

Validating protein structure using kernel density estimates

Charles C. Taylor^{a,1}, Kanti V. Mardia^a, Marco Di Marzio^b, Agnese Panzera^b

^a*Department of Statistics, University of Leeds, Leeds LS2 9JT, UK.*

^b*DMQTE, Università di Chieti-Pescara, Viale Pindaro 42, 65127 Pescara, Italy.*

Abstract

Measuring the quality of determined protein structures is a very important problem in bioinformatics. Kernel density estimation is a well-known nonparametric method which is often used for exploratory data analysis. Recent advances, which have extended previous linear methods to multi-dimensional circular data, give a sound basis for the analysis of conformational angles of protein backbones, which lie on the torus. By using an energy test, which is based on interpoint distances, we initially investigate the dependence of the angles on the amino acid type. Then by computing tail probabilities which are based on amino-acid conditional density estimates, a method is proposed which permits inference on a test set of data. This can be used, for example, to validate protein structures, choose between possible protein predictions and highlight unusual residue angles.

Key words: Circular kernel, Conformational angle, Probability contour, Variable bandwidth, von Mises density.

1. Introduction

Determination of protein structures is often carried out using X-ray crystallography, which leads to a set of co-ordinates of all the atoms — measured within some resolution. Such structures are typically made available in the Protein Data Bank, but they can be of variable quality. Currently, there is a validation suite [10] of software which provides a set of tools to validate and check structure data. A more recent approach, based on conformational angles, has also been proposed by [12], and this paper builds on their approach.

Email addresses: charles@maths.leeds.ac.uk (Charles C. Taylor), k.v.mardia@leeds.ac.uk (Kanti V. Mardia), mdimarzio@unich.it (Marco Di Marzio), agnesepanzera@yahoo.it (Agnese Panzera)

¹Corresponding author

A *circular* observation can be seen as a point on the unit circle, and represented by an angle $\theta \in [-\pi, \pi)$. It is periodic, *i.e.* $\theta = \theta + 2m\pi$ for $m \in \mathbb{Z}$, which sets apart circular statistical analysis from standard real-line methods. Recent accounts are given by [7] and [13]. Concerning nonparametric density estimation, there exist a few contributions focused on data lying on the circle or on the sphere ([1], [2], [9], [17]). Recently, [6] obtained general results for kernel estimation of densities (and their partial derivatives) defined on the d -dimensional torus $\mathbb{T} := [-\pi, \pi]^d$.

Data on the (two-dimensional) torus are commonly found in descriptions of protein structure. Here, the protein backbone is given by a set of atom co-ordinates in \mathbb{R}^3 which can then be converted (without any loss of information) to a sequence of *conformation angles*. The sequence of angles can be used to assign [8] the structure of that part of the backbone (for example α -helix, β -sheet) which can then give insights into the functionality of the protein. A potential higher-dimensional example is provided by NMR data which will give replicate measurements, revealing a dynamic structure of the protein. For shorter peptides the modes of variability could be studied by an analysis of the replicates, requiring density estimation on a high-dimensional torus. In Section 2.1 we introduce toroidal kernels for kernel density estimation and review a simple way to select the smoothing parameter. Our application, of conformational angles in a protein backbone, is introduced in Section 3, and in Section 4 we investigate whether the bivariate distributions of angles are dependent on the amino acid type. Various *validation* scores, which can be used for “new” proteins”, are introduced in Section 5. We conclude with a discussion.

2. Density estimation on the torus

2.1. Toroidal kernels

A *kernel density estimate* on the circle is easily constructed by adopting a circular density (with mean zero, and concentration parameter λ) for the kernel function. In this case, given angles $\theta_1, \dots, \theta_n$, the kernel density estimate is simply

$$\hat{f}_\lambda(\theta) = \frac{1}{n} \sum_{i=1}^n K_\lambda(\theta - \theta_i)$$

where $\lambda > 0$ is the (inverse of the) smoothing parameter, and $K_\lambda(\cdot)$ is a circular (symmetric) probability density function.

On the torus, we can use a d -fold product $K_C := \prod_{s=1}^d K_{\lambda_s}$, where $C := (\lambda_s \in \mathbb{R}_+, s = 1, \dots, d)$ is a set of smoothing parameters. Most kernels are continuous and symmetric about the origin, so the d -fold products of von Mises, wrapped normal and wrapped Cauchy distributions are all valid. However, we note that the cardioid density (which was used by [12]): $(2\pi)^{-1}\{1 + 2\lambda \cos(\cdot)\}$ with $|\kappa| < 1/2$, $\theta \in \mathbb{T}$ gives a very inefficient [5] kernel relative to the von Mises and wrapped normal kernels.

