
Article:
Abushilah, SF, Taylor, CC orcid.org/0000-0003-0181-1094 and Gusnanto, A (2020) Geometry-based Distance for Clustering Amino Acids. Journal of Applied Statistics, 47 (7). pp. 1235-1250. ISSN 0266-4763

https://doi.org/10.1080/02664763.2019.1673324

ARTICLE TEMPLATE

Geometry-based Distance for Clustering Amino Acids

Samira F. Abushilah[a,b], Charles C. Taylor[a], and Arief Gusnanto[a]

[a]Department of Statistics, University of Leeds, Leeds LS2 9JT, UK; [b]Department of Mathematics, Faculty of Education for Girls, University of Kufa, Najaf, Iraq

ARTICLE HISTORY
Compiled October 11, 2019

ABSTRACT
Clustering amino acids is one of the most challenging problems in functional and structural prediction of protein. Previous studies have proposed clusters based on measurements of physical and biochemical characteristics of the amino acids such as volume, area, hydrophilicity, polarity, hydrogen bonding, shape, and charge. These characteristics, although important, are less directly related to the protein structure compared to geometrical characteristics such as dihedral angles between amino acids. We propose using the p-value from a test of equality of dihedral-angle distributions as the basis of a distance measure for the clustering. In this novel approach, an energy test is modified to deal with bivariate angular data and the p-value is obtained via a permutation method. The results indicate that the clusters of amino acids have sensible interpretation where Glycine, Proline, and Asparagine each forms a distinct cluster. A simulation study suggests that this approach has good working characteristics to cluster amino acids.

KEYWORDS
Circular distance; Squared Euclidean distance; Permutation two-sample test; Energy statistic; Hierarchical clustering; Similarity indices

1. Introduction

Clustering amino acids is a challenging problem in protein bioinformatics. This problem is important because, within a bigger context, it is strongly related to the problem of protein structure prediction. Since a protein's function is determined by its structure, researchers are interested to understand e.g. whether a substitution of one amino acid with another has a substantial impact on its structure, and hence its function. Since the identification of protein structure is expensive and very time consuming [21], amino acid clustering becomes crucial as it enables us to better understand how they are related to protein structure.

In recent years, different procedures have been used to cluster amino acids. Georgiu *et al.*[8] published a study of amino acids classification via a fuzzy clustering technique. They employ two different distance measures: the Minkowski distance and the fuzzy distance metric and they rely on several physical properties of the amino acids in this classification. Kosiol *et al.*[12] developed a criterion and grouping method to classify amino acids. This criterion depends on the description of protein evolution by a Markov process. Albatineh and Razeghifard[2] presented a clustering approach to clas-

CONTACT A. Gusnanto. Email: a.gusnanto@leeds.ac.uk

sify amino acids depending on physicochemical properties of amino acids such as volume, area, hydrophobicity, polarity, hydrogen bonding, shape and charge. Stanfel[22] produced a new approach to clustering amino acids depending on their physicochemical properties.

Regardless of the clustering methods employed, these methods rely on dissimilarity measures between amino acids that are informed by their physical, biological, biochemical, or physicochemical properties. Some of the physicochemical properties, for example, are not related to the 3-dimensional structure of the protein, while the others are only related indirectly. These are volume, area, polarity, charge, and shape [22]. The term 'shape' here refers to the shape characteristics of each amino acid based on the atom configuration, and does not refer to protein structure that are shaped by the amino acids. When protein structure and function is the main interest, we regard this as sub-optimal. It is therefore intuitive and appropriate to consider dissimilarity measures that are informed directly by geometrical properties of amino acids in a protein structure.

The structure of a protein can be described as a collection of dihedral angles along the backbone of protein that connects the amino acids [5]. The backbone of a protein, with m amino acids, consists of a sequence of atoms $\mathrm{N}_1 - \mathrm{C}_1^\alpha - \mathrm{C}_1 - \mathrm{N}_2 - \mathrm{C}_2^\alpha - \mathrm{C}_2 - \ldots - \mathrm{N}_m - \mathrm{C}_m^\alpha - \mathrm{C}_m$, where N and C denote nitrogen and carbon atoms, respectively. The difference between C^α and C atoms are simply in the types of other atoms that are bonded with them. The angle around the bond $\mathrm{N}_i - \mathrm{C}_i^\alpha$, denoted ϕ_i, and the angle around the bond $\mathrm{C}_i^\alpha - \mathrm{C}_i$, denoted ψ_i, are the i-th pair of dihedral angles of the protein, for $i = 1, \ldots, m-1$ [4].

In this study, we propose a novel approach where the dissimilarity measures between amino acids are informed directly by their distribution of dihedral angles. When a pair of amino acids are similar, i.e. in the same cluster, then we expect the dihedral angles of the first amino acid across proteins will have the same distribution as those in the second amino acid. Similarly, when the two amino acids are not similar, i.e. in different clusters, then we expect the distribution of dihedral angles to be different. It is therefore natural to use p-value of a test of equality of the two distributions as the basis of a measure of dissimilarity.

To test the equality of the distribution of dihedral angles, we construct a test where we modify the energy distance [25] as a test statistic to deal with the bivariate nature of dihedral angles and its significance is assessed using a permutation method. Once the p-value for each pair of amino acids is obtained and transformed to a dissimilarity measure, *any* distance matrix-based clustering method can be employed. In our study we consider commonly used clustering methods as illustrations on the use of the new dissimilarity measures: hierarchical clustering with single, complete, and average linkage, and Ward's method to illustrate our application. We also consider a simulation study to confirm the proposed method's characteristics.

This paper is organised as follows. A motivating dataset is described in Section 2. Section 3 describes the main methods and a simulation study. The results of simulation study and an application to real dataset are illustrated in Section 4. Section 5 will contain discussion of our new approach.

2. Motivating datasets

We consider the dataset from the Kinemage database,
http://kinemage.biochem.duke.edu/databases/top500.php, where a selection of 500 pro-

teins from the Protein Data Bank (PDB) are catalogued. The 500 proteins were selected because of their high quality, low homology and high resolution (1.8 Å or better) [13]. The dataset contains the information on dihedral angles ϕ and ψ, protein ID, position of amino acids in the protein, and type of amino acid in this position. It is important to note that each amino acid in each position in each protein is associated with a pair of dihedral angles.

