Biometrics NA, 1–20 DOI: 10.1111/j.1541-0420.2005.00454.x
NA NA

A statistical model for helices with applications

Kanti V Mardia[1,2,*], Karthik Sriram[3,], and Charlotte M Deane[1,***]**

[1]Department of Statistics, University of Oxford, Oxford, UK

[2]Department of Statistics, School of Mathematics, University of Leeds, Leeds LS2 9JT, UK

[3]Quantitative Methods area, Indian Institute of Management Ahmedabad, Ahmedabad, Gujrat, India.

*_email:_ K.V.Mardia@leeds.ac.uk

**_email:_ karthiks@iima.ac.in

***_email:_ deane@stats.ox.ac.uk

Summary: Motivated by a cutting edge problem related to the shape of $\alpha-$helices in proteins, we formulate a parametric statistical model, which incorporates the cylindrical nature of the helix. Our focus is to detect a "kink", which is a drastic change in the axial direction of the helix. We propose a statistical model for the straight $\alpha-$helix and derive the maximum likelihood estimation procedure. The cylinder is an accepted geometric model for $\alpha-$helices, but our statistical formulation, for the first time, quantifies the uncertainty in atom-positions around the cylinder. We propose a change point technique "Kink-Detector" to detect a kink location along the helix. Unlike classical change point problems, the change in direction of a helix depends on a simultaneous shift of multiple data points rather than a single data point, and is less straightforward. Our biological building block is crowdsourced data on straight and kinked helices; which has set a gold standard. We use this data to identify salient features to construct Kink-Detector, test its performance and gain some insights. We find the performance of Kink-Detector comparable to its computational competitor called "Kink-Finder". We highlight that identification of kinks by visual assessment can have limitations and Kink-Detector may help in such cases. Further, an analysis of crowdsourced curved $\alpha-$helices finds that Kink-Detector is also effective in detecting moderate changes in axial directions.

Key words: Change point; Crowdsourced data; Helix fitting; Kink detection; Membrane protein; Protein structure.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Proteins are considered the workhorses of life. There have been recent studies into the statistical aspects of the 3D atomic configuration of protein structures (see Mardia, 2013). The shape of a protein plays an important role in its function, and proteins contain some specific shapes (secondary structures) such as helices. Among helices, the most common is the $\alpha-$helix, which is of a right-handed spiral shape. In membrane proteins, it is known that $\alpha-$helices appear frequently with a "kink". These are called kinked helices in contrast to straight helices. These kinked helices are often functionally important (see Sansom and Weinstein, 2000), particularly in cellular processes and drug targets. Figure 1 shows a plot of a straight helix and a kinked helix (for a schematic diagram of a membrane protein with kinked and straight helices, see Web Figure 1 Web Appendix A). Kinks have been commonly studied in soluble proteins, for which more structural data is available (see Yohannan et al., 2004; Meruelo et al., 2011; Bansal et al., 2000). In this paper, we use data on $\alpha-$helices for membrane proteins originally from Kneissl et al. (2011).

A large number of methods exist to identify kinks in $\alpha-$helices in proteins. Such structure-based kink identification methods use the three-dimensional atomic coordinates of the C_α atoms as a basis for kink identification. Wilman, Ebejer, Shi, Deane, and Knapp (2014) have given an overview together with what softwares are available. To name a few softwares: ProKink (Visiers et al., 2000), TMkink (Meruelo et al., 2011), Helanal (Bansal et al., 2000), Helanal-Plus (Kumar and Bansal, 2012), McHelan (Langelaan et al., 2010) and Kink-Finder (Wilman, Shi, and Deane, 2014). These methods identify the $\alpha-$helices within a protein using a variety of approaches and then analyze these $\alpha-$helices to identify kinks. They differ in a number of areas, but the main differences are in the way they fit helix axes, the length and the segments of the helix to which they fit the axes. In general, they identify the position of the kink based on the residue with the largest angle.

Crowdsourced Data: Wilman, Ebejer, Shi, Deane, and Knapp (2014) have carried out a comparison of competitive methods after creating a gold standard data based on crowdsourcing. The crowdsourced data has 300 helices, which have been drawn randomly from the manually annotated data of 1014 helices in Kneissl et al. (2011). Kneissl annotation uses three classes: Straight, Kinked and Curved. Web Table 1 in Web Appendix B shows the Kneissl classification together with the crowdsourcing classification. We note that 48 of the 300 helices remained unclassified in the crowdsourcing exercise, thus leaving 252 helices classified: 129 straight, 64 kinked and 59 curved helices. Wilman, Ebejer, Shi, Deane, and Knapp (2014) have chosen three competitive methods for a comparison with Kink-Finder, viz. McHelan, Helanal-Plus and manual annotation by Kneissl et al. (2011). They concluded that the Kink-Finder and Kneissl et al. (2011) classifications are more consistent with the crowdsourced classifications than the other two methods. Note that the Kneissl classification is based on manual annotation and crowdsourcing "refines" the classification of the 300 helices, leading to the gold standard data or crowdsourced data, and therefore we are left with Kink-Finder as our benchmark.

Blundell et al. (1983) and Barlow and Thornton (1988) initiated the work using curvatures of the helices and used main classification as Straight, Kinked and Curved. There are several ways to define these classifications but it is better to recall how Wilman, Ebejer, Shi, Deane, and Knapp (2014) formulated these for their experiments.

• Kinked: There is a clear location where the direction of the helix changes. Only a small part of the helix is involved in this.

• Curved: There is a slow but steady change of the direction of the helix. This can happen over a large part or even all of the helix.

• Straight: There is no change in the overall direction of the helix.

The main point to note is that so far there is no method based on a statistical model. We

will take Kink-Finder as the main benchmark, noting that the comparison of a model–based versus a computational method is gratifying but our novel statistical model stands on its own and provides additional insights.

[Figure 1 about here.]

This paper has two main methodological objectives: the first is to fit a curve (helix) on a cylinder and another is to find the 'change point' on the curve. Fitting a curve on a manifold (namely fitting a small circle on a sphere) has been studied initially by Mardia and Gadsden (1977) and very recently by Jung et al. (2011) among others. There are some similarities, but here we are dealing with a curve on a cylinder, where the data points are ordered.
The data here consists of three coordinates; so it can be thought of as a change point problem in multivariate analysis. In multivariate analysis, the change point problem using a Gaussian model has been studied initially by Srivastava and Worsley (1986) and recently by Siegmund et al. (2011) among others. While their focus is on shifting of mean, our problem is in the non-Euclidean setting relating to change in the axial direction of a cylinder. The problem of a change point on a manifold is not straightforward, see for example, change point on a circular manifold in Rueda et al. (2016). Furthermore, our problem is not simply a change point problem on a manifold, in the sense that the change point is not a single point, but changes in the direction of the helix axis (which depends on multiple points). So, we can describe our problem as a "regional change point problem".