2.2. A plug-in rule for the von Mises kernel

The performance of a kernel density estimate is usually measured by the integrated mean squared error

$$\text{IMSE} = \int \mathbb{E} (\hat{f}_\lambda(\theta) - f(\theta))^2 d\theta$$

which seeks a trade-off between the bias-squared and variance. In the case that $d = 2$ and a multiplicative von Mises kernel function is adopted, we have a kernel density estimate of $f(\phi, \psi)$ given by

$$\hat{f}_\lambda(\phi, \psi) = \{n(2\pi)^2 I_0(\lambda)^2\}^{-1} \sum_{i=1}^n \exp\{\lambda \cos(\phi - \phi_i) + \lambda \cos(\psi - \psi_i)\}$$

where

- the bivariate data is given by $(\phi_i, \psi_i), i = 1, \dots, n$
- $I_r(\lambda)$ is the modified Bessel function of order r
- λ is the (inverse of the) smoothing parameter (assumed here to be equal for both variables).

Note that distance between two angles is measured by taking the cosine of the difference, which is important when the data may be distributed around the torus.

When f is assumed to be a bivariate von Mises distribution, with independent components, and common concentration κ , then we can approximate [6] the asymptotic integrated variance of the kernel density estimate as

$$\lambda / (4n\pi)$$

with asymptotic integrated bias-squared as

$$\kappa [3\kappa I_0(2\kappa)^2 - I_0(2\kappa)I_1(2\kappa) + \kappa I_1(2\kappa)^2] / (32\pi^2 I_0(\kappa)^4 \lambda^2).$$

As usual, we see a trade-off between bias-squared and variance: as λ increases (corresponding to less smoothing) the bias decreases whilst the variance increases, but when λ decreases the bias increases whilst the variance decreases. In this setting (assuming von Mises data) we can obtain an *asymptotic* choice for λ to minimize the asymptotic IMSE (integral of bias-squared plus variance). We obtain a plug-in rule

$$\lambda^* = [n\hat{\kappa} \{3\hat{\kappa} I_0(2\hat{\kappa})^2 - I_0(2\hat{\kappa})I_1(2\hat{\kappa}) + \hat{\kappa} I_1(2\hat{\kappa})^2\} / (4\pi I_0(\hat{\kappa})^4)]^{(1/3)} \quad (1)$$

where $\hat{\kappa}$ is an estimate of the concentration of the data.

As will be seen in the next section, angles associated with protein structure do not follow a von Mises distribution so we will not adopt (1). In some cases they can be modelled by a mixture of von Mises densities with an EM algorithm being used to fit the components [15]. In the case of data which are not von Mises, [17] investigates robust ways to obtain useful estimates of κ which can be used in (1), though cross-validation provides a more objective approach to the choice of λ . In this case, we choose

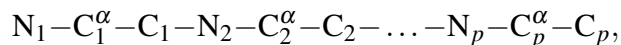
$$\lambda_{\text{CV}} = \arg \min_{\lambda} \prod_{i=1}^n \hat{f}_{\lambda}^{(i)}(\phi_i, \psi_i) \quad (2)$$

where

$$\hat{f}_{\lambda}^{(i)}(\phi_i, \psi_i) = \{n(2\pi)^2 I_0(\lambda)^2\}^{-1} \sum_{j \neq i}^n \exp\{\lambda \cos(\phi_j - \phi_i) + \lambda \cos(\psi_j - \psi_i)\}.$$

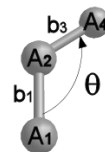
3. Conformational Angles

The *backbone* of a protein comprises a sequence of atoms



By choosing 4 atoms with A_3 directly behind A_2 , and A_1 directly below A_2 we can specify 3 *dihedral angles*: ϕ , ψ , ω .

