J Math Chem (2015) 53:2184–2190
DOI 10.1007/s10910-015-0540-4

ORIGINAL PAPER

Use of circular variance to quantify the deviation of a macromolecule from the spherical shape

Mihaly Mezei[1]

Received: 1 July 2015 / Accepted: 6 August 2015 / Published online: 20 August 2015

Abstract It is shown that the extent of deviation of a molecular shape from spherical can be characterized by comparing the distribution of the circular variances, a measure originally proposed to quantify angular spread, of the vectors from each atom to the rest of the molecule to the circular variance of a collection of atoms filling the unit sphere. Different measures for quantifying the difference between distribution are proposed and compared.

Keywords Sphericity · Circular variance · Moments · Accessible surface

Mathematics Subject Classification 51P05 · 92C05

1 Introduction

Molecular shape is an important property in determining the behavior of molecules [1]. For example, the molecular shape would affect the hydrodynamic properties of proteins. One important shape descriptor is the deviation of the shape from spherical, called sphericity.

Since the sphere is the shape with the smallest surface for a give volume, deviation from sphericity traditionally is measured by the ratio of the surface of the object in question and the surface of the sphere whose volume matches the volume of the object. However, it is hard to apply this definition to a protein (or, in general to a macromolecule) as such molecules are likely to have a number of internal cavities,

✉ Mihaly Mezei
Mihaly.Mezei@mssm.edu

[1] Department of Structural and Chemical Biology, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA

pockets, invaginations. This approach has been applied to proteins by a judicious definition of molecular surface [2]. A recent work [3] introduced a novel idea: measure the sphericity by the work required to transform the surface of the molecule into a sphere.

Circular variance (CV), a measure of directional spread [4], has been shown to be useful for characterizing macromolecular topography [5] and molecular surfaces [6]. The aim of this paper is to show that CV can also be used to characterize molecular sphericity. It is suggested that the distribution of the CV values of the set of vectors drawn from each atom of a macromolecule (typically a protein) to the rest of the atoms be compared to the same distribution of a set of uniformly distributed points filling a sphere; the difference between the two distributions will be a measure of the difference between that molecule's shape and a sphere.

2 Methods

The circular variance C of vectors $\vec{v}_i$ is defined [4] as

$$C = 1 - \left| \sum_{i=1}^{n} \frac{\vec{v}_i}{|\vec{v}_i|} \right| / n; \quad 0 \le C \le 1.0 \tag{1}$$

When all $\vec{v}_i$ vectors are parallel, $C = 0$. It has been proposed in [5] that the CV of vectors drawn from a test point $\vec{R}$ to atoms $\vec{r}_\imath$ of a macromolecule provides a 0–1 scale for the test point being in the middle of the macromolecule or it being way outside:

$$C = 1 - \left| \sum_{i=1}^{n} \frac{\vec{r}_i - \vec{R}}{|\vec{r}_i - \vec{R}|} \right| / n; \quad 0 \le C \le 1.0 \tag{2}$$

In addition, an analog of CV whose properties are very similar to CV, called weighted circular variance (C_{w}), has also been defined [5]:

$$C_w = 1 - \frac{|\sum_i \vec{r}_i - \vec{R}|}{\sum_i |\vec{r}_i - \vec{R}|}; \quad 0 \le C_w \le 1.0 \tag{3}$$

In the present work it is proposed that the distribution of the C or C_{w} values generated for all protein atoms with respect to all other atoms can be used to derive a measure of the sphericity of that molecule. This measure can be obtained via the comparison of various scalar measures of the CV distribution calculated for the atoms of the molecule in question and of a sufficiently fine uniformly distributed set of points within the unit sphere. Several different measures were considered for quantifying the difference between two CV distributions: (a) direct comparison of the distributions: $\int(\Delta\rho)^2$, $\int(\Delta P)^2$, $\int|\Delta\rho|$, $\int|\Delta P|$ where the integrals were approximated by sums over the discrete bins the distributions were calculated (50 bins in the [0, 1] interval) and (b) comparison of the power sums P^m

$$P^m = \sum_{i=1}^{n} \sum_{j=1}^{m} C_i^j \tag{4}$$

or moments M^m

$$M^m = \sum_{i=1}^{n} \sum_{j=1}^{m} (C_i - \bar{C})^j \tag{5}$$

where n is the number of points that contributed to the CV distributions and $\bar{C}$ is the average of C over the set of points. Note, that there is an important difference in the two types of measures: comparison of the explicit distributions gives a discrete measure since the distributions are calculated using finite bins while the measures using the power or moment sums are continuous as a function of the coordinates.