In Section 2, we describe our statistical model for a straight helix and give the maximum likelihood estimation procedure in Section 3. Using this procedure, we train our model on a crowdsourced data of straight helices (Wilman, Ebejer, Shi, Deane, and Knapp, 2014) in Section 4; this data provides a "gold standard". Thus we have the parameter values of the model for straight helices and hence a full specification of the distribution under the null hypothesis of a straight helix. We propose to detect the presence and position of a kink based

on a particular departure from the null model, where there is a definite gradual change point. The proposed method, which we call "Kink-Detector", is discussed in Section 5. In Section 6, we test the Kink-Detector on the crowdsourced data and find its performance comparable to Kink-Finder. Here, we also demonstrate an interesting finding that identification of straight or kinked helices by visual assessment in the experimental data has some limitation when the kink angles are small, and our method provides some insight into such cases. Further, based on an analysis of curved helices we find that Kink-Detector is more accurate in detecting moderate changes in axial directions than Kink-Finder. We conclude with a discussion in Section 7. The web supplement included with this paper provides supporting details relating to the helix structure, estimation procedure and output from Kink-Detector when applied to crowdsourced data. We use some specific helices (numbered Helix 1 to Helix 9) from the crowdsourced data in this paper and its web supplement. Helices 1 to 7 are cases of misclassification by Kink-Detector while classifying helices into "kinked" versus "straight". Helices 8 and 9 are used as examples in the web supplement to illustrate the estimation procedure and its application. For ease of reference, Table 1 lists these helices and provides a brief description of the context in which they are discussed.

[Table 1 about here.]

2. The Straight Helix Model

Let $(x_i, y_i, z_i), i = 1, \ldots, n$ be n consecutive points on a helix. For the $\alpha-$helix, we take these n points for our study here to be C_α atoms which are the key atoms of a protein chain (see for example Mardia, 2013). The C_α is the central atom of each residue and is commonly used to trace an $\alpha-$helix in display programmes. It is easier to first formulate the straight helix model for an "aligned $\alpha-$helix", i.e. where the axis of the helix is aligned with the $z-$axis.

We then extend it to an "unaligned $\alpha-$helix", i.e. where the axis direction is not aligned with $z-$axis.

Aligned $\alpha-$helix: We begin with a model formulation for a straight helix where its axis is known. Without any loss of generality, we take it to be the z-axis. For the aligned straight helix, our model formulation with random errors is as follows.

$$x_i = a\cos t_i + \epsilon_{1i}, y_i = a\sin t_i + \epsilon_{2i}, z_i = ct_i + \epsilon_{3i};\ i = 1, \ldots, n;\ a > 0, c > 0, \qquad (1)$$

where the axis of the aligned straight helix is the z–axis, a is the radius of the cylinder, and $2\pi c$ is the pitch (i.e. the vertical distance between consecutive turns of the helix). Further, we assume the errors $\{\epsilon_{1i}, \epsilon_{2i}, \epsilon_{3i}\}$ to be independent and normally distributed, i.e. $N(0, \sigma^2\boldsymbol{\mathcal{I}})$, where $\boldsymbol{\mathcal{I}}$ is the 3-dimensional identity matrix. Web Figure 2 in Web Appendix C plots an aligned helix with the model parameters given in equation (1).

For the aligned $\alpha-$helix, based on previous empirical studies, the "ideal" (accepted) value for a is taken as 2.3 Å and for the pitch 5.4 Å (see for example Dickerson and Geis, 1969, pp 26–28), where we note that the distance between two atoms is measured in Ångström (Å) and 1Å$= 10^{-10}$m. Note that the $\{t_i\}$ are in radians and for the ideal α–helix, the coordinates on the cylinders move in steps of 100 degrees or $\frac{2\pi}{3.6}$ radians. In other words, there are 3.6 atoms for every full turn of 2π radians. Hence, it follows that $t_i = \frac{2\pi i}{3.6}$ and $c = \frac{5.4}{2\pi} = 0.86$Å. This means that here $t_i = \beta i$ with the "ideal" value of β is 1.75 radians or 100°. So, the ideal values of $\alpha-$helix parameters in (1) are:

$$a = 2.3, c = 0.86, \beta = 1.75, t_i = \beta i, i = 1, \ldots, n; \qquad (2)$$

so $\{t_i\}$ are known except for the parameter β. Later, in Section 4, we show that the estimates for (a, c, β) obtained by applying our statistical model on the crowdsourced data for straight helices provide support for the aforementioned ideal values. Furthermore, based on our statistical model, for the first time, we provide additional insight on the uncertainty around the parameter σ^2.

Unaligned $\alpha-$helix: If the axis of the helix is not already aligned, then one needs to work with the unaligned data, say, $(x_{0i}, y_{0i}, z_{0i}), i = 1, \ldots, n$. We write

$$\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots \mathbf{X}_n), \mathbf{X_0} = (\mathbf{X}_{01}, \mathbf{X}_{02}, \ldots \mathbf{X}_{0n}) \tag{3}$$

where $\mathbf{X}_i = (x_i, y_i, z_i)^T$ and $\mathbf{X}_{0i} = (x_{0i}, y_{0i}, z_{0i})^T, i = 1, \ldots, n$. Then the $3 \times n$ data matrices $\mathbf{X}$ and $\mathbf{X_0}$ are related by a rigid transformation, say,

$$\mathbf{X} = \mathbf{A}\mathbf{X_0} + \mathbf{B}, \tag{4}$$

where $\mathbf{A}$ is a 3×3 rotation matrix and $\mathbf{B} = \mathbf{b}\mathbf{1}_n^T$ where $\mathbf{1}_n$ is an n dimensional vector of 1's and $\mathbf{b}$ is a translation vector. Further, we write the parameter matrices $\mathbf{\Delta}$ and $\mathbf{D}$ as follows:

$$\mathbf{\Delta} = diag(a, a, c), \mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2, \ldots \mathbf{d}_n). \tag{5}$$

where $\mathbf{d}_i = (\cos t_i, \sin t_i, t_i)^T, i = 1, \ldots, n$, Then, the model for the unaligned helix becomes

$$\mathbf{A}\mathbf{X_0} + \mathbf{B} = \mathbf{\Delta}\mathbf{D} + \boldsymbol{\epsilon}, \tag{6}$$

where $\boldsymbol{\epsilon}$ is $3 \times n$ matrix with its elements ϵ_{ij}. Equivalently, we have

$$\mathbf{X_0} = -\mathbf{A}^T\mathbf{B} + \mathbf{A}^T\mathbf{\Delta}\mathbf{D} + \boldsymbol{\epsilon_1}, \tag{7}$$

where $\boldsymbol{\epsilon_1} = \mathbf{A}^T\boldsymbol{\epsilon}$. Since the matrix $\mathbf{A}$ is orthonormal and each column vector of $\boldsymbol{\epsilon}$ is $N(0, \sigma^2\boldsymbol{\mathcal{I}})$, each column vector of $\boldsymbol{\epsilon_1}$ will also be $N(0, \sigma^2\boldsymbol{\mathcal{I}})$.