θ	A_1	A_2	A_3	A_4
ϕ_i	C_{i-1}	N_i	C_i^{α}	C_i
ψ_i	N_i	C_i^{α}	C_i	N_{i+1}
ω_i	C_{i-1}^{α}	C_{i-1}	N_i	C_i^{α}



The angle ω is usually restricted to be about zero. The remaining angles (ϕ , ψ) are measured between $-\pi$ and π . Scatter plots of the (ϕ , ψ) angles for a given protein are known as *Ramachandran plots*; for further details, see [11]. For any protein, it would be possible to compute a kernel density estimate with λ being chosen by cross-validation. The kernel density estimates can be used: (i) to indicate sub-groups in the data; (ii) for

classification purposes [8]; (iii) for estimation of quantiles; and (iv) for clustering. However, it should be noted that — in general — the observations $(\phi_i, \psi_i), i = 1, \dots, n$ will not be independent, and so the usual considerations of IMSE, and general principles underlying cross-validation, may not hold. This lack of independence has been investigated by [3] and has also been modelled by [4] using a Markov model — with bivariate von Mises mixture components; a similar approach might be possible here.

4. Amino acid dependence and inference

Note that each pair (ϕ_i, ψ_i) is associated with an amino acid. There are twenty amino acids, each coded by a single letter — for example Alanine (A) — and we use the letter Z to denote a pre-proline amino acid. For a large database of proteins we can collect all bivariate angles associated with each amino acid. Then we can estimate the probability density for each, say

$$\hat{f}_A(\phi, \psi), \hat{f}_C(\phi, \psi), \hat{f}_E(\phi, \psi), \hat{f}_F(\phi, \psi), \dots$$

It could be of interest to visually compare the distributions and this can be shown graphically using contour representations for the densities.

To make formal comparisons between the densities of angles for two amino acids is feasible using bootstrap methods, or circular analogues of the Kolmogorov-Smirnov test [13]. Such tests reveal that it is possible to detect (statistically significant) differences in the distribution between most amino acids. That is, for most pairs of amino acids we can formally reject the hypothesis $H_0 : f_k(\phi, \psi) = f_j(\phi, \psi) \ k \neq j$. In an all-against-all comparison we can obtain a test statistic based on “energy” [16], and associated p-value for each pair of amino acids. (Note that in such a multiple comparison situation, the threshold for an interesting p-value would be much less than the usual 0.05.) These matrices have the potential for use as additional information within a substitution matrix. Here we use the test statistics as a distance matrix to which multidimensional scaling [14] can be applied. This allows a graphical representation of which amino acids are more similar; the results are shown in Figure 1. It is interesting to note that the energy test cannot detect a difference between the distributions of Alanine (A) and Glutamic Acid (E) even though these have different side-chain polarities and charges. However, given that very few pairings are found to have similar distributions, it seems important *not* to pool together angles from different amino acids.

The kernel density estimate can be converted to *probability* contours as follows. Given α , with $(0 \leq \alpha \leq 1)$ define the set $B_\alpha = \{(\phi, \psi) \mid \hat{f}(\phi, \psi) \geq t(\alpha)\}$ where $t(\alpha)$ is a threshold determined so that

$$\iint_{(\phi, \psi) \in B_\alpha} \hat{f}(\phi, \psi) d\phi d\psi = 1 - \alpha.$$

Protein Similarity and Cliques (isoMDS)

