# How new shape analysis and directional statistics are advancing modern life-sciences

Professor Kanti V. Mardia
*Senior Research Professor,*
*University of Leeds*
*Department of Statistics*
*Leeds LS2 9JT,UK*
*E-mail: k.v.mardia@leeds.ac.uk*

## 1 Introduction

If the last century in Science belongs to Physical Sciences then this century must belong to Life Sciences with many breakthroughs starting from DNA and proteins! The proteins are biological macromolecules that are of primary importance to all living organisms and there are various open problems including the Nobel-Prize-type problem related to protein folding. Some of these questions mainly depend on the 3-dimensional shape of the protein, which can be summarized in terms of either the configuration of points (landmarks) or more compactly by some dihedral angles (conformational angles). A similar comment applies to RNA. We need some new appropriate tools in Statistical Shape Analysis and Directional Statistics to answer various statistical problems for such data. In Statistical Shape Analysis, the aim could be to align different proteins using unlabelled landmarks so as to understand the function of proteins. In this case, new statistical alignment methods have appeared but how do these methods compete with deterministic methods which are faster? Does a hybrid approach provide a solution efficiently? In Directional Statistics, we need plausible multivariate circular distributions to model such angular data. The problem is not as simple as the distributions have unknown finite support in many dimensions. Also, there are clusters so mixture models are appropriate. Further, the data in general are along a sequence so time series models are appropriate leading to hidden Markov model with circular distributions as nodes. The statistical distributions are now coming up but there are serious challenges, for example, what directional distributions are workable? In this paper,we will mainly review new tools available so far in the two statistical areas; these tools have a potential use in protein design, protein folding and drug discovery.

## 2 Shape Analysis

### 2.1 Background

Statistical shape analysis is concerned with extracting the shape information contained in a random sample of physical objects, where shape is defined to be all the geometrical information about an object that is invariant under a particular transformation of interest. Therefore, an important stage in the comparison of the shapes of two or more objects is to align them in some optimal sense, under some geometric transformation, so that the information which remains is the shape information of interest. Here we will deal with **size-and-shape, or Form,** where only rotation and translation (rigid body transformations) are filtered out. This is of interest here as the objects are molecules, where the bond lengths between atoms should be preserved, so only the rigid body transformations are allowed.

It is common to represent the shape of an object as a configuration of points, where the points represent the locations of landmarks chosen to represent the object. Landmarks in "labelled" shape analysis are uniquely defined and "homologous" for similar objects. An important problem is that of unlabelled shape analysis, where the correspondence between landmarks on different objects is

not known. The problem is then to simultaneously identify the matches between landmarks and the transformation between the configurations in some optimal way. Figure 1 gives an example of three unlabelled points which are aligned to a configuration with four unlabelled points under a rigid transformation. The same principle applies to configurations in 3 dimensions such as the atomic coordinates of proteins.

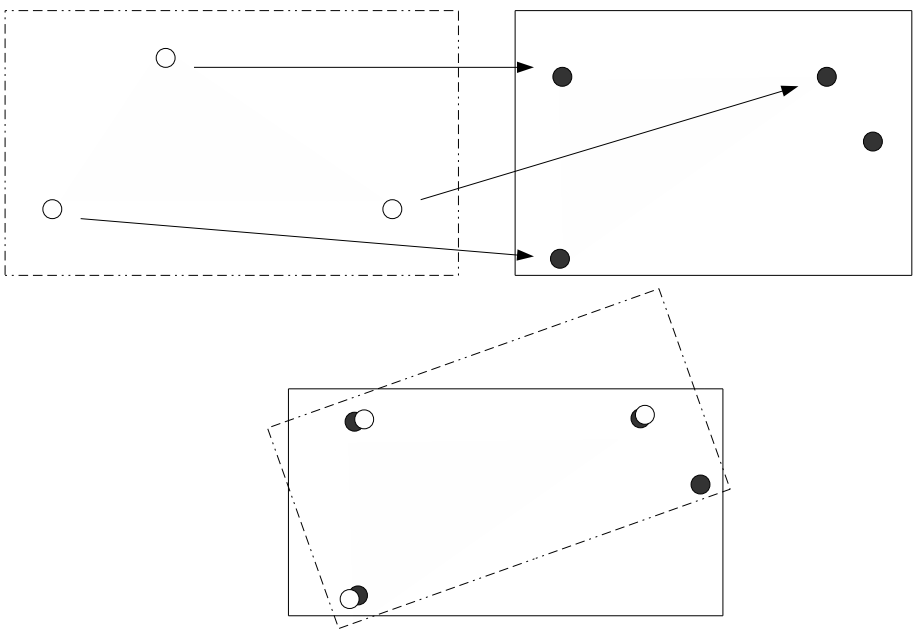

Figure 1: Top: Two unlabelled configurations of points. Bottom: Alignment under a rigid transformation (Figure by Chris Fallaize).