3. Maximum Likelihood Estimation

We now obtain the maximum likelihood estimates (MLE) $(\widehat{a}, \widehat{c}, \widehat{\mathbf{b}}, \widehat{\mathbf{A}}, \{\widehat{t_i}\}, \widehat{\sigma^2})$ for the parameters $(a, c, \mathbf{b}, \mathbf{A}, \{t_i\}, \sigma^2)$ of the general model given in (7). There are no closed form expressions for the parameters and hence the estimation involves iterations. Here, we derive the system of mathematical equations required for the iterations, and defer the implementation details to Web Appendix D.

The $-2\times$ log likelihood function (except for a constant) for the data matrix $\mathbf{X_0}$ is given by

$$3n \log \sigma^2 + Trace(\mathbf{X_0} + \mathbf{A}^T\mathbf{B} - \mathbf{A}^T\mathbf{\Delta}\mathbf{D})^T(\mathbf{X_0} + \mathbf{A}^T\mathbf{B} - \mathbf{A}^T\mathbf{\Delta}\mathbf{D})/\sigma^2.$$

First we note that given all other estimates except σ^2 and the aligned coordinates as $\mathbf{X} = \widehat{\mathbf{A}}\mathbf{X_0} + \widehat{\mathbf{B}}$, we have

$$\widehat{\sigma^2} = \sum_{i=1}^{n} \left\{ \left(x_i - \widehat{a}\cos\widehat{t_i}\right)^2 + \left(y_i - \widehat{a}\sin\widehat{t_i}\right)^2 + \left(z_i - \widehat{c}\,\widehat{t_i}\right)^2 \right\} /(3n). \tag{8}$$

This is the mean squared deviation of the aligned coordinates from the expected helix position. Next, for the remaining parameters, rotation matrix $\mathbf{A}$, translation vector $\mathbf{b}$, helix parameters $\mathbf{\Delta}$ (equivalently a, c) and $\{t_i\}_{i=1}^n$, we solve the following minimization problem.

$$\min_{(a,c,\mathbf{b},\mathbf{A},\{t_i\},\sigma^2)} Trace(\mathbf{X_0} + \mathbf{A}^T\mathbf{B} - \mathbf{A}^T\mathbf{\Delta}\mathbf{D})^T(\mathbf{X_0} + \mathbf{A}^T\mathbf{B} - \mathbf{A}^T\mathbf{\Delta}\mathbf{D}). \tag{9}$$

Since $\mathbf{A}$ is an orthonormal matrix, this is equivalent to solving the following problem.

$$\min_{(a,c,\mathbf{b},\mathbf{A},\{t_i\},\sigma^2)} Trace(\mathbf{A}\mathbf{X_0} + \mathbf{B} - \mathbf{\Delta}\mathbf{D})^T(\mathbf{A}\mathbf{X_0} + \mathbf{B} - \mathbf{\Delta}\mathbf{D}). \tag{10}$$

Further, we have $\mathbf{X} = \mathbf{A}\mathbf{X_0} + \mathbf{B}$, so (10) reduces to

$$\sum_{i=1}^{n} \left\{ (x_i - a\cos t_i)^2 + (y_i - a\sin t_i)^2 + (z_i - ct_i)^2 \right\}. \tag{11}$$

Differentiating this with respect to a, c and t_i gives the equations for the respective MLE as

$$\widehat{a} = \sum_{i=1}^{n} \left(x_i\cos\widehat{t_i} + y_i\sin\widehat{t_i} \right) /n, \tag{12}$$

$$\widehat{c} = \frac{\sum_{i=1}^{n} z_i\widehat{t_i}}{\sum_{i=1}^{n} \widehat{t_i^2}}, \tag{13}$$

$$\widehat{a}\ x_i\sin\widehat{t_i} - \widehat{a}\ y_i\cos\widehat{t_i} + \widehat{c}^2\ \widehat{t_i} = \widehat{c}z_i. \tag{14}$$

The estimates for $\widehat{\mathbf{\Delta}}$ and $\widehat{\mathbf{D}}$ can now be obtained substituting $\widehat{a}, \widehat{c}, \{\widehat{t_i}\}$ in equation (5). For the $\alpha-$helix, we have pointed out that $t_i = \beta i$; so $\{t_i, i = 1, 2, \ldots, n\}$ are known except for the parameter β. Substituting $t_i = \beta i$ in equation (11) and differentiating with respect to β, we obtain the equation for the MLE of β as

$$\hat{a}\sum_{i=1}^{n} i\ x_i\sin(\hat{\beta}i) - \hat{a}\sum_{i=1}^{n} i\ y_i\cos(\hat{\beta}i) + \hat{c}^2\hat{\beta}\sum_{i=1}^{n} i^2 = \hat{c}\sum_{i=1}^{n} iz_i. \tag{15}$$

Furthermore, substituting $\hat{c}$ from equation (13), this expression simplifies to

$$\sum_{i=1}^{n} i\ x_i\sin(\hat{\beta}i) = \sum_{i=1}^{n} i\ y_i\cos(\hat{\beta}i). \tag{16}$$

We need to iterate between (12), (13) and (16) to obtain $(\widehat{a}, \widehat{c}, \widehat{\beta})$ and these values are substituted in (8) to obtain $\widehat{\sigma^2}$. To obtain $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{b}}$, we go back to equation (10). By minimizing this equation with respect to $\mathbf{b}$, we find that

$$\widehat{\mathbf{b}} = \left(\widehat{\mathbf{\Delta}}\widehat{\mathbf{D}} - \widehat{\mathbf{A}}\mathbf{X_0}\right) \times \mathbf{1}_n/n, \quad \text{and } \widehat{\mathbf{B}} = \widehat{\mathbf{b}} \times \mathbf{1}_n^T. \tag{17}$$

Let $\widehat{\mathbf{M}} = \widehat{\mathbf{\Delta}}\widehat{\mathbf{D}} - \widehat{\mathbf{B}}$ and now it remains to obtain $\widehat{\mathbf{A}}$ by minimizing $Trace(\mathbf{A}\mathbf{X_0} - \widehat{\mathbf{M}})^T(\mathbf{A}\mathbf{X_0} - \widehat{\mathbf{M}})$ with respect to $\mathbf{A}$. This is the standard orthonormal Procrustes problem (see for example Mardia et al. 1979, pp 416-417 and Dryden and Mardia 2016, pp 125-132) . Suppose the singular value decomposition of $\mathbf{X_0}\widehat{\mathbf{M}}^T = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T$, where $\mathbf{U}, \mathbf{V}$ are orthonormal and $\mathbf{\Lambda}$ is a diagonal matrix. Then, the solution to the problem is given by

$$\widehat{\mathbf{A}} = \mathbf{V}\mathbf{U}^T. \tag{18}$$

In summary, the MLE of $(\sigma^2, a, c, \beta, \mathbf{b}, \mathbf{A})$ are obtained by iteratively solving (8),(12), (13), (16),(17), and (18) respectively. In particular, since the axis for the aligned straight $\alpha-$helix is just the $z-$axis, the estimated axis line for the unaligned $\alpha-$helix is given by:

$$\widehat{\mathbf{axis}} = [0\ 0\ z] \times \widehat{\mathbf{A}} - \widehat{\mathbf{b}}^T \times \widehat{\mathbf{A}},\ z \in \mathbb{R}^1. \tag{19}$$

The above procedure estimates 10 unknown parameters: 4 for the aligned $\alpha-$helix (viz. a, c, β, σ), and additional 6 for the unaligned $\alpha-$helix (namely, 3 for the translation vector $\mathbf{b}$ and 3 for the rotation matrix $\mathbf{A}$). The implementation details along with an illustrative example are provided in Web Appendix D. We make some remarks on identifiability, existence of solutions and properties of the MLE in Web Appendix E.

