold (17). This procedure yields fixed-point equations, $y = f(y)$, which are solved by numerical iteration, $y_{n+1} = f(y)$. After solving the fixed-point equations, GeoPCA

provides projection of the data onto the first two principal component geodesics and the corresponding Cartesian coordinates of the data points projected on the unit sphere in 3D space to enable plots to be made using standard plotting packages. Thus, GeoPCA allows visualization of the output data along the great circles, which accounts for most of the data variability.

Data set

To validate the principal component geodesic approach, it was used to cluster a set of RNA conformations derived from a published database of 73 RNA structures containing 7407 nt (19). We did not update this database so that clusters obtained from a plot of the first two principal component geodesics, which are characterized by two 'principal' angles, can be compared with the clusters found in an η - θ plot of all non-helical nucleotides with C3'-endo sugar pucker from the published database [see Figure 4 in Wadley *et al.* (19)]. The latter yielded six clusters of non-helical C3'-endo nucleotides, which were labeled as I, II, III, IV, V and VI by Wadley *et al.* (19). We chose to include in our data set non-helical C3'-endo nucleotides in clusters I and II, as they have the highest density in the η - θ plot, ensuring that they are statistically significant and should be detected by any effective clustering method. Furthermore, cluster I contains nucleotides that are often constituents of S1 and S2 motifs, while cluster II contains nucleotides that serve as the second bases in GNRA/GNRA-like tetraloops or in T-loop motifs. Unlike the nucleotides in cluster I or cluster II, nucleotides in cluster V do not belong to structural motifs. Hence, cluster V was also included in our data set to verify if the principal component geodesic approach, like the η/θ plot, also predicts this cluster.

The non-helical C3'-endo nucleotides belonging to clusters I, II and V were extracted from the published data set of 7407 nt as follows: First, the η/θ values corresponding to the peak density of a given cluster was found from an initial guess of the 'peak' η/θ values from the η - θ plot of all non-helical C3'-endo nucleotides and refining them to yield the maximum number of nucleotides for a given cluster. Then, all C3'-endo nucleotides with η/θ values within $\pm 15^\circ$ of the 'peak' η/θ values were extracted; i.e. (147–187°)/(330–360°) for cluster I, (15–45°)/(225–255°) for cluster II and (299–329°)/(216–246°) for cluster V. This yielded 59, 88 and 43 nt for clusters I, II and V, respectively (Supplementary Table S1). This data set was used to test whether the principal component geodesic approach can yield the three clusters found in the η - θ plot.

RESULTS AND DISCUSSION

dPCA using dihedral angles

Using our database (see above), we first examined if dPCA ('Introduction' section) using the seven standard dihedral angles (α , β , γ , δ , ϵ , ζ and χ in Figure 1), represented by cosine and sine values but neglecting the $\cos^2 + \sin^2 = 1$ correlation, could yield the three distinct clusters

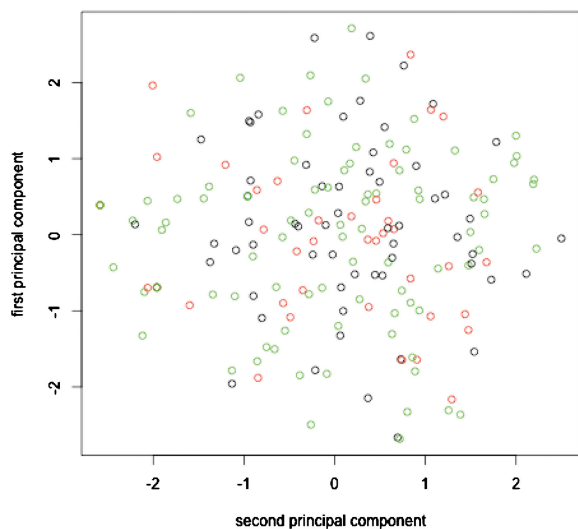


Figure 3. Plot based on the angles describing the first and second principal components obtained using dPCA. The black, green and red circles denote non-helical C3'-endo nucleotides belonging to clusters I, II and V found in an η - θ plot of all non-helical C3'-endo nucleotides in previous work.