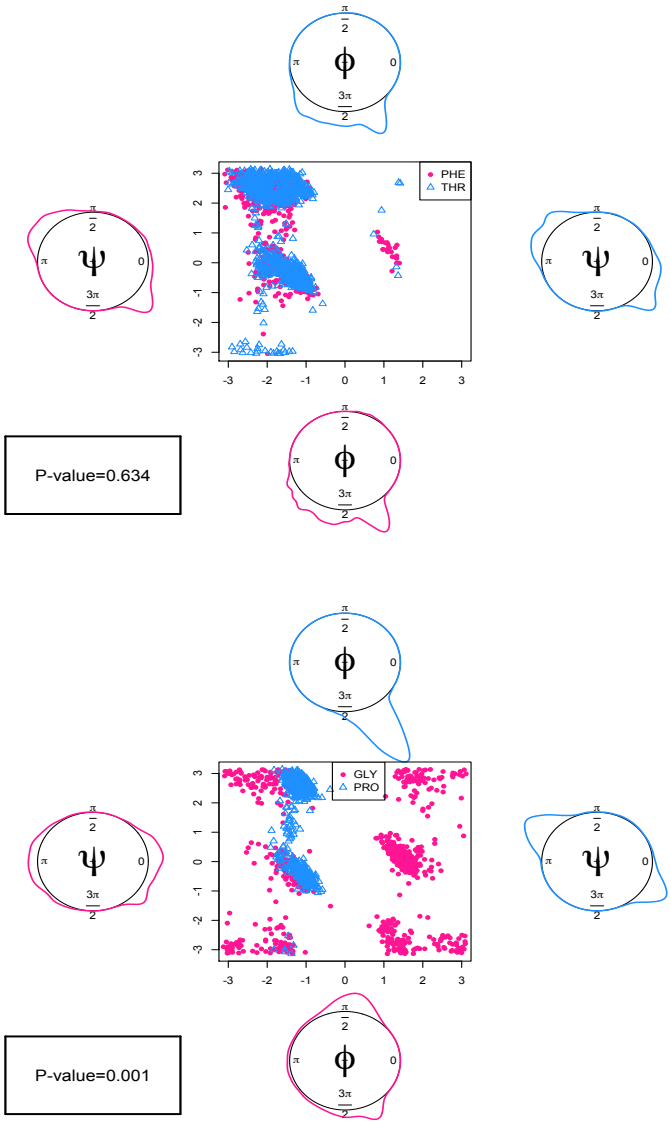

Figure 1. Distribution of dihedral angles ϕ (horizontal axis) and ψ (vertical axis) of amino acids PHE (red dot) and THR (blue triangle) in the top panel and GLY (red dot) and PRO (blue triangle) in the bottom panel. Please note that opposite edges in the figures are the same points because they are circularly wrapped. Circular plots illustrate the marginal circular density for the dihedral angles. The p-values are from test of equality of distribution of dihedral angles as described in Section 3.

Some of the data are presented in Figure 1 for illustration. The figure shows the distribution of dihedral angles between a pair of amino acids in the data. In the top panel of the figure, we compare the distribution of the dihedral angles betweed PHE

and THR that we later find to be similar (in terms of distribution) from the test proposed in Section 3. The bottom panel of the figure shows the comparison between GLY and PRO that we later find to be different. These examples provide a quite straightforward conclusion on the equivalence of the distribution of dihedral angles between two amino acids. In other pairs of amino acids, it is not straightforward to identify that. Therefore, we need the help of a statistical test as described in the next section. Further description of this dataset is available in the Supplementary Material.

In addition to the above dataset, we consider a second independent dataset, also from the Protein Data Bank (PDB). We select *all* proteins from PDB that have less than 1.5Å in resolution, and published (uploaded to the repository) from 2015 to May 2019 (inclusive). In the end, we have 1,259 proteins in this dataset, and none of them are overlapping with the first dataset above. The description of this second dataset is presented in more details in the Supplementary Material. Some more comparison results with this new dataset are also presented, mainly in the Supplementary Material.

3. Methodology

In this section, we present main method to obtain the dissimilarity measures between pairs of amino acid as the p-value of equality test of distribution of dihedral angles. We also present how this information is utilised to cluster amino acids. A simulation study is also considered to see the working characteristics of the proposed method.

3.1. Notation and setting

Before we discuss the test, it is important that we set out the notation and context that we use in this manuscript. With regard to the motivating dataset above, let $x_1, x_2, \ldots, x_{n_1}$ be n_1 pairs of dihedral angles from the first amino acid across proteins and positions (in proteins). Similarly, let $y_1, y_2, \ldots, y_{n_2}$ be n_2 pairs of dihedral angles from the second amino acid. We assume that the x_i's and y_j's are samples or realised values of random variables X and Y with cummulative distribution function F_X and F_Y, respectively. Given this, our setting is as follows.

(1) Each x_i contains two angular measurements ϕ_{xi} and ψ_{xi} for $i = 1, 2, \ldots, n_1$. Similarly, each y_i contains two angular measurements ϕ_{yj} and ψ_{yj} for $j = 1, 2, \ldots, n_2$. In other words, each x_i and y_j is bivariate.

(2) We assume that x_i and $x_{i'}$ are independent for all $i \neq i'$. Similarly, we also assume that y_j and $y_{j'}$ are independent for all $j \neq j'$. When we are considering a pair of amino acids to be tested for equality in their distribution of dihedral angles, it is possible that a particular x_i of the first amino acid and a particular y_j of the second amino acid are adjacent in a protein, i.e. they are dependent. However, the dependencies are relatively weak; the correlation between adjacent amino acids is approximately between -0.1 and 0.1. Relative to the size of the data, where this occurence happens in less than 18 adjacent amino acids among approximately 1,600 to 9,000 pairs of amino acids, this dependence is too weak to have practical importance. Some more details are presented in the Supplementary Material on this note. Furthermore, the significance of the test will be calculated through permutation that can deal with this weak dependency[9], hence it is not a problem.