4. Crowdsourced Data and the Null Distribution for Straight Helices

We now implement the estimation procedure for the $\alpha-$helix on the crowdsourced data of straight helices. This data has set a "gold standard" and Wilman, Ebejer, Shi, Deane, and Knapp (2014) have given a comparative study for various computational methods. The crowdsourced data has 129 straight, 64 kinked and 59 curved helices. Helices with different

lengths are represented in the data (length = number of C_α atoms). Web Tables 3, 4 and 5 in Web Appendix F give the frequency distribution of lengths of helices.

In this section, we will concentrate on the straight helix data. We train our model on 129 straight helices and thus estimate the parameters of the null distribution, i.e. when the helix is straight. Since we will be interested later on in the angle between successively computed axes for the kink detection, which is invariant under rotation of the co-ordinate system, it suffices to work with the null model for the aligned helix (i.e. with the helix-axis aligned with the $z-$axis), as in equation (1). Therefore, we need a set of parameters a, c, β and σ^2 that describe the aligned null distribution as in equation (1), with $t_i = \beta i$. That is, we need to compute the MLE of the parameters a, c, β and σ^2 for these helices as in Section 3. The estimation procedure (see Web Appendix D) was applied to the 129 straight helices and estimates for $a,\ c,\ \sigma^2$ and β were obtained. The values for c and β were more well-behaved and distributed around 0.86 and 1.75 respectively. However, while most of the values obtained for a were distributed around the expected ideal value 2.3, there was a visible set of outliers taking values less than 0.5. Similarly, while most values of σ^2 happened to be clustered near 0.05, there were many outliers as well. For reference, the plots of the 129 estimated values for $a,\ c,\ \sigma^2$ and β are given in Web Figure 5 in Web Appendix G.

On further investigation of the outliers, we found that these were caused by some single outlier atom positions, typically atoms at each of the end points of the helices. To obtain "robust" maximum likelihood estimates, for each helix of length n, we computed the MLE with three different sequences of atoms (i) by considering atoms 1 to (n-2), (ii) atoms 2 to (n-1) and (iii) atoms 3 to n. Then among these three sets we picked the set of that MLE corresponding to the smallest value of estimated σ^2. Figure 2 shows the plot of estimates obtained through this robust MLE procedure. We note that the cluster of the outliers is now not seen and all the estimates are well behaved. Table 2 gives the mean and standard errors

calculated using the 129 different values of these robust MLE for all the four parameters a, c, σ^2 and β.

We note from Table 2 that the estimates of a, c and β are close to the ideal values. Further, they have small standard errors. Therefore, for the null distribution, we set $a = 2.3$Å, $c = 0.86$Å and $\beta = 1.75$, i.e. the ideal values in equation (2). While there is no such known ideal value for σ^2, we set the estimate based on crowdsourced data on straight helices of $\sigma^2 = 0.056$ for the null distribution. We note that for $\widehat{\sigma^2}$, the standard error is relatively larger than that of other parameters. The large value is expected for the $\alpha-$helix since it is well known that the atoms do not strictly sit on the cylinder (see for example Wilman, 2014). Although the cylinder is an established geometric model for $\alpha-$helices, an important point to note is that our statistical formulation, for the first time, quantifies the uncertainty about σ^2. To sum up, the parameter values of our null distribution for the straight $\alpha-$helix are fixed as follows:

$$a = 2.3, c = 0.86, \beta = 1.75, \sigma^2 = 0.056. \tag{20}$$

Further in the following,when we treat the unaligned case, the axis under this null distribution will be estimated from (19).

[Figure 2 about here.]

[Table 2 about here.]

5. "Kink-Detector": A Procedure for Detecting Kinks

We name our proposed method for detecting kinks in an $\alpha-$helix as the "Kink-Detector". The method is based on detecting statistically significant changes in the axis angle as we move along the $\alpha-$helix. Our idea is to estimate the angle, say θ, between axes fitted to two successive sets of $k = 6$ C_α atoms, and check whether there is a "critical deviation", i.e. whether θ exceeds an "angle-threshold" $T = 11$ degrees. We found it useful sometimes

to work with $\cos\theta$ rather than θ itself. The axes are computed using the null distribution specified in the last section. Our method is sequential, in that we work with a moving window of $k = 6$ successive C_α atoms along the $\alpha-$helix of length n, thus obtaining a sequence of angles $\{\theta_k, \theta_{k+1}, \ldots, \theta_{n-k+1}\}$. The angle θ_k is between axis fitted to C_α atoms $\{1, \ldots, k\}$ and $\{k+1, \ldots, 2k\}$, angle θ_{k+1} is between axis fitted to C_α atoms $\{2, \ldots, k+1\}$ and $\{k+2, \ldots, 2k+1\}$, and so on (see Web Appendix H for details on the computation of angle between axes). We call k the "atom-window size". Further, since kink detection is a regional change point problem, we find it important to check not one but a clustered set of critical deviations. Towards this, we consider every run of critical deviations, i.e. every maximal set of consecutive angles that exceed the angle-threshold, and we denote $F =$ length of this run. We set a "critical-deviation-run-threshold" $r = 4$, and declare the presence of a kink whenever $f \geqslant r$. If there is no such run, we conclude that there is no kink in the helix.

Kink Location: Since detecting a kink is a regional change point problem, there is strictly no unique atom position corresponding to a kink. However, if one insists on a kink change-point as some papers do (for example Wilman, Shi, and Deane 2014; Sansom and Weinstein 2000), we recommend the point with the maximum angle (or minimum cosine value) within any run of critical deviations with length $f \geqslant r$. See Web Figure 6 in Web Appendix H for an illustration of changing axis direction, which depends on multiple points rather than a single point, and Web Table 6 for assigning kink location. Next, we discuss our rationale for the choice of values for (k, T, r).

5.1 *Choice of tuning parameters*

The atom-window size (k), the angle-threshold (T) and critical-deviation-run-threshold (r) are the tuning parameters required for Kink-Detector. We now describe below how our choice of tuning parameters is motivated by contextual considerations, or by some experimentation carried out with a few randomly selected straight and kinked helices from crowdsourced data,

or by both. The adequacy of these choices is further confirmed by a sensitivity analysis of classification accuracy on the full crowdsourced data.