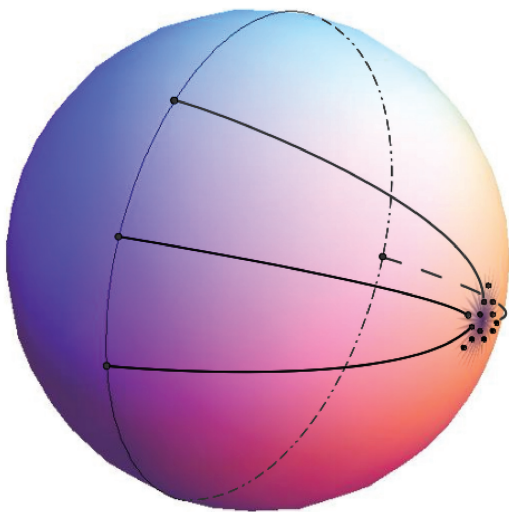


Figure 4. An example of a possible arrangement of the cluster along the circle after projection of the 2D sphere onto 1D circle.

(I, II and V, see 'Data set' section) found in an η - θ plot in previous work (19). A 2D plot of the first two principal components in Figure 3 shows that dPCA cannot distinguish non-helical C3'-endo nucleotides belonging to cluster I (black circles), II (green circles) and V (red circles).

Clustering patterns in non-Euclidean space

Before presenting the clustering results using the principal component geodesic approach, we first highlight some differences between the clustering patterns in Euclidean and non-Euclidean space. Whereas data points belonging to a cluster appear close to one another in Euclidean space,

they may be dispersed along a circle in non-Euclidean geometry. The latter becomes evident if data points clustered at the pole of a sphere are projected onto the big circle (equator) of that sphere. As shown in Figure 4, although points around the pole are close to each other, their projections on the equator may cover the entire big circle. Hence, points lying on a circle can be identified as forming a cluster. This property has the advantage that points lying on different circles can be unambiguously assigned to different clusters except for those at the intersection of circles.

Geodesic PCA using dihedral angles

Geodesic PCA was performed using the seven standard dihedral angles for each nucleotide in our data set. Figure 5 shows the projection of the non-helical C3'-endo nucleotides on a sphere and the first two principal component geodesics (blue great circles). The results in Figure 5 show that geodesic PCA can separate the nucleotides into three distinct clusters, as observed in a η - θ plot. The C3'-endo nucleotides from cluster I (black points) and cluster II (green points) lie close to the first two principal geodesic components (Figure 5a) and are at different distances from the sphere center, so they are well separated from each other. However, the C3'-endo nucleotides from cluster V (red points) are not visible from this viewpoint, but become evident from another viewpoint (Figure 5b). Although the red circles do not form a compact cluster, they are nevertheless clearly separated from the nucleotides in the other two clusters.

That the C3'-endo nucleotides form three clusters are also shown when the data points on the sphere are projected onto a plane. Figure 6a shows the non-helical C3'-endo nucleotides as a function of the first two principal component geodesics, which can be described by two 'principal' angles. The C3'-endo nucleotides from cluster V (red points) are located at the top and bottom of the 2D plot. They are well separated from the C3'-endo nucleotides from cluster II (green points) and cluster I (black points), which lie along two great circles.

Geodesic PCA using pseudotorsion angles

Since the input data for the principal geodesics approach (seven torsion angles) differs from that for the η - θ plot (two pseudotorsion angles), the outcome from these two methods would not be expected to be identical. Indeed, the three distinct clusters found herein do not contain exactly the same nucleotides as clusters I, II and V from an η - θ plot in previous work (19). For example, C3'-endo nucleotides from clusters I (black circles) and V (red circles) are found along the big circle encompassing C3'-endo nucleotides from cluster II (green circles), as shown in Figure 6. To verify that this discrepancy is not due to limitations/errors in the GeoPCA program, geodesic PCA was performed with the two pseudotorsion angles, η and θ , used to derive clusters I, II and V in previous work (19). Note that the two 'principal' angles describing the first two principal component geodesics do not correspond to η and θ . Thus, although the 2D plot in