(3) The test of equality of the distribution of dihedral angles between two amino

acids is a test on two independent samples with n_1 and n_2 bivariate observations. These two samples are not paired, i.e. x_1 is not paired with y_1, although within each x_i and y_j, the angles ϕ_{xi} is paired with ψ_{xi} and ϕ_{yj} is paired with ψ_{yj}.

(4) The context of this investigation is in identifying clusters of 20 amino acids, each of which has a distribution of dihedral angles. The context is not in clustering dihedral angles themselves. Each amino acid has a bivariate distribution of dihedral angles and our interest lies in developing a dissimilarity measure between amino acids so that *any* distance matrix-based clustering method can be applied. Therefore, the focus of investigation is to cluster *groups* of observations that are represented by a single p-value, rather than the (individual) observations themselves. With this in mind, the dissimilarity matrix will be between amino acids and of size 20×20, although one amino acid in the above data can have more than 5,000 bivariate dihedral angles. As such, the dissimilarity between two different amino acids is represented by a single number that summarises how similar the distribution of dihedral angles in both amino acids. In this case, a p-value of less than 0.05 indicates that the distribution of of dihedral angles between the two amino acids are different at the 5% significance level.

We now discuss the test of equality of distribution of dihedral angles between two samples (two amino acids) as presented in the next section.

3.2. Bivariate angular permutation test (BAPT)

As described above, we aim to obtain a p-value from a test of equality distribution of dihedral angles between two amino acids as a dissimilarity measure. We are therefore interested in testing the null hypothesis

$$H_0 : F_X = F_Y$$

against

$$H_1 : F_X \neq F_Y,$$

where F_X is bivariate distribution of dihedral angles from one amino acid and F_Y is from the second amino acid.

It is intuitive to consider the distance between two probability distribution as the Cramer distance [6]

$$\int_0^{2\pi} \{F_X(z) - F_Y(z)\}^2 \, dz, \tag{1}$$

which is equal to zero if $F_X = F_Y$. In the context of Euclidean space, Szekely[23] showed that this distance is half of the *energy distance*

$$D^2(F_X, F_Y) = 2E||X - Y|| - E||X - X'|| - E||Y - Y'||, \tag{2}$$

where the cummulative distribution function of X' is F_X and that of Y' is F_Y, E is the expected value, and $||\cdot||$ is the norm of vector.

In metric spaces, the above energy distance is defined as *energy statistic* (e-statistic), which is a function of distances between statistical observations [19]. It can be applied

to measure the difference between distributions of two or more samples with arbitrary dimension not necessarily equal [24]. Since our context is bivariate distribution (of dihedral angles), we consider the energy statistic as defined by Szekely and Rizzo[25] in testing the equivalence of distribution in higher dimension:

$$\mathcal{E}(X,Y) = \frac{n_1 n_2}{n_1+n_2}\Big(\frac{2}{n_1 n_2}B_1 - \frac{1}{n_1^2}B_2 - \frac{1}{n_2^2}B_3\Big) \tag{3}$$

where

$$B_1 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} d(x_i, y_j), \qquad B_2 = \sum_{i=1}^{n_1}\sum_{i'=1}^{n_1} d(x_i, x_{i'}), \qquad B_3 = \sum_{j=1}^{n_2}\sum_{j'=1}^{n_2} d(y_j, y_{j'})$$

and, in the Euclidean space, $d(x_i, y_j)$ is usually defined as

$$d(x_i, y_j) = ||x_i - y_j||. \tag{4}$$

From the above formulation, it is immediately clear that if the two datasets are exactly identical (i.e. $x_i = y_j$ for all i, j), then $d(\cdot)$ will be zero for all ϕ's and ψ's, and so will the energy statistic $\mathcal{E}(\cdot)$. When there is a discrepancy in the distribution of dihedral angles in the first amino acid from those in the second amino acid, then $\mathcal{E}(\cdot)$ will increase to be greater than zero in the positive real line.

However, this definition of distance (4) is not directly applicable to our case with bivariate angular data. For the application of energy statistic in our context with $x_i \equiv (\phi_{xi}, \psi_{xi})$ and $y_j \equiv (\phi_{yj}, \psi_{yj})$, we define $d(x_i, y_j)$ as

$$d(x_i, y_j) = \Delta(\phi_{xi}, \phi_{yj})^2 + \Delta(\psi_{xi}, \psi_{yj})^2, \tag{5}$$

for $i = 1, \ldots, n_1$, and $j = 1, \ldots, n_2$, where, for two angles θ and ϑ,

$$\Delta(\theta, \vartheta) = 1 - \cos(\theta - \vartheta). \tag{6}$$

It is important to note that the objective of this study to cluster of amino acids will be based on the dissimilarity measure obtained from the p-value of the test statistic $\mathcal{E}$, and not to identify clusters from the distance in Eq. (5). The distance (5) measures the angular distance between dihedral angles from a pair of amino acids.

To have a sense of whether the observed energy statistic $\mathcal{E}$ is "far enough" from that under the null hypothesis, we calculate the p-value of the test as a measure of significance, which will be considered as a measure of dissimilarity between amino acids. Under the null hypothesis, the distribution of the energy statistic $\mathcal{E}$ does not follow a simple form [24], even when we consider Euclidean distance. So, in our context where the distance is from dihedral angles, the distribution is not tractable. To deal with this, we consider the calculation of p-value of the test via permutation method, as suggested by Szekely and Rizzo[24].

With this new definition of distance for bivariate angular data, it is important to note that $\mathcal{E}$ may take a small negative value. Our simulation study indicates that, under the null hypothesis, it is possible to have a small fraction of simulated test statistics to be negative, although their magnitude is very small. The main reason is because the definition of distance between two bivariate angular points in Eq. (5) is calculated as a direct line that 'pierces' through the torus of bivariate angular distribution [3, 14, 16],

instead of the shortest line along the surface of the torus. Because of this, the quantities $B_2/n_1^2 - B_3/n_2^2$ may be slightly larger than $2B_1/(n_1 n_2)$ in Eq. (3). However, this is not a problem in our analysis because, first, we consider p-value as the dissimilarity measure that is non-negative and monotonically decreasing and, second, the p-value is obtained from a permutation method where the distribution of $\mathcal{E}$ is obtained empirically. It requires, however, a careful consideration when interpreting the magnitude of the observed energy statistic from real data.