Atom-window size ($k = 6$) – Contextual considerations: The atom-window size $k = 6$ is necessary to avoid degeneracy. It is the minimum number of atoms needed for two full turns of the helix and is therefore required to estimate the axis of the helix. Using larger window size will necessarily require a larger number of atoms on either side of the kink, thus limiting the applicability of the procedure to only larger helices. So, we fix the atom-window size at the smallest meaningful choice of $k = 6$. Indeed, this value of $k = 6$ has been used in Bioinformatics, see for example Kink-Finder by Wilman, Shi, and Deane (2014). Kink-Finder (2014) is the latest software to estimate kinks and we will give some comparative details in the next section.

Angle-threshold ($T = 11$)– Experimentation: The value of $\cos\theta$ is bounded above by 1, and this happens when the angle between the axes is zero (i.e. $\theta = 0$). If the helix is straight then we would expect that the sequence of $\cos\theta$s will stay close to 1. Since there will be some randomness in the coordinates of the atoms, the cosine values will not be exactly equal to 1 and would be subject to some random fluctuations. The aim here is to obtain a lower cutoff value for the cosine or equivalently an upper threshold T for the angle. Unlike k, the value of T cannot be contextually fixed. We simulate a large number $(100,000)$ straight helices of size $2k$ from the null model in equation (1) with the parameter values as specified in (20). For each simulated helix, we compute the angle between the estimated axis based on the first k atoms and the estimated axis based on atoms $k+1$ to $2k$, which gives $100,000$ simulated values of the cosine angle from the null model. We then computed various percentiles (e.g. $5^{th}, 1^{st}, 0.1^{th}$) from this data as possible choices for the lower cutoff value for cosine. After some experimentation with a few kinked and straight helices data, we chose the 0.1^{th} percentile of 0.9818 as the lower cut-off value, which corresponds to an

angle threshold of approximately $T = 11$ degrees. A typical example of our experimentation is described in Web Appendix H, in particular see Web Figure 6. The chosen cutoff suggests that the probability of incorrectly classifying a straight helix (Type I error) is at most 0.1%. We analyzed the classification accuracy on crowdsourced straight and kinked helices. This is discussed in Section 6.

Critical-deviation-run-threshold ($r = 4$) Experimentation and contextual considerations: The presence of the kink at a position not only causes the $\cos\theta$ at that position to significantly deviate from 1, but also has a similar effect on the $\cos\theta$ computations at the neighboring positions near the kink. For example, if there is a kink at position j, the estimated axis obtained with C_α atoms $\{j-3, j-2, j-1, j, j+1, j+2\}$ also is influenced by the presence of the kink and is not going to be aligned with the axis computed from C_α atoms $\{j+3, j+4, j+5, j+6, j+7, j+8\}$. As a result the $\cos\theta$ at position $j+2$, computed based on the two estimated axes, is also going to deviate from 1. In effect, when a kink is present, we would see a cluster of cosine values that significantly deviates from 1, with the maximum dip in cosine value (or maximum peak in angle) happening close to the kink position. This is illustrated with a typical example in Web Appendix H (see Web Table 6 part (b) and Web Figure 6). To see a clear dip in cosine values, one needs to check at least 3 consecutive deviations. Based on our experimentation with a few kinked and straight helices, we determined $r = 4$ as a reasonable choice.

5.2 *Sensitivity analysis of choice of tuning parameters*

To check the sensitivity of Kink-Detector to the choice of tuning parameters,we first record the accuracy of classification of crowdsourced 129 straight and 64 kinked helices, based on the chosen values of the tuning parameters viz. $(k = 6, T = 11, r = 4)$, and then study the change in accuracy if we change the tuning parameters away from the chosen values. In particular, we study the impact on classification accuracy, when we increase the atom-window size from

the minimum value of $k = 6$ to $k = 7$, increase (decrease) the angle-threshold from $T = 11$ to $T = 12$ ($T = 10$) degrees, and increasing (decreasing) the critical-deviation-run-threshold from $r = 4$ to $r = 5$ ($r = 3$). While changing each tuning parameter, we keep the other two parameters fixed at the chosen values. The detailed results of this analysis are described in Web Appendix I and Web Table 7. To sum up, our choice of $(k = 6, T = 11, r = 4)$ is driven by contextual considerations and experimentation. The sensitivity analysis on the crowdsourced data further confirms that these choices are indeed reasonable as any change leads to significant deterioration in the classification accuracy either for kinked or for straight helices.

6. Testing of Kink-Detector on the Crowdsourced Data

In this section, we test the performance of Kink-Detector on the crowdsourced data. Table 3 shows the summary of results from the testing. Of the 129 crowdsourced straight helices, Kink-Detector classified 124 helices as "straight". Of the 64 crowdsourced kinked helices, Kink-Detector classified 62 as "kinked". The overall performance of Kink-Detector is satisfactory as only 7 helices are misclassified.

The performance compares well with its competitor Kink-Finder (details are given below). The detailed output of Kink-Detector including the kink position and angle for the 64 kinked helices of crowdsourced data is provided in Web Table 8 of Web Appendix J. We next analyze the misclassified cases in detail.

Misclassified Helices: Seven straight and kinked helices were misclassified by Kink-Detector, namely, Kinked Helices 1-2, and Straight Helices 3-7. We have further examined these helices, and it seems that these misclassifications could be a result of some limitation of visual perception. Table 1 (see cases 1 to 7) provides a brief description of these helices.

We first discuss Kinked Helices 1 and 2, which were classified as straight by Kink-Detector. Web Figure 8 in Web Appendix J shows a plot of the 2 kinked helices. On visual inspection, it

may appear that the kink is present near to one of the extreme positions for both helices (near position 4 for Helix 1 and near position 3 for Helix 2). However, we note that visual inspection could be misleading, for the visual appearance of a kink can be considerably weakened by a perturbation of just a single atom position. In Web Figure 8 of Web Appendix J, we also illustrate how a perturbation of a single atom position (position 4 in Helix 1 and position 3 in Helix 2) weakens the appearance of a kink. The points are further elaborated in Web Appendix J.

Next, we discuss Straight Helices 3, 4, 5, 6 and 7, which are classified as kinked by Kink-Detector. For these five helices, Table 4 (parts (c) to (g)) shows the cosine angles computed between axes based on the atom-window size of 6. Web Figure 9 in Web Appendix K shows a plot of the 5 helices. The change in axis direction near the kink is also shown in Web Figure 9. A visual inspection after marking the kink position and axis change based on Kink-Detector suggests that the method seems to identify the position of the kinks correctly (Web Appendix H: Web Figure 6 provides another example where the method identifies kinks correctly). It may therefore be argued that these five cases (i.e. Helices 3, 4, 5, 6 and 7) have kinks that are perhaps not prominent enough to be easily detected by visual inspection. The kink positions as detected by Kink-Detector for helices 3, 4, 5, 6 and 7 are near positions 13, 7, 7, 10 and 13 respectively. The corresponding angles (in degrees) are 18.69, 15.87, 22.82, 20.39 and 20.04 respectively. So, the angles of the kinks detected in all these four cases happen to be not as large. Visual identification of a kink is subject to some limitation when the angle change is small. Web Table 8 in Web Appendix J, which contains all the 64 helices identified as kinked in crowdsourcing, suggests that kinks are visually identified when the angle is above 23.66 degrees (the minimum shown in the table). The points are further expanded in Web Appendix K. The most important message here is that there is some limitation to what we can see visually, therefore we can miss out moderately-kinked helices.