For the reference point set a uniformly distributed set of points were generated in the positive octant of the sphere in 20 layers of uniform thickness and transformed into the other seven octant. All together 1,600,000 points were generated.

These measures were calculated for two different data sets, using both the original definition and the weighted version. The first set consists of the same grid that was used to obtain the reference distributions for the sphere but stretched along one axis by a factor $S (1.0 \leq S \leq 2.0)$. The second dataset is the 533 proteins extracted from the CATH database in [3]. For each protein an approximate measure of the surface area (A) and volume (V) was calculated, allowing the estimate of their sphericity (Sph) as:

$$Sph = \pi^{1/3} (6V)^{2/3} / A. \tag{6}$$

followed by the calculation of the rank order correlation between the resulting sphericities and the measures proposed here. Note, that $Sph \geq 1.0$; $Sph = 1.0$ for a sphere.

The volume of a protein was approximated as the sum of the heavy-atom spheres defined by the respective VdW radii and its surface area was obtained by the sum of the exposed surface areas of the same spheres (calculated by an extension of the program Gepol [7]). As in the earlier publication [6], surface atoms were defined as having at least 3 % exposed *accessible* surface (based on a probe radius of 1.4 Å) and having a CV value (w.r.t. the rest of the atoms) <0.8.

The CV distributions of the proteins and of a sphere, as well as the various difference measures discussed were calculated by the program CVDISTR (written in Fortran-77). It is available from the http://inka.mssm.edu/~mezei/cvdistr.

3 Results and discussion

The probability densities $\rho(C)$, $\rho(C_w)$ and cumulative probability distributions $P(C)$, $P(C_w)$ calculated from the uniformly distributed random points in the unit sphere are shown on Fig. 1. The same distributions were also calculated on a regular lattice of comparable size. However, while the resulting distributions were close to the ones obtained from the random points, they were significantly less smooth and thus the random point distributions were used as the reference.

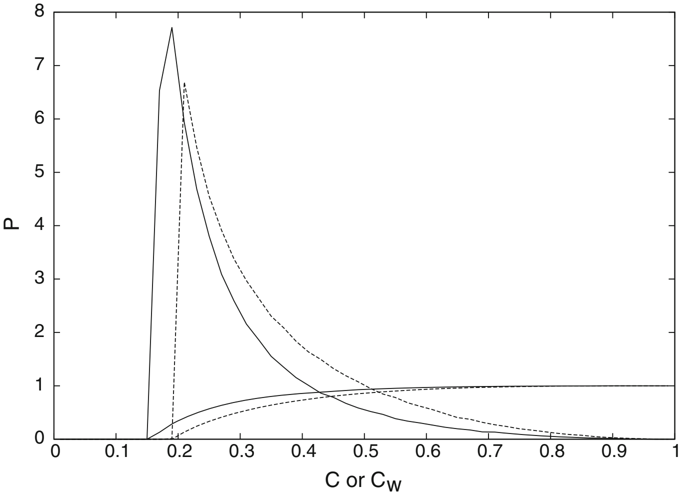

Fig. 1 Probability density ρ(C) (*full line*, curve with peak), cumulative probability distribution $P(C)$ (*full line*, no peak), probability density ρ(C_w) (*broken line*, curve with peak), cumulative probability distribution $P(C_w)$ (*broken line*, no peak) of the circular variances calculated over a 1.6×10^6 uniformly distributed random points within a sphere