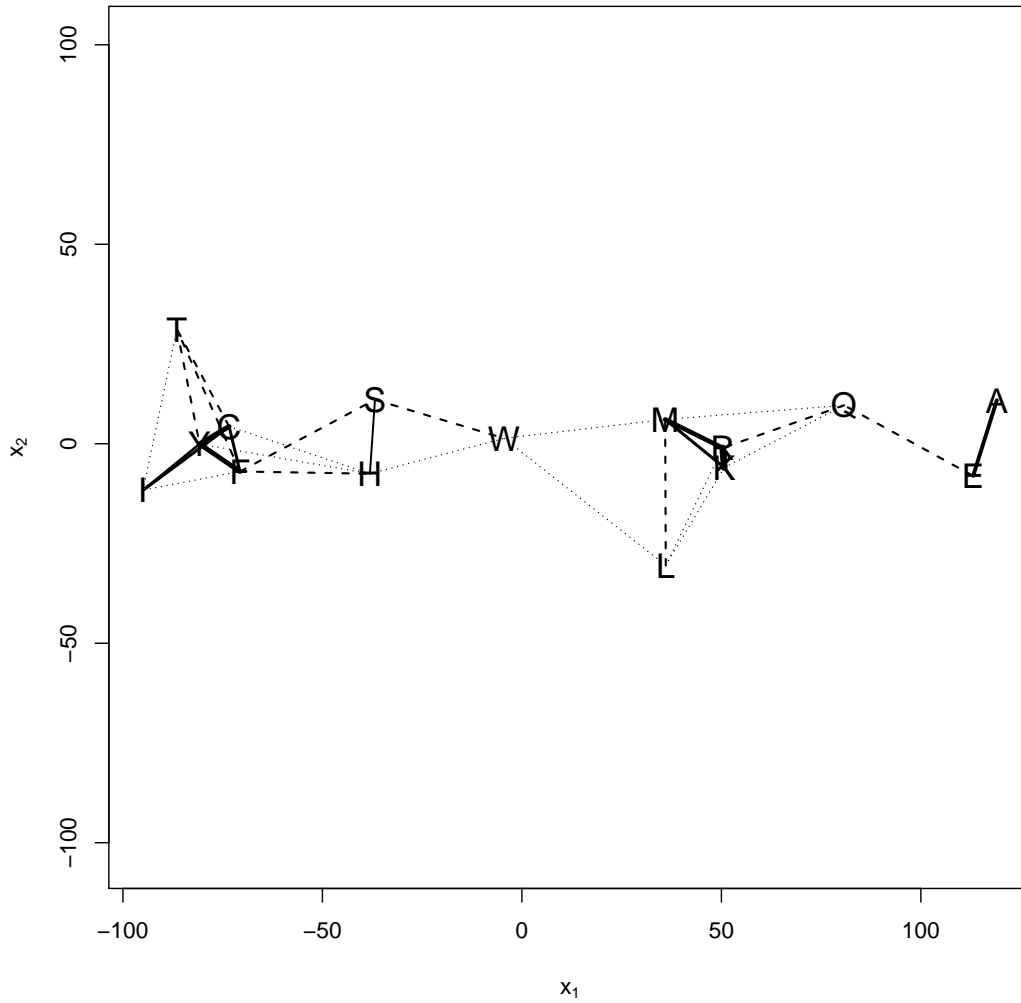


Figure 1: Kruskal's multidimensional scaling for the bootstrap test statistic to measure similarity between the distributions of each pair of amino acids. Those connected by a continuous line have distributions which are not significantly different at 5%; dashed lines are similar at 1%, and dotted lines at 0.1%. Those amino acids which are not connected — and those which are not shown (D, G, N, P, V and pre-proline) — have no similarities with any other (at 0.1% level).

A α -probability contour can then be drawn at the boundary of B_α . Some examples are shown in Figure 2 (for similarities refer to Figure 1). Conversely, given a specific point (ϕ_0, ψ_0) a contour passing through this point will have a specific value of $\alpha = \alpha_0$, say which can be interpreted as a “probability” of occurrence at that location. Such probabilities can be used for validation as described in the next section.

5. Validation for new proteins

Given existing density estimates for each amino acid: $\hat{f}_A(\phi, \psi), \hat{f}_C(\phi, \psi), \dots$ these can be converted (as above) to “probability functions”, say $\hat{P}_A(\phi, \psi), \hat{P}_C(\phi, \psi), \dots$. Then for a “new” n -residue protein (which did not contribute training data to the density estimates) with $\{(\mathcal{A}_i, \phi_i, \psi_i), i = 1, \dots, n\}$ we can compute a probability for the i th residue conditional on the amino acid type $\mathcal{A}_i, i = 1, \dots, n$. We can then create an overall measure of quality using the geometric mean:

$$\left\{ \prod_{\text{jth amino acid type}} \prod_{\{i:\mathcal{A}_i=j\}} \hat{P}_j(\phi_i, \psi_i) \right\}^{1/n} \quad (3)$$

Alternatively (or in addition), we can consider $\min_{i,j} \hat{P}_j(\phi_i, \psi_i)$, or the list $p = (p_1, \dots, p_{21})$ where

$$p_j = \min_{\{i:\mathcal{A}_i=j\}} \hat{P}_j(\phi_i, \psi_i).$$

When a previously validated training dataset is used, a leave-one-out approach can be adopted to provide a benchmark for the above quantities, by which future data can be compared.

Using a cleaned up subset [12] of the top 500 proteins from the Kinemage² database, we can obtain a validation probability (using leave-one-protein-out cross-validation) for each protein. This dataset has 74,414 bivariate angles, with frequencies for each amino acid given in Table 1.