[Table 3 about here.]

Comparison with Kink-Finder: We now compare the performance of Kink-Detector to Kink-Finder (Wilman, Shi, and Deane, 2014), which is the most competitive algorithm for finding Kinks as pointed out in Section 1, and further details on Kink-Finder are given Web Appendix L. Kink-Finder misclassified 3 of the 129 straight helices from crowdsourced data as kinked and 2 of the 64 kinked helices from crowdsourced data as straight. The set of misclassified helices is not exactly the same for Kink-Detector and Kink-Finder. Helix 5 is misclassified as kinked by both Kink-Finder and Kink-Detector. Helix 1 is misclassified as straight by both Kink-Finder and Kink-Detector. We conclude that the accuracy of statistical model based Kink-Detector is comparable to that of computational approach based Kink-Finder.

We note that the higher accuracy of Kink-Finder in classifying straight helices (126 out of 129) is perhaps at the cost of over-classifying helices with moderate changes in axial directions as straight. We noted earlier in this section that the straight Helices 3-7 that were classified as kinked by Kink-Detector had been estimated with an angle close to 20 degrees. A further support to this observation is obtained when we study the classification accuracy of Kink-Finder and Kink-Detector on curved helices. As noted in the introduction of this paper, curved helices have a slow but steady change in axial direction; so they are not straight. We expressly clarify that neither Kink-Detector nor Kink-Finder are designed to differentiate curved from kinked helices. However, the methods can at least be used to check whether curved helices are correctly classified as "not straight". In addition to 129 straight and 64 kinked helices, the crowdsourced data includes 59 curved helices. Again, from Table 3, we note that Kink-Detector provides good power in classifying curved helices when compared to Kink-Finder. Namely, Kink-Detector classifies 71% as not straight while Kink-Finder classifies only 51% of the 59 curved helices as not straight. Web Table 9 in Web Appendix L summarizes these findings in the form of confusion matrices and some standard

classification measures (viz. sensitivity, specificity, accuracy and precision). In Web Appendix L, we provide several differences between Kink-Finder and Kink-Detector methods, some of which help explain the above results. In summary, Kink-Detector is the first approach based on a statistical model, that can provide further insights into helical structures and perhaps is more guarded against classifying helices with moderate angular changes as straight. Kink-Finder on the other hand is a purely computational approach, and some features of the method can lead to over-classification of helices with moderate change in axial directions, into straight helices.

7. Discussion

In this paper, we have developed a kink detection algorithm "Kink-Detector" based on a plausible statistical model for a straight $\alpha-$helix. Our estimation of the model on crowdsourced straight $\alpha-$helices confirms previous empirical findings on the radius and pitch of $\alpha-$helices. Further, for the first time, it provides a quantification of variability in atom positions around the cylinder. The effectiveness of Kink-Detector in detecting the presence or absence of kinks has been demonstrated using the "gold standard" crowdsourced data on straight and kinked helices. In particular, we indicate how the visual identification of kinks in helices can have some limitations. Furthermore, the performance of statistical model based Kink-Detector is comparable to its computational competitor Kink-Finder. While our discussion is only for $\alpha-$helix, this work should be applicable to other types of helices. A natural future extension of our work will be to explore a direct statistical model formulation for kinked and curved helices. A potential approach is to extend the exploratory work of Mardia et al. (1999) on torsion and curvature of curves including helices.

8. Supplementary Materials

Web Appendices, Tables and Figures referenced in Sections 2, 3, 4, 5.1, 5.2 and 6 are available with this article at the Biometrics website in Wiley online library. The R codes implementing the proposed approach are available with this paper at the Biometrics website in Wiley Online Library.

Acknowledgements

We wish to express our thanks to Jeanine Houwing-Duistermaat for very helpful comments and to Eleanor Law for helping us with the crowdsourced data and the Kink-Finder software. The first author also wants to thank the Indian Institute of Management Ahmedabad for hospitality during his visits to carry out some of this work. The authors also thank the anonymous reviewers and the associate editor for several useful comments which improved this manuscript.

References

Bansal, M., Kumar, S., and Velavan, R. (2000). Helanal: A program to characterize helix geometry in proteins. Journal of Biomolecular Structure and Dynamics **17,** 811–819.

Barlow, D. J. and Thornton, J. M. (1988). Helix geometry in proteins. Journal of Molecular Biology **201,** 601–619.

Blundell, T., Barlow, D., Borkakoti, N., and Thornton, J. (1983). Solvent-induced distortions and the curvature of $\alpha-$ helices. Nature **306,** 281–283.

Dickerson, R. E. and Geis, I. (1969). The Structure and Action of Proteins. W.A. Benjamin, California.

Dryden, I. L. and Mardia, K. V. (2016). Statistical Shape Analysis, with Applications in R. Wiley, Chichester.

Jung, S., Foskey, M., and Marron, J. S. (2011). Principal arc analysis on direct product manifolds. The Annals of Applied Statistics **5,** 578–603.

Kink-Finder (2014). Software is accessible at `http://www.mybiosoftware.com/kink-finder-1-01-find-kinks-in-helices.html`.

Kneissl, B., Mueller, S. C., Tautermann, C. S., and Hildebrandt, A. (2011). String kernels and high-quality data set for improved prediction of kinked helices in $\alpha-$helical membrane proteins. Journal of Chemical Information and Modeling **51,** 3017–3025.

Kumar, P. and Bansal, M. (2012). Helanal-plus: a web server for analysis of helix geometry in protein structures. Journal of Biomolecular Structure and Dynamics **30,** 773–783.

Langelaan, D. N., Wieczorek, M., Blouin, C., and Rainey, J. K. (2010). Improved helix and kink characterization in membrane proteins allows evaluation of kink sequence predictors. Journal of Chemical Information and Modeling **50,** 2213–2220.

Mardia, K. V. (2013). Statistical approaches to three key challenges in protein structural bioinformatics. Journal of Royal Statistical Society: Series C **62,** 487–514.

Mardia, K. V. and Gadsden, R. J. (1977). A small circle of best-fit for spherical data and areas of vulcanism. Journal of Royal Statistical Society: Series C **26,** 238–245.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). Multivariate Analysis. Wiley, Academic Press.

Mardia, K. V., Morris, R. J., Walder, A. N., and Koenderink, J. J. (1999). Estimation of torsion. Journal of Applied Statistics **26,** 373–381.

Meruelo, A. D., Samish, I., and Bowie, J. U. (2011). Tmkink: A method to predict transmembrane helix kinks. Protein Science **20,** 1256–1264.

Rueda, C., Fernández, M. A., Barragán, S., Mardia, K. V., and Peddada, S. D. (2016). Circular piecewise regression with applications to cell-cycle data. Biometrics **72,** 1266–1274.

Sansom, M. S. and Weinstein, H. (2000). Hinges, swivels and switches: the role of prolines in signalling via transmembrane alpha-helices. Trends in Pharmacological Sciences **21,** 445–451.