The first test of the CV-based sphericity measure involved the same set of sphere-filling point that was used to generate the reference distributions but stretched along one of the axes by an increasingly larger factor. Table 1 gives the absolute and squared differences between the probability densities and between the cumulative probabilities for stretching factors 1.1, 1.2, 1.5 and 2.0, the absolute differences between the CV power sums and moment sums to several different orders, as well as the change (w.r.t. the sphere) in sphericity of the distorted spheres calculated from the surface and volume of the ellipsoids these points fill, along with the Pearson correlation coefficients between the sphere sphericity change and the measures proposed. All measures proposed, as well as the sphericity change, increase as the grid is increasingly distorted. This means that the Spearman (rank order) correlation between the sphericity and the proposed measure is 1.0, supporting the basic hypothesis of this paper. The calculated Pearson correlations are also very high. For the measures based on explicit comparison of the distributions they are between 0.97 and 0.99 and uniformly 0.99 for all power and moment sums shown. The highest correlation among the distribution-based measures was found for $\int |\Delta P|$. No notable difference was observed between measures using C or C_w in this test.

Next, the approximate sphericities of 533 protein domains were calculated and their CV distributions $\rho(C)$, $P(C)$, (C_w) and $P(C_w)$ were calculated. The Spearman (rank order) and Pearson correlations between the sphericities and the proposed measures are shown in Table 2. Among the measures based on the explicit distributions the best performance was (based on both test) found with $\int |\Delta P|$; no systematic difference was found between C and C_w based measures. Measures based on the moment sums converged at $m = 2$ and correlated with the sphericity significantly better using the original C-based measure. On the other hand, the power sum based measures showed

Table 1 Comparison of the different CV-based sphericity measures on progressively distorted spheres

Sphericity measure	CV type	Stretching factor				P-corr
		1.1	1.2	1.5	2.0	
$\int(\Delta\rho)^2$	C	0.0015	0.0031	0.0069	0.0110	0.966
$\int(\Delta P)^2$	C	0.0010	0.0051	0.0476	0.2133	0.977
$\int\lvert\Delta\rho\rvert$	C	0.0743	0.1362	0.3018	0.5028	0.974
$\int\lvert\Delta P\rvert$	C	0.0907	0.2905	0.1930	2.8947	0.990
$\int(\Delta\rho)^2$	C_w	0.0013	0.0046	0.0125	0.0208	0.963
$\int(\Delta P)^2$	C_w	0.0125	0.0099	0.1014	0.4238	0.981
$\int\lvert\Delta\rho\rvert$	C_w	0.0707	0.1607	0.4116	0.6936	0.970
$\int\lvert\Delta P\rvert$	C_w	0.1136	0.3833	1.6291	3.9337	0.990
ΔP^2	C	0.00045	0.00168	0.00854	0.02528	0.990
ΔM^2	C	0.00008	0.00059	0.00141	0.00411	0.990
ΔP^5	C	0.00079	0.00291	0.01474	0.04395	0.989
ΔM^5	C	0.00009	0.00035	0.00177	0.00524	0.990
ΔP^{10}	C	0.00091	0.00334	0.01695	0.05073	0.989
ΔM^{10}	C	0.00009	0.00035	0.00177	0.00526	0.990
ΔP^{20}	C	0.00094	0.00348	0.01768	0.05303	0.989
ΔM^{20}	C	0.00009	0.00035	0.00177	0.00526	0.998
ΔP^2	C_w	0.00035	0.00131	0.00679	0.02030	0.989
ΔM^2	C_w	0.00013	0.00050	0.00258	0.00779	0.990
ΔP^5	C_w	0.00065	0.00242	0.01253	0.03810	0.989
ΔM^5	C_w	0.00015	0.00055	0.00286	0.00866	0.989
ΔP^{10}	C_w	0.00074	0.00276	0.01430	0.04388	0.989
ΔM^{10}	C_w	0.00015	0.00055	0.00289	0.00876	0.989
ΔP^{20}	C_w	0.00076	0.00285	0.01484	0.04571	0.989
ΔM^{20}	C_w	0.00015	0.00055	0.00289	0.00877	0.989
	ΔSph	0.0016	0.0057	0.0266	0.707	

Stretching factor: ratio of the extended axis to the corresponding sphere radius; The integrals expressing sphericity measures are calculated over the [0, 1] interval and involve the absolute or squared differences of P(C), ρ(C), P(C_W), ρ(C_W), resp., ΔP^m and ΔM^{m} are the differences of P^m and M^{m}, calculated by Eqs. (4) and (5), resp.; P-corr: Pearson correlation between ΔSph and the measure; C or C_w in the second column indicates whether the CV was calculated using Eq. (2) or Eq. (3); ΔSph: 1.0-Sphericity calculated from the ellipsoid volume and surface

progressively lower correlation with the spericity as m was increased but still correlated better than the moment sum based measures. Furthermore, for the power sums the C_w-based measures outperformed those that were based on C. The overall best correlation was observed for ΔP^2 using C_w. Note that the conclusions above hold for both types of correlations calculated.