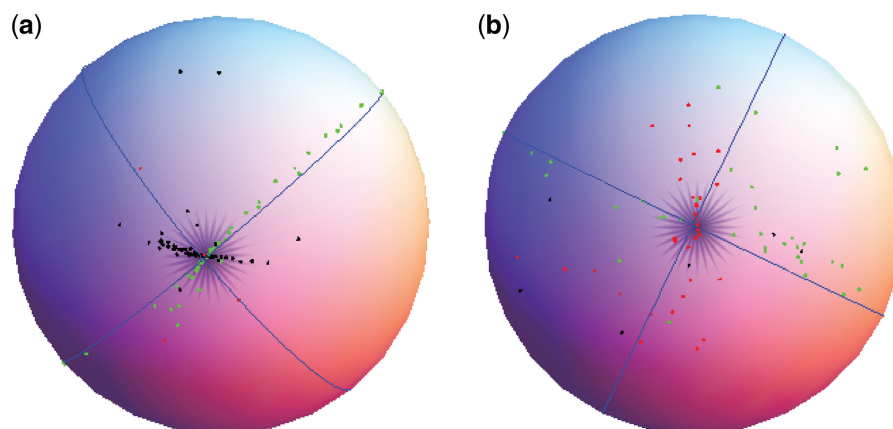


Figure 5. C3'-endo nucleotides on a sphere. The first two principal geodesic components are shown in blue. The (a) shows C3'-endo cluster I (black points) and cluster II (green points), while the (b) shows the C3'-endo cluster V (red points).

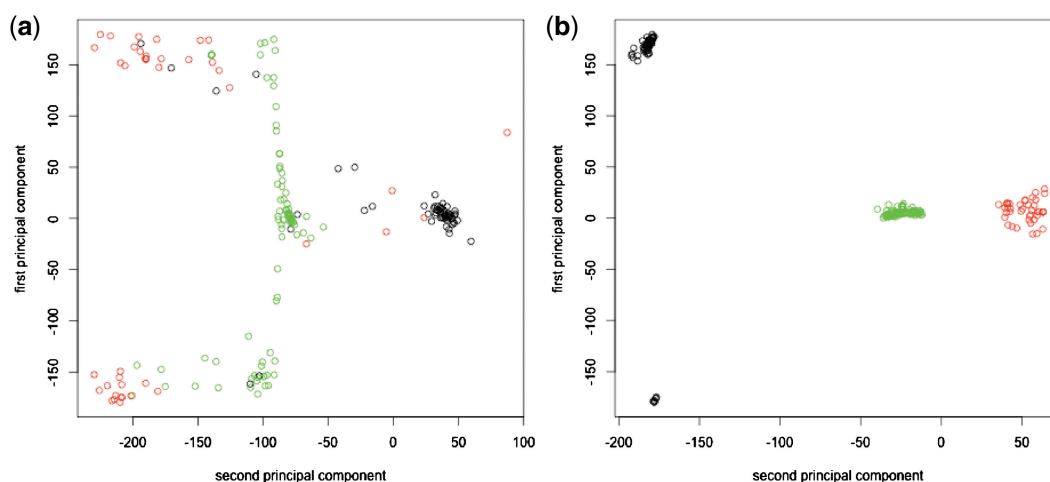


Figure 6. 2D plot based on the angles describing the first and second principal component geodesics using (a) the seven conventional torsion angles and (b) the two pseudotorion angles, η and θ (Figure 1). The black, green and red points denote non-helical C3'-endo nucleotides belonging to clusters I, II and V found in an η - θ plot of all non-helical C3'-endo nucleotides in previous work. Principal component geodesics are lines parallel to the axes, passing through zeros of the respective axes and described by dotted lines.

Figure 6b is not the same as an η - θ plot, the same three clusters found in an η - θ plot are found.