To obtain the p-value, we consider the following algorithm. This p-value is calculated for each and every pair of amino acids.

(1) Given a pair of amino acids, we calculate the observed test statistic $\mathcal{E}_{\text{obs}}$ from Eq. (3) using the distance according to Eqs. (5) and (6).
(2) We pool the bivariate data $x_1, x_2, \ldots, x_{n_1}$ and $y_1, y_2, \ldots, y_{n_2}$ to have a pooled dataset with n_1+n_2 observations. Under the null hypothesis, we randomly shuffle the group labels and assign them to the pooled data to obtain $x_1^*, x_2^*, \ldots, x_{n_1}^*$ and $y_1^*, y_2^*, \ldots, y_{n_2}^*$. In this step, the dihedral angles within each x_i and y_j are kept intact and fixed.
(3) Calculate the energy statistic for the shuffled data $x_1^*, x_2^*, \ldots, x_{n_1}^*$ and $y_1^*, y_2^*, \ldots, y_{n_2}^*$, denoted $\mathcal{E}^*$, from Eq. (3) using the distance according to Eqs. (5) and (6).
(4) We repeat Steps 2 and 3 n_{perm} times to obtain $\mathcal{E}_1^*, \mathcal{E}_2^*, \ldots, \mathcal{E}_{n_{\text{perm}}}^*$. In our study, we use $n_{\text{perm}} = 1000$, although this can be set at a higher number if one wishes to have higher precision on the estimated p-value. The $\mathcal{E}^*$'s represent the distribution of the energy statistic under the null hypothesis.
(5) The p-value of the test is calculated as

$$\frac{\sum_{h=1}^{n_{\text{perm}}} I\left(\mathcal{E}_h^* \geq \mathcal{E}_{\text{obs}}\right)}{n_{\text{perm}}},$$

where $I(\cdot)$ is a function that is equal to one if the statement in the brackets is true and zero otherwise.

The above algorithm is run for each and every pair of amino acids to obtain the dissimilarity matrix. The end result of this step is a 20×20 dissimilarity matrix of amino acids. After this step, then any distance matrix-based clustering methods can be employed. To illustrate the application of clustering amino acids based on the dissimilarity matrix, we consider hierarchical clustering (single linkage, complete linkage, and average linkage), and the Ward's method. We touch on them briefly in the next section for completeness.

3.3. Clustering methods based on dissimilarity matrix

We consider several well known methods to cluster amino acids based on the dissimilarity matrix obtained using the above algorithm. The purpose of considering them in this paper is only to illustrate the application of clustering amino acids based on the dissimilarity matrix produced by the above method. These methods are hierarchical clustering with single, complete, and average linkage, and Ward's method[15, 18]. For a reference on these clustering methods, the reader may refer to any cluster analysis textbook, for example [7]. In determining the number of clusters, we consider to make a 'cut' at a reasonable height in a dendogram. There are several criteria to do this

(e.g. [15, 18]). However, since the purpose is only to illustrate the use of newly defined dissimilarity matrix for clustering, one may consider different ways of determining the number of clusters.

One important note in the application of clustering is that the p-value from the test on a pair of amino acids is transformed to a distance by taking the negative of the log of p-value. Let M be a 20×20 distance matrix between amino acids. The entries of the matrix, denoted $m_{kk'}$, is defined as

$$m_{kk'} = -\log(p_{kk'}) \tag{7}$$

for $k = 1, \ldots, 20$ and $k' = 1, \ldots, 20$ where $p_{kk'}$ is the p-value of the test between k-th and k'-th amino acids. For $k = k'$, we set $p_{kk'} = 1$.

After clustering, we will consider the Rand index to see whether there is agreement between two clustering methods [17]. There are other indices to identify how good the agreement between two different clustering methods but we consider the Rand index for simplicity and easy interpretation on the agreement. A value of zero in Rand index indicates that the two clustering methods do not agree while a value of one indicates that both methods produce exactly the same clusters.

3.4. Simulation study

We perform a simulation study to understand the working characteristics of the test of equality of distribution of dihedral angles between two amino acids, which characterises whether the two amino acids shall be in the same cluster or not. The purpose of this simulation study is two-fold. First, we are interested to understand whether the proposed test has an appropriate control of false positive (Type-I error). For this purpose, we generate two dihedral-angle distributions under the null hypothesis that they are equal to identify the control of Type-I error rate. This simulation is critical to identify that when we say that the significance of the test is 5%, then the test actually controls the probability of false positive at 5%. In other words, we wish to confirm that the estimated p-value is accurate, as it is going to be the basis for a measure of dissimilarity between amino acids.

Second, we are interested to identify that the test is able to distinguish two amino acids when they are truly from different distributions, and at what amount of differences that the test is able to distinguish the two. We perform a simulation study under the alternative hypothesis, where there is a difference in the distribution between two amino acids either in terms of mean or concentration. Given that there are six parameters from bivariate distribution that can vary in this simulation (two mean parameters and four concentration parameters), we consider to only vary one mean parameter and two concentration parameters that are constrained to be equal for simplicity. Both purposes of the simulation can be constructed in a single simulation framework as described below.

In the simulation study, we perform the following:

(1) For the first amino acid, we generate n_1 bivariate data points under the bivariate von Mises distribution with mean vector $(0, 0)^T$ and concentration parameters

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The number of simulated data points n_1 is set at $n_1 = 100, 200, 500$ and 1000, to understand the effect of different number of observations per sample in the simulation.

(2) In the case where we vary the mean, we generate n_2 =1000 bivariate data points under the bivariate von Mises distribution for the second amino acid with mean vector $(0, \delta)^T$ and concentration parameters

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The values of δ that we consider range from 0 to 2.5 or $\delta \in \{0, 0.25, 0.5, 1.0, 1.5, 2.0, 2.5\}$.