The extent of correlation between the proposed measure and the protein sphericities is similar to the correlation observed in [3] between the measure introduced there

Table 2 Correlation between the calculated sphericity measures and the approximate sphericities calculated from the molecular volumes and surfaces

Sphericity measure	Spearman corr.		Pearson corr.	
	C	C_w	C	C_w
$\int(\Delta\rho)^2$	0.490	0.507	0.529	0.490
$\int(\Delta P)^2$	0.532	0.539	0.575	0.532
$\int\lvert\Delta\rho\rvert$	0.587	0.604	0.606	0.587
$\int\lvert\Delta P\rvert$	0.645	0.635	0.650	0.649
ΔP^2	0.608	0.709	0.608	0.776
ΔM^2	0.661	0.582	0.661	0.598
ΔP^5	0.598	0.674	0.615	0.744
ΔM^5	0.668	0.594	0.687	0.610
ΔP^{10}	0.592	0.660	0.614	0.733
ΔM^{10}	0.669	0.594	0.688	0.610
ΔP^{20}	0.589	0.653	0.611	0.723
ΔM^{20}	0.669	0.594	0.688	0.594

The integrals expressing sphericity measures are calculated over the [0, 1] interval and involve the absolute or squared differences of P(C), ρ(C), P(C_W), ρ(C_W), resp., ΔP^m and ΔM^m are the differences of P^m and M^m, calculated by Eqs. (4) and (5), resp.; C and C_w indicates whether the CV was calculated using Eq. (2) or Eq. (3)

and the sphericity. Some difference from the full correlation can be attributed to the approximate calculation of the sphericities but certainly not all. The rest of the difference is ‘real’, reflecting the fact that quantification of the concept of sphericity of an irregular shape is not a uniquely defined process.

The major advantage of the proposed sphericity measure is that there is no need to define the surface of the molecule and there are no restrictions on the shape of the molecule; it only needs the atomic coordinates. While the complexity of the proposed measure is quadratic in the number of atoms (quite high) there is a natural limit of molecule size where such measure is likely to be of interest, typically for a single protein domain.

The measures based on the CV power sums or moment sums are, unlike the measures based on explicit calculation of distributions, continuous as a function of the atomic coordinates. This means that only members of this class of measures are amenable to be used for restraining the molecular shape to a certain level of sphericity during simulations.

Acknowledgments Computational resources and staff expertise was provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai. Prof. Patrice Koehl is thanked for providing the CATH533 database of protein domain structures used in testing.

References

1. P.G. Mezey, *Shape in Chemistry: An Introduction to Molecular Shape and Topology* (VCH Publishers, New York, 1993)
2. D.S. Kim, J.K. Kim, C.I. Won, C.M. Kim, J.Y. Park, J. Bhak, Sphericity of a protein via the β – complex. J. Mol. Graph. Model. **28**, 636–649 (2010)

3. J. Hass, P. Koehl, How round is a protein? Exploring protein structured for globularity using conformal mapping. Front. Mol. Biosci. **1**, 1–11 (2014)
4. K.V. Mardia, P.E. Jupp, *Directional Statistics* (Wiley, Chichester, 2000)
5. M. Mezei, A new method for mapping macromolecular topography. J. Mol. Graph. Model. **21**, 463–472 (2003)
6. M. Mezei, Statistical properties of protein–protein interfaces. Algorithms **8**, 92–99 (2015)
7. E. Silla, F. Villar, O. Nilsson, J.L. Pascual-Ahuir, O. Tapia, Molecular volumes and surfaces of biomacromolecules via GEPOL: a fast and efficient algorithm. J. Mol. Graph. **8**, 168–172 (1990)