The probabilities can be plotted (vs number of residues to improve clarity) and this plot, together with a histogram, is shown in Figure 3. Given the nature of the data, it is not surprising that the smallest probability is about 0.158 (for protein 1tgsIH), which gives no cause for concern. It is also reassuring that these probabilities do not seem to depend on size.

We now consider a hypothetical “new” protein 1xb1 — which is not in the training database — using the proposed validation procedure, and compare the outcomes with Procheck [10], which is a commonly used tool for the validation of protein structures.

²<http://kinemage.biochem.duke.edu/databases/top500.php>

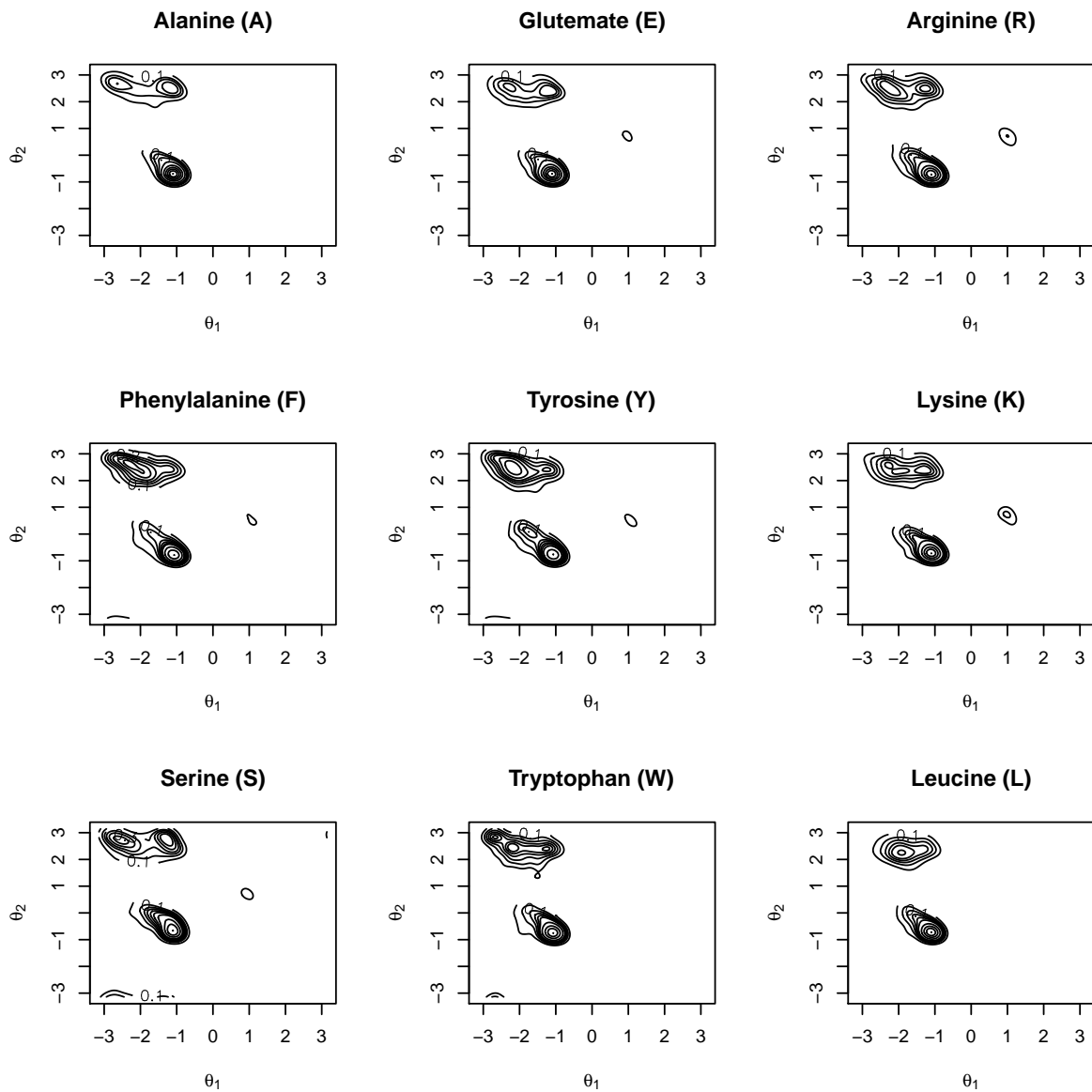


Figure 2: Contour (probability) plots of kernel density estimates for some example proteins. Comparison with Figure 1 shows that the distributions of pairs (A,E), (R,K), (F,Y) are indistinguishable ($p > 0.05$), the distributions of (S,W), (F,S) are very similar ($0.05 > p > 0.01$), and the distributions of (K,L), (R,L), (L,W) are probably distinct ($0.01 > p > 0.001$).