SUMMARY

This work introduces a new tool, based on principal component geodesics, for conformational analysis using circular data such as bond, torsion and pseudotorion angles. It shows how our approach could aid structural analysis like analyses of η - θ plots and counterintuitive consequences of non-Euclidean geometry (e.g. points lying on a circle belong to the same cluster). The web interface of GeoPCA, which implements the principal component geodesics approach described herein, requires as input, a file with angular data. It yields as output: (i) Cartesian coordinates of the data points projected on the first and second principal component geodesics of a sphere (orthogonal great circles on the sphere) and (ii) the values of two angles representing corresponding distances to first and second principal component geodesics for each data

point. To the best of our knowledge, this is the first method to automatically reduce a multidimensional analysis of several angles to only two angles containing most of the information. GeoPCA thus provides a useful way of visualizing, analyzing and predicting conformations of complex macromolecules with many degrees of freedom.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1.

ACKNOWLEDGEMENTS

We would like to thank Kevin S. Keating and Anna Marie Pyle for providing information on the nucleotides used in our data set. The authors thank Karine Mazmanian for her help in preparation of the figures.

FUNDING

Funding for open access charge: National Science Council, Taiwan (grant NSC 95-2113-M-001-038 and NSC 95-2113-M-001-001 to C.L.).

Conflict of interest statement. None declared.

REFERENCES

- Jolliffe, I.T. (2002) *Principal Component Analysis*. Springer, New York.
- Mu, Y., Nguyen, P.H. and Stock, G. (2005) Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *PROTEINS: Structure, Function, and Bioinformatics*, **58**, 45–52.
- Yang, L.W., Eyal, E., Bahar, I. and Kitao, A. (2009) Principal component analysis of native ensembles of biomolecular structures (PCA NEST): insights into functional dynamics. *Bioinformatics*, **25**, 606–614.
- Ichiye, T. and Karplus, M. (1991) Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins*, **11**, 205–271.
- Bahar, I. and Rader, A. (2005) Coarse-grained normal mode analysis in structural biology. *Curr. Opin. Struct. Biol.*, **15**, 586–592.
- Altis, A., Nguyen, P.H., Hegger, R. and Stock, G. (2007) Dihedral angle principal component analysis of molecular dynamics simulations. *J. Chem. Phys.*, **126**, 244111.
- Hinsen, K. (2006) Comment on: energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins*, **64**, 795–797.
- Mu, Y., Nguyen, P. and Stock, G. (2006) Reply to the comment on Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins*, **64**, 798–799.
- Petersen, P. (2006) *Riemannian Geometry*. Springer, Berlin.
- Mardia, K.V. and Jupp, P. (2000) *Directional Statistics*, 2nd edn. John Wiley & Sons Ltd, Chichester.
- Reijmers, T.H., Wehrens, R. and Buydens, L.M.C. (2001) Circular effects in representations of an RNA nucleotides data set in relation with principal components analysis. *Chemometrics and Intelligent Laboratory Systems*, **56**, 61–71.
- Kohonen, T. (2001) *Self-organizing Maps*. Springer, Berlin.
- Kégl, B. (1999) *Principal Curves: Learning, Design, and Applications*. Concordia University, Canada.
- Schölkopf, B., Smola, A. and Müller, K.-R. (1997) Kernel principal component analysis. *Artificial Neural Networks—ICANN'97, Lecture Notes in Computer Science*, **1327**, 583–588.
- Tenenbaum, J.B., de Silva, V. and Langford, J.C. (2000) A global geometric framework for nonlinear dimensionality reduction. *Science*, **290**, 2319–2323.
- Coifman, R.R., Lafon, S., Lee, A.B., Maggioni, M., Nadler, B., Warner, F. and Zucker, S.W. (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl Acad. Sci. USA*, **102**, 7426–7431.
- Huckemann, S. and Ziezold, H. (2006) Principal component analysis for Riemannian manifolds, with an application to triangular shape spaces. *Adv. Appl. Prob.*, **38**, 299–319.
- Duarte, C.M. and Pyle, A.M. (1998) Stepping through an RNA structure: a novel approach to conformational analysis. *J. Mol. Biol.*, **284**, 1465–1478.
- Wadley, L.M., Keating, K.S., Duarte, C.M. and Pyle, A.M. (2007) Evaluating and learning from RNA pseudotorsional space: quantitative validation of a reduced representation for RNA Structure. *J. Mol. Biol.*, **372**, 942–957.
- Nash, J. (1956) The imbedding problem for Riemannian manifolds. *Ann. Math.*, **63**, 20–63.