Siegmund, D., Yakir, B., and Zhang, N. R. (2011). Detecting simultaneous variant intervals in aligned sequences. The Annals of Applied Statistics **5,** 645–668.

Srivastava, M. S. and Worsley, K. J. (1986). Likelihood ratio tests for a change in the multivariate normal mean. Journal of American Statistical Association **81,** 199–204.

Visiers, I., Braunheim, B. B., and Weinstein, H. (2000). Prokink: A protocol for numerical evaluation of helix distortions by proline. Protein Engineeriing, Design and Selection **13,** 603–606.

Wilman, H. R. (2014). Computational studies of protein helix kinks. DPhil. University of Oxford. Thesis is accessible at `http://ora.ox.ac.uk/objects/uuid:21225f0e-efed-49c6-af27-5d3fe78fa731`.

Wilman, H. R., Ebejer, J. P., Shi, J., Deane, C. M., and Knapp, B. (2014). Crowdsourcing yields a new standard for kinks in protein helices. Journal of Chemical Information and Modeling **54,** 2585–2593.

Wilman, H. R., Shi, J., and Deane, C. M. (2014). Helix kinks are equally prevalent in soluble and membrane proteins. Proteins: Structure, Function and Bioinformatics **82,** 1960–1970.

Yohannan, S., Faham, S., Yang, D., Whitelegge, J. P., and Bowie, J. U. (2004). The evolution of transmembrane helix kinks and the structural diversity of g protein-coupled receptors. Proceedings of the National Academy of Sciences of the United States of America **101,** 959–963.

Received May 2016. *Revised NA . Accepted NA.*

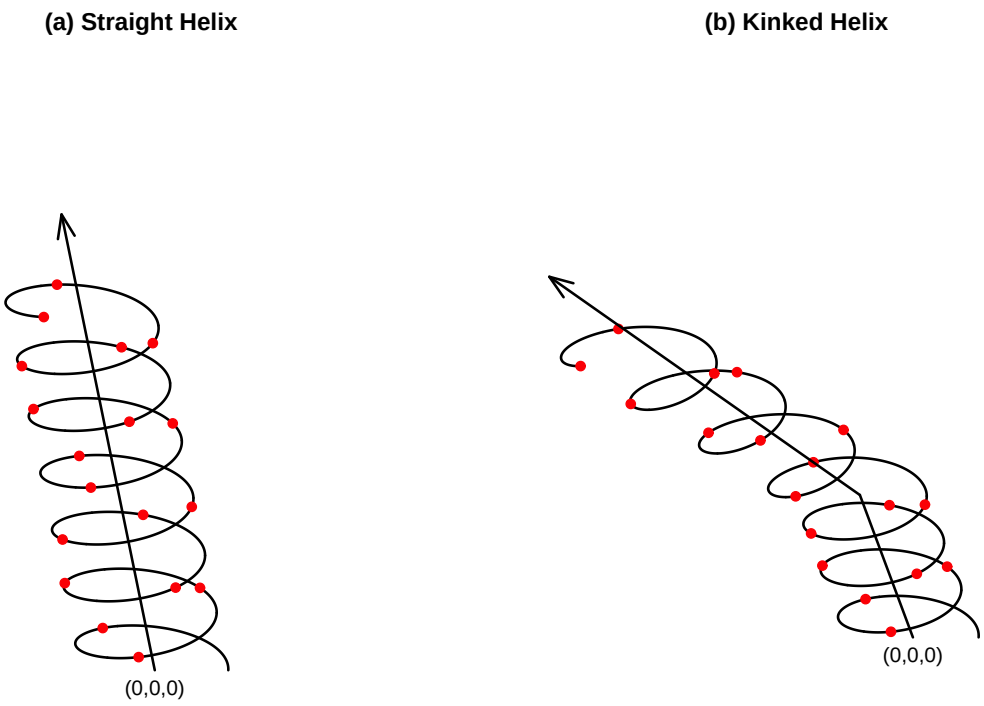

Figure 1. A diagram showing (a) a straight helix and (b) a kinked helix with their atoms (shown as dots) and axis/axes. This figure appears in color in the electronic version of this article.