A	7506	C	1299	D	4711
E	4132	F	3313	G	6865
H	1579	I	3772	K	3482
L	5968	M	1399	N	3083
P	1898	Q	2177	R	2621
S	4767	T	4639	V	5607
W	1153	Y	2838	Z	1605

Table 1: Frequencies of amino acids in the database of [12], with “Z” denoting pre-proline

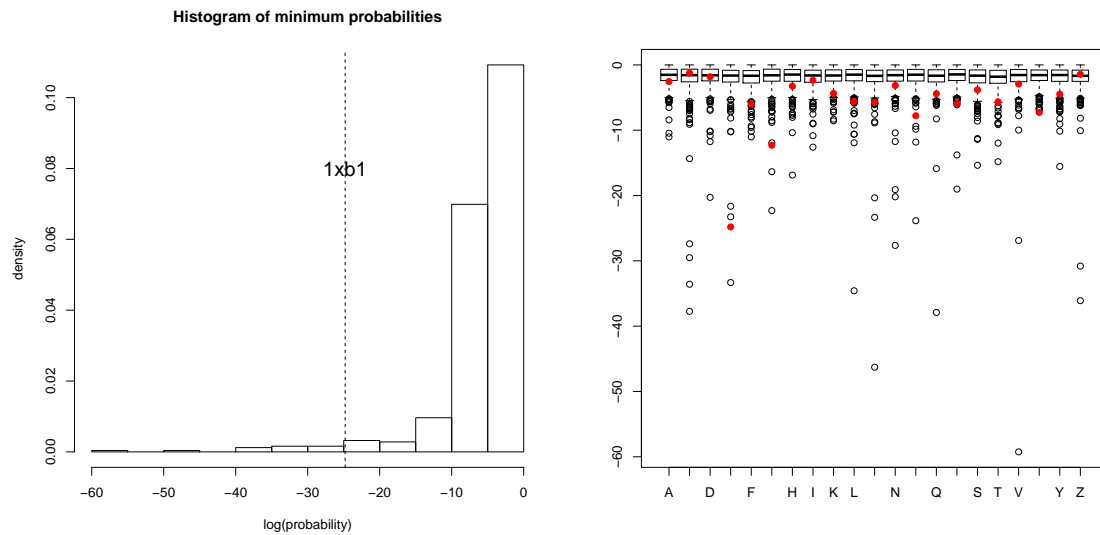


Figure 4: Left: Histogram of log of the minimum tail probabilities for each of the Kinemage proteins, with minimum of the “test” protein 1xb1. Right: Boxplots of the log of the minimum tail probabilities by amino acid for the database, with red points corresponding to 1xb1.

The geometric mean for the angles is 0.28, which is well within a “normal” range of overall validation probabilities (Figure 3). The minimum \hat{P} is 1.7×10^{-11} , which — at first sight — does look significant, and it is this residue that Procheck associates with a “disallowed region”. However, if we consider a similar calculation for each of the 500 Kinemage proteins, then this minimum ranks 14 — see Figure 4. Boxplots for each $\log p_j$ over the Kinemage database, with a comparison of the corresponding values for protein 1xb1 provides another visual check (Figure 4). More formal tests which compare the distribution of the minimum amino acid tail probabilities of the Kinemage proteins with 1xb1 give p-values ranging around 0.1.

6. Discussion

The above probabilities all critically depend on the choice of smoothing parameter — λ for the von Mises kernel. In general, the smaller is λ (which corresponds to more smoothing), the larger is the validation probability, and the less sensitive is the test.

In principle, the above method could be used on any dataset. Ideally, one would like *training data* which consists of a large database of independent bivariate observations which are known to be “correct”. (Note that our use of the Kinemage database does not have the sought-after independence.) Probability estimates $\hat{P}_A(\phi, \psi), \hat{P}_C(\phi, \psi), \dots$ for each amino acid require selection of the λ ’s which could be obtained by cross-validation (or a plug-in rule). Having stored the \hat{P} ’s (on a reasonably fine grid on the torus) then formula (3) could be used to validate any new protein structure.