(3) In the case where we vary the concentration, we generate 1000 bivariate data points under the bivariate von Mises distribution for the second amino acid with mean vector $(0, 0)^T$ and concentration parameters

$$\begin{bmatrix} \tau^2 & 0 \\ 0 & \tau^2 \end{bmatrix}.$$

The values of τ^2 that we consider range from 1.25 to 4 or $\tau^2 \in \{1.25, 1.5, 2.0, 3.0, 4.0\}$.

(4) Test the null hypothesis of equal distribution between the two amino acids and record the p-value of the test.

(5) We repeat Steps 1-2-4 1000 times and Steps 1-3-4 200 times for each of δ and each τ^2 so that we have 1000 p-values under each setting.

In the Steps 1-2-4, there is a setting where we can investigate the control of Type-I error, which is when $\delta = 0$. In this setting, we are generating data from the null hypothesis and we can investigate whther the distribution of p-value follow a uniform distribution. For the other settings, whether the mean or concentration that varies, we can investigate the sensitivity of the test as the mean (or variance) starts to differ from the null.

4. Results

4.1. Simulated data

The results of simulation study are presented in Figure 2. The figure indicates that the proposed BPAT test has a proper control of type-I error rate at 0.05 for the nominal 0.05 significance level. This is confirmed further by looking into the quantile-quantile plot of the p-values under H_0 in Figure 3. More details are presented in the Supplementary Material. Figure 2 (left panel) indicates that, as the mean difference increases (and concentration remains the same), the test is able to start identifying that the two distributions of dihedral angles are different. On the right panel of the figure, the test is also able to start identifying the difference as the ratio of concentration between the two distributions increases.

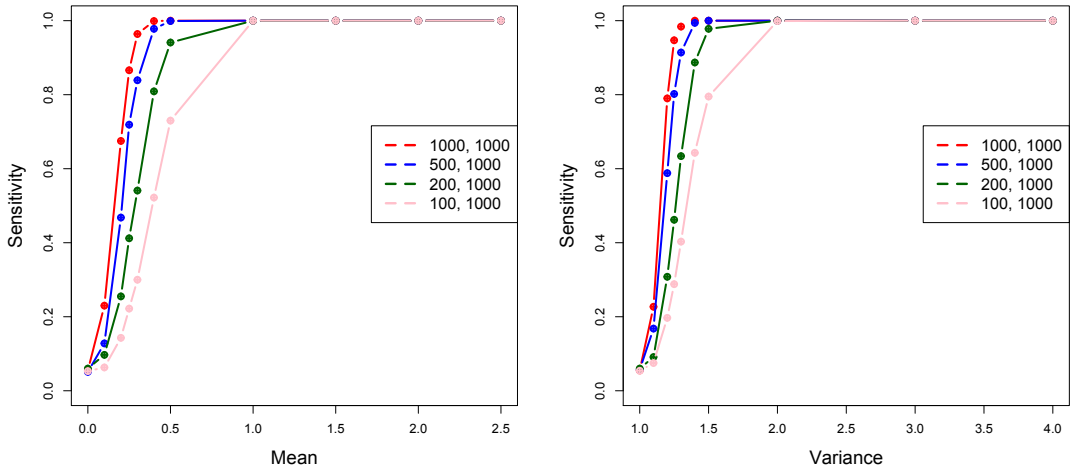

Figure 2. Type-I error rate and sensitivity of the energy test between two simulated distributions of dihedral angles as a function of mean difference (left panel) and concentration ratio (right panel) for different number of observations in the first group: 100, 200, 500, 1000. The number of observations in the second group remains the same (1000). The value for mean difference zero or concentration ratio one corresponds to type-I error rate, while the other values correspond to sensitivity.

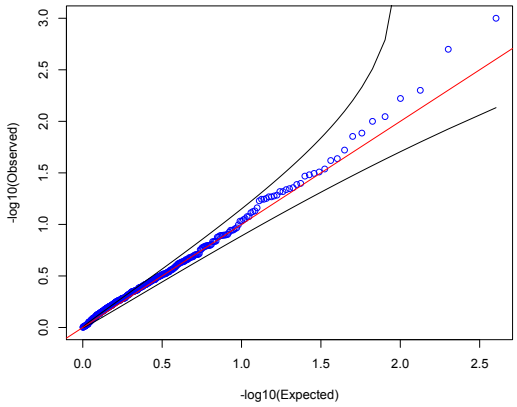

Figure 3. Quantile-quantile plot of -log(p-value) under the null hypothesis of equal distribution between two datasets in the simulation study.

4.2. Real data: test results

The results of testing equality of distribution of dihedral angles between pairs of amino acids are presented in Figure 4. The figure shows pairs of amino acids with p-value of the test more than 0.05 connected by solid black lines. This indicates that those pairs of amino acids have similar distribution of dihedral angles. The figure also shows pairs of amino acids with p-value of the test less than 0.05 not connected by the solid black lines. This indicates that those pairs of amino acids have somehow different distribution of dihedral angles. The figure suggests that GLY, PRO, and ASN have different distributions of dihedral angles to any of the other amino acids.

Figure 4 suggests a natural clustering of amino acids, even before any distance-

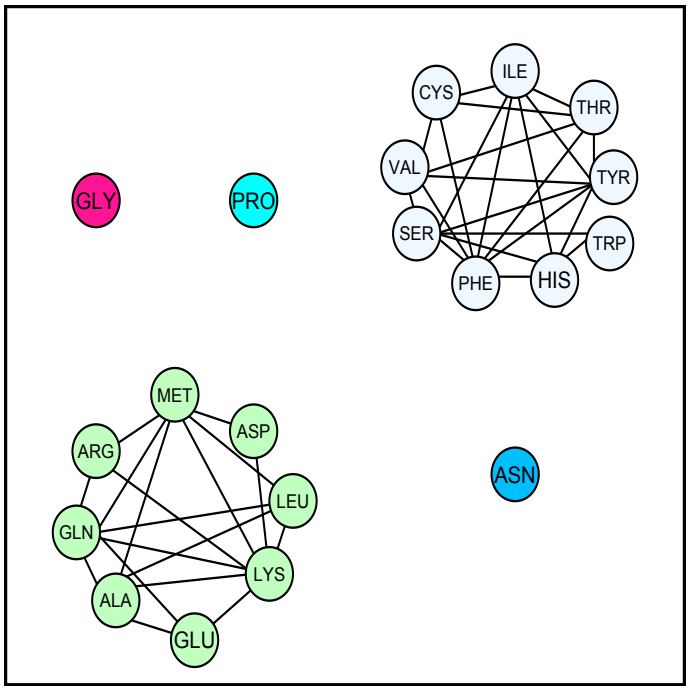

Figure 4. Results of test of equality of distribution of dihedral angles between pairs of protein. The lines between pairs of amino acids denote that the p-value of the test between the pairs is more than 0.05 indicating similarities in the distribution, while the absence of lines incidate that the p-value is less than 0.05 indicating dissimilarity.

based clustering methods is employed. Let us for the moment consider the (natural) clustering of amino acids as suggested in Figure 4. For the two big amino-acid clusters in the figure, the details of p-values of the test between the pairs are presented in Table 1.