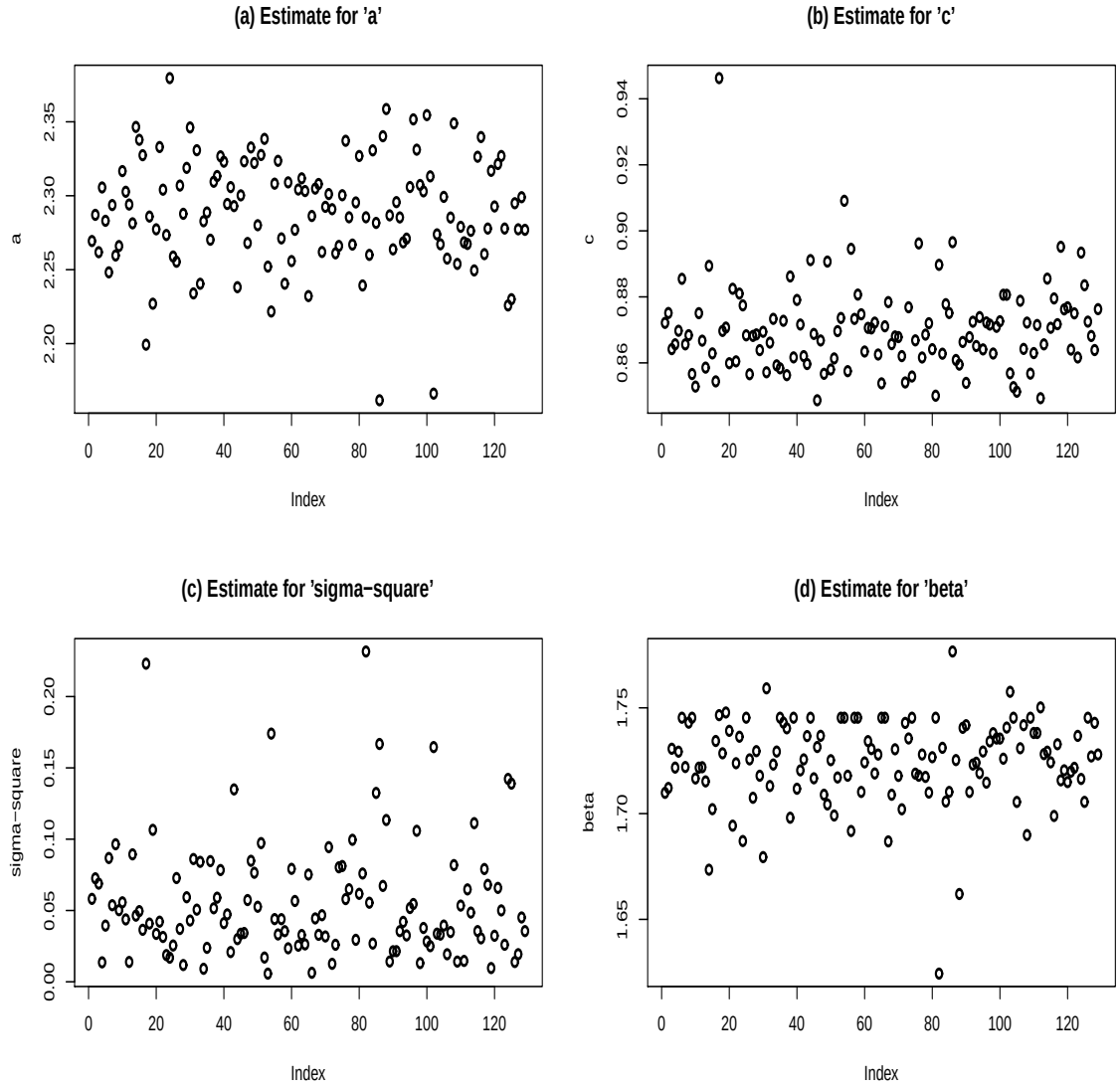

Figure 2. Parts (a), (b), (c) and (d) plot the robust MLE for a, c, σ^2 and β based on 129 straight helices of the crowdsourced data. In comparison to Web Figure 5, all the estimates are well behaved and there are no clustered outliers. Note that this figure uses a different scale on the y-axes compared to Web Figure 5.

Table 1

Summary of different helix examples discussed in the paper and the web supplement. Helices 1 to 7 are cases of misclassification by Kink-Detector while classifying crowdsourced helices into "kinked" versus "straight". Helices 8 and 9 are used to illustrate the estimation procedure and its application. Section 6 in the paper and Web Appendices J-K deal with Helices 1-7, and Web Appendices H deals with Helices 8 and 9.

Case	Reference	PDB name (chain)	Atoms	Length	Context in which discussed
Helix 1	Section 6, Table 4 parts (a)-(b), Web Appendix J: Web Figure 8.	1v54_B (16-46)	1 to 31	31	These two helices were identified as kinked helices in crowdsourced data but were classified as straight by Kink-Detector. We observe that the visual appearance of kink for these helices is weakened just by perturbation of a single atom position (see Web Appendix J: Web Figure 8). Thus, Kink-Detector is cautious in classifying them as kinked.
Helix 2		2gfp_A(322-345)	1 to 24	24	
Helix 3	Section 6, Table 4 parts (c)-(g), Web Appendix K: Web Figure 9.	1c3w_A (38-61)	1 to 24	24	These 5 helices were identified as straight helices in crowdsourced data, but were identified as kinked by Kink-Detector. It may be argued that some of these cases actually have kinks but that these are perhaps not prominent enough to be easily detected by visual inspection (see Table 4 , Web Appendix K: Web Figure 9).
Helix 4		1rc2_A(37-53)	1 to 17	17	
Helix 5		1wpg_A(965 to 988)	1 to 24	24	
Helix 6		3mp7_A(148 to 170)	1 to 23	23	
Helix 7		2fyn_A(362 to 380)	1 to 19	19	
Helix 8	Web Appendix D : Web Figure 4, Web Table 2, Web Appendix H: Web Table 6 part (a)	3bz1_B(203-217)	1 to 15	15	A straight protein helix in crowdsourced data selected to demonstrate estimation of model parameters and axis. Web Table 6 part (a) shows the values of cosine between successive axes based on 6-atom moving window.
Helix 9	Web Appendix H: Web Figures 6 and 7, Web Table 6- part (b)	2a65_A(44-70)	1 to 27	27	A kinked protein helix in crowdsourced data selected to highlight kinks along with changing axis direction. Due to the helical structure, the kink change point is not a specific point but is a change in the direction of the helix axis, which depends on multiple points (see Web Appendix H: Web Figure 6 for an illustration). Web Table 6 part (b) shows the values of cosine between successive axes based on a moving 6-atom window. Our method looks for four consecutive critical deviations to detect a kink.

Table 2

Mean and standard error (SE) for the robust MLE of the parameters (a, c, β, σ^2) obtained across 129 straight helices of the crowdsourced data. Also shown are the ideal values of the parameters and our final choice of the parameter values of the null model.

	a	c	β	σ^2
Mean	2.29	0.87	1.72	0.056
SE	0.04	0.01	0.02	0.041
Ideal value	2.3	0.86	1.75	Unknown
Null Model	2.3	0.86	1.75	0.056

Table 3

Comparison of classification of crowdsourced straight, kinked and curved helices by Kink-Detector and Kink-Finder.

		Kink-Detector		Kink-Finder	
Crowdsourcing	Total	Straight	Kinked	Straight	Kinked
Straight	129	124	5	126	3
Kinked	64	2	62	2	62
Curved	59	17	42	28	31

Table 4

Cosine values between successive axes based on 6-atom windows for crowdsourced helices misclassified by Kink-Detector. Helices 1 and 2 are identified as kinked in crowdsourcing but classified as straight by Kink-Detector. Helices 3 to 7 are straight in crowdsourcing but classified as kinked by Kink-Detector. The cosine values that fall below the cutoff value 0.9818 (i.e. angle > 11 degrees) are highlighted in **bold** *and the value at the kink position is underlined. For Kink-Detector, at least 4 consecutive values in* **bold** *indicate the presence of a kink, and the kink position is taken to be the one with minimum cosine value. Web Figures 8 and 9 in the Web Appendices J and K respectively, show plots of these helices.*

(a) Helix 1

C_α position	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
cos angle	**0.97**	**0.96**	**0.98**	1.00	**0.97**	**0.97**	**0.97**	1.00	**0.97**	1.00	1.00	1.00	0.99	1.00	**0.97**	1.00

(b) Helix 2

C_α position	6	7	8	9	10	11	12	13	14	15	16	17	18
cos angle	**0.90**	**0.96**	1.00	0.99	1.00	1.00	**0.98**	**0.98**	0.99	**0.97**	0.99	0.99	0.99

(c) Helix 3

C_α position	6	7	8	9	10	11	12	13	14	15	16	17	18
cos angle	**0.95**	**0.97**	**0.98**	0.99	1.00	**0.97**	**0.97**	**0.95**	**0.97**	**0.96**	1.00	**0.98**	1.00

(d) Helix 4

C_α position	6	7	8	9	10	11
cos angle	**0.98**	**0.96**	**0.98**	**0.97**	**0.97**	0.99

(e) Helix 5

C_α position	6	7	8	9	10	11	12	13	14	15	16	17	18
cos angle	**0.98**	**0.92**	**0.94**	**0.92**	0.99	0.99	0.99	**0.94**	**0.93**	0.99	0.98	0.99	1.00

(f) Helix 6

C_α position	6	7	8	9	10	11	12	13	14	15	16	17
cos	1.00	0.99	**0.98**	**0.97**	**0.94**	**0.98**	1.00	1.00	1.00	0.99	0.99	0.99

(g) Helix 7

C_α position	6	7	8	9	10	11	12	13
cos	1.00	0.99	1.00	**0.97**	**0.95**	**0.97**	**0.97**	**0.94**