An alternative to cross-validation for selection of the λ ’s is to consider an *adaptive* kernel bandwidth. Theoretical results — see, for example [18] — suggest that using a separate bandwidth for each observation, with a bandwidth that depends on density, will have better theoretical properties. This approach has been adopted by [12], although they have used an inefficient Cardioid kernel, and an adaptation that is not consistent with theory which suggests that, for a von Mises kernel with concentration λ , the adaptive bandwidth for observation i should satisfy $\lambda_i \propto f(\phi_i, \psi_i)$. However, limited simulation experiments suggest that $\lambda_i \propto f(\phi_i, \psi_i)^\gamma$ with $\gamma \approx 0.7$ may work better for large datasets. Further work will investigate this more closely, as well as examining which of the above proposed scores are more useful.

Acknowledgements

We are grateful to Simon Lovell for providing the filtered data from the “top 500” proteins. The paper was improved after initial discussions with Andy Neuwald and Wally Gilks, both of whom also made helpful comments on an early draft.

References

- [1] Z. D. Bai, R. C. Rao, and L. C. Zhao. (1988). Kernel estimators of density function of directional data. *Journal of Multivariate Analysis*, 27:24–39.
- [2] R. Beran. (1979). Exponential models for directional data. *The Annals of Statistics*, 7:1162–1178.
- [3] D. S. Berkholz, M. V. Shapovalov, R. L. Dunbrack, Jr. and P. A. Karplus (2009) *Structure*, 17:1316–1325.
- [4] W. Boomsma, K. V. Mardia, C. C. Taylor, J. Ferkinghoff-Borg, A. Krogh, and T. Hamelryck. (2008) A generative, probabilistic model of local protein structure. *PNAS*, 105:8932–8937.
- [5] M. Di Marzio, A. Panzera and C.C. Taylor. (2009). Local polynomial regression for circular predictors. *Statistics and Probability Letters*, 79:2066–2075.
- [6] M. Di Marzio, A. Panzera and C.C. Taylor. (2011). Kernel density estimation on the torus. *Journal of Statistical Planning and Inference*, 141:2156–2173.
- [7] S. R. Jammalamadaka and A. SenGupta. (2001). *Topics in Circular Statistics*. World Scientific, Singapore.
- [8] W. Kabsch and C. Sander. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577–2637.
- [9] J. Klemelä. (2000). Estimation of densities and derivatives of densities with directional data. *Journal of Multivariate Analysis*, 73:18–40.
- [10] R. A. Laskowski, M. W. MacArthur, D. S. Moss, J. M. Thornton (1993). PROCHECK - a program to check the stereochemical quality of protein structures. *J. App. Cryst*, 26:283–291.
- [11] A. M. Lesk (2010) *ntroduction to protein science : architecture, function, and genomics. 2nd edition* Oxford: Oxford University Press.
- [12] S.C. Lovell, I.W. Davis, W.B. Arendall III, P.I.W. de Bakker, J.M. Word, M.G. Prisant, J.S. Richardson, and D.C. Richardson. (2003) Structure Validation by $C\alpha$ Geometry: ϕ , ψ and $C\beta$ Deviation. *Proteins: Structure, Function and Genetics*, 50: 437–450.

- [13] K. V. Mardia and P. E. Jupp. (1999). *Directional Statistics*. John Wiley, New York, NY.
- [14] K. V. Mardia, J. T. Kent and J. M. Bibby (1979). *Multivariate Analysis*. London: Academic Press.
- [15] K.V. Mardia, C.C. Taylor, and G.K. Subramaniam. (2007). Protein bioinformatics and mixtures of bivariate von mises distributions for angular data. *Biometrics*, 63: 505–512.
- [16] M. L. Rizzo (2007) *Statistical Computing with R*. Boca Raton: Chapman & Hall.
- [17] C. C. Taylor. (2008). Automatic bandwidth selection for circular density estimation. *Computational Statistics & Data Analysis*, 52:3493–3500.
- [18] G. R. Terrell, and D. W. Scott. (1992). Variable kernel density estimation. *Annals of Statistics*, 20:, 1236–1265.