4.3. Clustering of amino acids

We now illustrate the application of some methods to cluster amino acids based on the dissimilarity matrix M that we obtained by transforming p-values into dissimilarity measures. Figure 5 shows the results of clustering based on M. With a natural clustering formed based on the p-values shown in Figure 4, then we expect that most, if not all, of distance-matrix based clustering methods will produce similar results. For example, the dendogram of hierarchical clustering under single linkage is quite similar to that under average linkage.

When we make a cut on the dendogram to form five clusters, then the clusters formed have the same members of amino acids as suggested by Figure 4. This is clearer to infer when we consider the single linkage in Figure 5 rather than the average linkage. In the average linkage, we may make a cut in the dendogram to make five or six clusters. When it is six clusters, then HIS and TRP form a different cluster separating from their previous main cluster as shown in Table 2.

Table 1. p-values for each pair of amino acids in two major cluster in Figure 4 as a result of test of equality of distribution of dihedral angles.

	Cluster 1								
	ILE	CYS	THR	TYR	TRP	HIS	PHE	SER	VAL
ILE	—	1.000	1.000	1.000	0.184	0.002	1.000	1.000	0.211
CYS	1.000	—	0.104	0.052	0.001	0.001	0.061	0.165	1.000
THR	1.000	0.104	—	0.274	0.001	0.001	0.634	0.001	1.000
TYR	1.000	0.052	0.274	—	0.003	0.092	0.171	1.000	0.220
TRP	0.184	0.001	0.001	0.003	—	0.070	0.014	1.000	0.001
HIS	0.002	0.001	0.001	0.092	0.070	—	0.026	0.469	0.001
PHE	1.000	0.061	0.634	0.171	0.014	0.026	—	1.000	0.024
SER	1.000	0.165	0.001	1.000	1.000	0.469	1.000	—	1.000
VAL	0.211	1.000	1.000	0.220	0.001	0.001	0.024	1.000	—

	Cluster 2							
	MET	ARG	GLN	ALA	LEU	LYS	GLU	ASP
MET	—	0.285	0.247	0.010	0.448	0.036	0.003	0.025
ARG	0.285	—	0.051	0.005	0.106	0.970	0.009	0.021
GLN	0.247	0.051	—	0.575	0.462	0.041	0.311	0.009
ALA	0.010	0.005	0.575	—	1.000	0.063	0.937	0.001
LEU	0.448	0.106	0.462	1.000	—	0.057	0.003	0.215
LYS	0.036	0.970	0.041	0.063	0.057	—	0.050	0.136
GLU	0.003	0.009	0.311	0.937	0.003	0.050	—	0.001
ASP	0.025	0.021	0.009	0.001	0.215	0.136	0.001	—

Table 2. Cluster memberships from hierarchical clustering with average linkage with five and six clusters.

Five Clusters	Six Clusters
{ILE,PHE,TYR,THR,HIS,SER,VAL,TRP,CYS}	{ILE,PHE,TYR,THR,SER,VAL,CYS}
{MET,LEU,ARG,ASP,LYS,GLN,GLU,ALA}	{MET,LEU,ARG,ASP,LYS,GLN,GLU,ALA}
{ASN}, {GLY},{PRO}	{ASN},{GLY},{PRO},{HIS,TRP}

4.4. Comparisons

Finally, we are interested to the see agreement between our clustering based on geometrical information of the amino acids and clustering based on the physicochemical and biochemical information that are well known (e.g. [2, 22]). The agreement between clustering based on the geometrical information and those other information are presented in Table 3. The table indicates that the agreement between clustering based on the geometrical information (BAPT) has the lowest agreement with the other clustering methods although the magnitude of agreement is moderate. This indicates that the geometry-based distance has some common and different information than those of physicochemical and biochemical information.

This indication is best seen when we consider clustering results based on a composite distance matrix that is calculated from weighting distance matrix based on geometric information and physicochemical information. The gradation of characteristics of clustering can clearly be seen when we vary the weight and it can be seen that some common features are shared between the two distance matrices and some other features are different (the results are presented in the Supplementary Material).

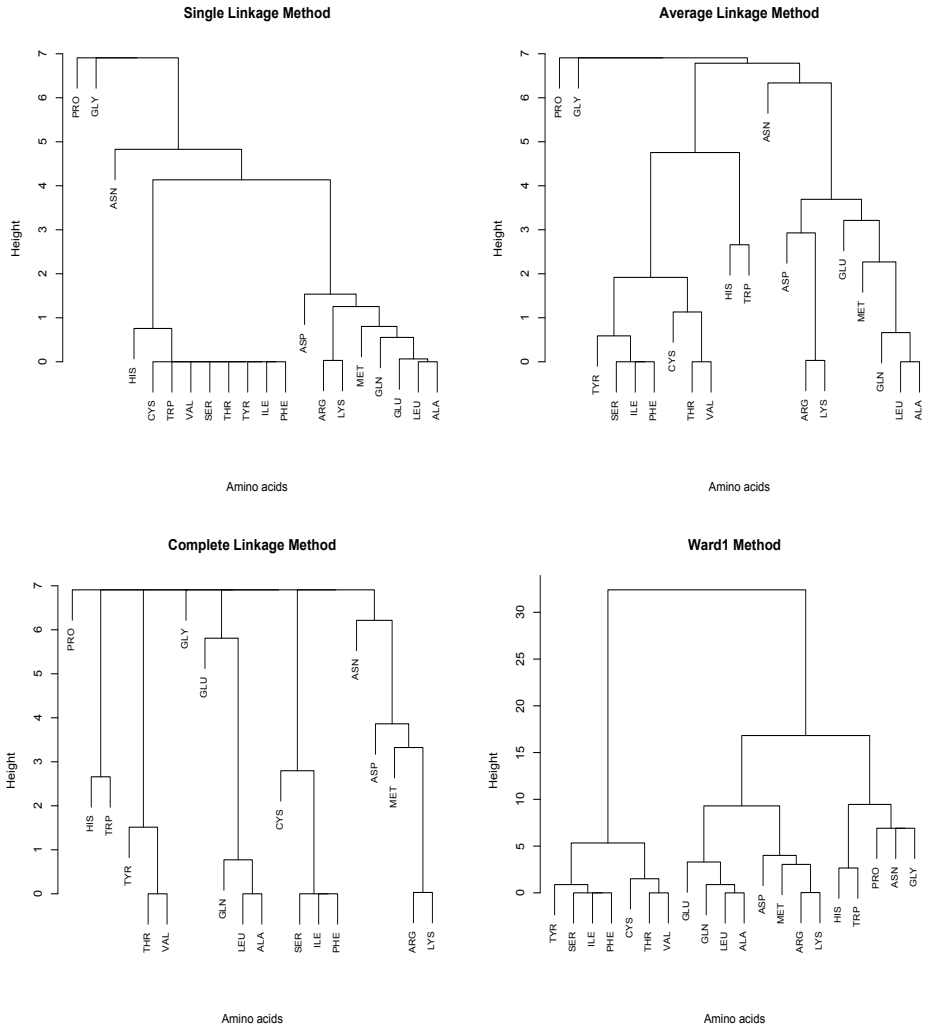

Figure 5. Dendogram of hierarchical clustering of amino acids under single, average, and complete linkage, and Ward's clustering based on the dissimilarity matrix M as defined in Equation (7) from p-values of test of equality of dihedral angle distributions between pairs of amino acids.

In summary, there are three clusters of amino acids that are relatively consistent between the two methods of distance matrix construction: {ASP, GLU, ASN, GLN}, {PHE, TRP, TYR}, and {HIS, LYS, ARG}, while the other ones are different.

5. Discussion

Clustering amino acids is a challenging problem in protein analysis. Extensive research has been done to cluster amino acids based on their physicochemical and biochemical properties. In this investigation we propose a geometry-based distance measure between amino acids for the clustering. This measure is more directly related to the protein structure, which is a critical determinant in protein function. Therefore, we expect that this new approach to clustering has a more direct relevance to the study of protein structure and function.

Table 3. The rate of agreement (Rand index) between clustering using geometry-based distance (BAPT) with clusters based on physicochemical and biochemical measures. Chemical[1]: aliphatic, hydroxyl, cyclic, aromatic, basic, or acidic. Chemical[2]: aliphatic, aromatic, sulfuric, hydroxyl, basic, acidic, and amide. Polarity[1]: non-polar, basic polar, polar, and acidic. Polarity[2]: polar and non-polar. Polarity[3]: non-polar, flexible, and polar. Hydrogen donor: donor, acceptor, donor & acceptor, and none. Charge: positive charged, negative charged, and uncharged.

	Chem.[1]	Chem.[2]	Polar.[1]	Polar.[2]	Polar.[3]	Hyd.	Charge
BAPT	0.642	0.616	0.542	0.495	0.532	0.553	0.469
Chem.[1]	1.000	0.932	0.732	0.611	0.616	0.721	0.532
Chem.[2]		1.000	0.800	0.637	0.611	0.789	0.558
Polar.[1]			1.000	0.837	0.695	0.905	0.737
Polar.[2]				1.000	0.858	0.795	0.574
Polar.[3]					1.000	0.674	0.484
Hyd.						1.000	0.642
Charge							1.000

For this purpose, we modify the energy statistic $\mathcal{E}$ to deal with bivariate angular data to test the null hypothesis that the distribution of dihedral angles between a pair of amino acids is equal. To assess the significance of the observed statistic, we calculate p-value of the test using a permutation method. One main challenge using the permulation method is in the burden of computation. However, we find that this is not a major hurdle as the permutation can be implemented in a home computer relatively easily. It is currently our active research to identify the null distribution of the energy statistics so that we can avoid the permutation and make the test more efficient.

Our simulation study indicates that the Type-I error control is appropriate for the proposed method. This indicates that the permutation test deals with potential dependencies between observations in the two samples. The simulation study also indicates that the test is able to distinguish two amino acids with different distributions, i.e. good sensitivity. Finally, we conclude that, although there are some information in the dissimilarity matrix that are shared with amino acids' physicochemical and biochemical properties, this new approach has new information. Clustering based on the proposed dissimilarity matrix indicates that each of GLY, PRO, and ASN is in its own individual cluster, and the remaining amino acids are grouped into two major clusters. The first two amino acids (GLY and PRO) are well known to not follow typical distribution in terms of their dihedral angles. Due to a ring formation connected to beta carbon, the dihedral angles around the peptide bond have less permissible degrees of rotation. Further research is needed to understand why ASN also has atypical distribution of dihedral angles. This is beyond the scope of our current study, since physical and biochemical properties of the amino acids need to be studied in more details.

Further extensions to this study are also possible. The context of our current study is in clustering *groups* of observations, rather than the observations themselves. It is currently our active research to investigate the clustering of dihedral angles (as observations) directly in each protein. This contributes to address an important research question: given *a* sequence of amino acids (protein), can we identify secondary structures of the protein (that contribute to the overall three-dimensional structure of protein)? In this context, recent advances in clustering can be considered to potentially be either extended or adapted to deal with bivariate angular data. For example,

[11] proposed a fuzzy c-means algorithm for clustering circular data by modifying the fuzzy c-means algorithm to be applicable to directional data. This is a simple and the most common clustering algorithm in all fuzzy clustering methods. Abraham *et al.* [1] presented a Bayesian model to cluster non-ordered multivariate circular data. They introduce a hierarchical model that combines a symmetrisation technique, projected normal distributions and a Dirichlet process. Huang *et al.* [10] proposed a novel clustering method termed multi-view space clustering (MVIC). The aim of the multi-view clustering is to combine information from multiple views in an unsupervised manner to discover a better clustering structure. For clustering circular-linear data, Roy *et al.*[20] proposed a mixture model-based clustering algorithm and they applied this algorithm for clustering hue and chroma information. It still remains to be seen whether the extension or adaptation would give a good clustering performance for bivariate dihedral angles.

6. Conclusion

This study proposes novel geometry-based distance, on which any distance-based clustering method can be utilised to cluster amino acids. This approach paves an alternative way of investigating clusters of amino acids based on a measure that are directly related to protein structure/function, in addition to those based on physicochemical and biochemical properties. The two-sample test involved has a good sensitivity and a proper control of type-I error.

Acknowledgements

SFA is supported by the Government of Iraq.

References

[1] C. Abraham, R. Servien, and N. Molinari, *A clustering bayesian approach for multivariate non-ordered circular data*, Statistical Modelling (2018), p. 1471082X18790420.
[2] A. Albatineh and R. Razeghifard, *Clustering Amino Acids Using Maximum Clusters Similarity*, in *International Conference on Bioinformatics, Computational Biology, Genomics and Chemoinformatics*. 2008, pp. 87–92.
[3] W. Boomsma, K.V. Mardia, C.C. Taylor, J. Ferkinghoff-Borg, A. Krogh, and T. Hamelryck, *A generative, probabilistic model of local protein structure*, Proceedings of the National Academy of Sciences 105 (2008), pp. 8932–8937. Available at http://www.pnas.org/content/105/26/8932.
[4] C.I. Branden and j. Tooze, *Introduction to Protein Structure*, Garland Science, 1999.
[5] M.M. Cox and D.L. Nelson, *Lehninger Principles of Biochemistry*, WH Freeman, 2008.
[6] H. Cramer, *On the composition of elementary errors*, Skandinavisk Aktuarietidskrift 11 (1928), pp. 141–180.
[7] B.S. Everitt, S. Landau, and M. Leese, *Cluster Analysis, 4th Edition*, Oxford University Press, 2001.
[8] D. Georgiou, T.E. Karakasidis, J. Nieto, and A. Torres, *Use of fuzzy clustering technique and matrices to classify amino acids and its impact to chou's pseudo amino acid composition*, Journal of Theoretical Biology 257 (2009), pp. 17–26.
[9] P. Good, *Permutation, Parametric, and Bootstrap Tests of Hypothe-*

ses, Springer Series in Statistics, Springer New York, 2006, Available at https://books.google.co.uk/books?id=tQtedCBEgeAC.

[10] L. Huang, H.Y. Chao, and C.D. Wang, *Multi-view intact space clustering*, Pattern Recognition 86 (2019), pp. 344–353.

[11] O. Kesemen, Ö. Tezel, and E. Özkul, *Fuzzy c-means clustering algorithm for directional data (fcm4dd)*, Expert Systems with Applications 58 (2016), pp. 76–82.

[12] C. Kosiol, N. Goldman, and N.H. Buttimore, *A new criterion and method for amino acid classification*, Journal of Theoretical Biology 228 (2004), pp. 97–106.

[13] S.C. Lovell, I.W. Davis, W.B. Arendall, P.I. de Bakker, J.M. Word, M.G. Prisant, J.S. Richardson, and D.C. Richardson, *Structure validation by C_α geometry: ϕ, ψ and C_β deviation*, Proteins: Structure, Function, and Bioinformatics 50 (2003), pp. 437–450.

[14] K.V. Mardia, *Statistics of directional data*, Journal of the Royal Statistical Society. Series B (Methodological) 37 (1975), pp. 349–393. Available at http://www.jstor.org/stable/2984782.

[15] K.V. Mardia, J.T. Kent, and J.M. Bibby, *Multivariate analysis / K.V. Mardia, J.T. Kent, J.M. Bibby*, Academic Press London ; New York, 1979, Available at http://www.loc.gov/catdir/toc/els031/79040922.html.

[16] K.V. Mardia and J. Frellsen, *Statistics of Bivariate von Mises Distributions*, in *Bayesian Methods in Structural Bioinformatics*, T. Hamelryck, K. Mardia, and J. Ferkinghoff-Borg, eds., Springer Berlin Heidelberg, Berlin, Heidelberg (2012), pp. 159–178.

[17] W.M. Rand, *Objective criteria for the evaluation of clustering methods*, Journal of the American Statistical Association 66 (1971), pp. 846–850. Available at http://www.jstor.org/stable/2284239.

[18] A.C. Rencher, *Methods of Multivariate Analysis*, 2nd ed., John Wiley & Sons, 2002, Available at http://www.worldcat.org/isbn/9780471461722.

[19] M.L. Rizzo and G.J. Székely, *Energy distance*, Wiley Interdisciplinary Reviews: Computational Statistics 8 (2016), pp. 27–38.

[20] A. Roy, A. Pal, and U. Garain, *Jclmm: A finite mixture model for clustering of circular-linear data and its application to psoriatic plaque segmentation*, Pattern Recognition 66 (2017), pp. 160–173.

[21] A. Secker, M.N. Davies, A.A. Freitas, J. Timmis, E. Clark, and D.R. Flower, *An artificial immune system for clustering amino acids in the context of protein function classification*, Journal of Mathematical Modelling and Algorithms 8 (2009), pp. 103–123.

[22] L.E. Stanfel, *A new approach to clustering the amino acid*, Journal of Theoretical Biology 183 (1996), pp. 195–205.

[23] G. Szekely, *E-Statistics: The energy of statistical samples*, Tech. Rep., Bowling Green State University, Department of Mathematics and Statistics, 2002.

[24] G.J. Székely and M.L. Rizzo, *Testing for equal distributions in high dimension*, InterStat 5 (2004), pp. 1–6.

[25] G.J. Székely and M.L. Rizzo, *Energy statistics: A class of statistics based on distances*, Journal of Statistical Planning and Inference 143 (2013), pp. 1249–1272.