## 2.2 Bioinformatics

Biological macromolecules, in particular proteins and nucleic acids, need specific structure in order to perform their specific functions (see, for examples, Branden and Tooze,1998). Changes to structure usually disrupt or change the way the macromolecules function. Also, homologous proteins are thought to have evolved from a common ancestor. This information can be extracted ,for example, by aligning proteins using their 3–dimensional structures. With the increasing number of protein structures becoming available and deposited in databases such as the Protein Data Bank, reliable methods for protein structure alignment are becoming very important. Many deterministic methods for alignment have been developed, such as DALI, CE, LGA, SSAP, and others. These methods are based on computational algorithms designed to find an optimal alignment in some sense, and do not give any indication of uncertainty in this optimum; for instance there may be high uncertainty in some areas of the alignment, and other areas where the alignment between the two structures is very good. Therefore, there is a need for probabilistic methods which allow sampling over the space of all possible alignments, allowing a statistical interpretation of the alignment uncertainty to be made. Here, we describe a fully Bayesian alignment model, which treats the registration parameters as unknowns about which to draw inference, allowing quantification of their uncertainty as well as the matches simultaneously.

### 2.3 The BALI Pairwise Alignment Algorithm

Green and Mardia (2006) have proposed the pairwise alignment of two configurations using a Bayesian hierarchical model. We call this BALI (Bayesian ALIgnment), following the well established DALI (Distance alignment) algorithm . Consider aligning a pair of configurations $x$ and $y$ in 3 dimensions under rigid body transformations. Denote the $j^{\text{th}}$ atom in the x configuration by $x_j, j = 1, \ldots, m$ and the $k^{\text{th}}$ atom in the y configuration by $y_k, k = 1, \ldots, n$. Let $A$ and $\tau$ denote the rotation matrix and translation vector to bring $y$ into alignment with $x$, with prior distributions $p(A)$ and $p(\tau)$. We denote the prior for $\sigma$, parameterising Gaussian noise in atomic positions for coordinates of $x_j, y_k$, by $p(\sigma)$. The joint posterior for the model is

$$(1) \qquad p(M, A, \tau, \sigma, x, y) \propto p(A)p(\tau)p(\sigma) \times \prod_{j,k:M_{jk}=1} \left[ \kappa \frac{\phi\{(x_j - Ay_k - \tau)/\sigma\sqrt{2}\}}{(\sigma\sqrt{2})^3} \right],$$

where $\phi(\cdot)$ is the standard normal probability density function and $\kappa > 0$ is a parameter representing the propensity of points to be matched. $M$ is an unknown matrix for matching between points on each configuration; here $M_{jk} = 1$ if $x_j$ is matched to $y_k$ , and $M_{jk} = 0$ otherwise. The priors recommended are

$$(2) \qquad A \sim \text{uniform}, \quad \tau \sim N(\mu_\tau, \sigma_\tau^2 I_3), \quad \sigma^{-2} \sim \Gamma(\alpha, \beta),$$

where $\sigma_\tau^2$ is large. A point estimate $\hat{M}$ of $M$ is found by minimising the error rates $P(\hat{M}_{jk} = 1|M_{jk} = 0)$ and $P(\hat{M}_{jk} = 0|M_{jk} = 1)$ and is controlled by the cost ratio, $K$, of falsely matching points. The posterior probability that the pair of points $(x_j, y_k)$ are a match, $p_{jk} = P(M_{jk} = 1|x, y)$, is given by the empirical frequency of this match from an MCMC run and $\hat{M}$ is a solution to a "linear assignment" problem with cost matrix $(p_{jk} - K)$. A standard linear assignment program is then used to find $\hat{M}$, with the cost matrix $(p_{jk} - K)_+$. All details including the MCMC implementation and protein examples are given in Green and Mardia (2006). A direct extension to multiple assignment is given in Ruffieux and Green (2009) and an alternative faster approach based on pairwise BALI is provided by Mardia et al (2010).

Green et al (2010) have given a review of various methods . The methods broadly speaking fall into two categories ;one treating x and y symmetrically as here, the other regressing x on y as in Dryden et al. (2007), Schmidler (2007) and Kent et al. (2010) . As suggested by Dryden (2007), approaches for dealing with the nuisance (transformation) parameters fall into two general classes: marginalisation and maximisation methods; the BALI falls into the former. In the maximisation approach, as in, for example, the approaches of Dryden et al. (2007) and Schmidler (2007), the joint distribution is maximised over the nuisance parameters (via a Procrustes alignment using the matched points), and inference for M is then performed conditionally. From a non-Bayesian perspective, Kent et al.(2010) have developed the EM algorithm, in which the likelihood is maximised over the transformation parameters given expected values of the unknown matching labels (M–Step), and the expectation of the labels is then taken with respect to the resulting maxima of the parameters (E–Step). That is,the labels are treated as missing data (and a coffin bin is introduced for unmatched landmarks ). More work is needed to compare the performance of these approaches in practical situations.

## 3 Directional Statistics

### 3.1 Bio-molecular Structure Validation

Validating the quality of new 3 dimensional structures of proteins and RNAs obtained from X-ray crystallography is one of the important issues in biosciences (see, for example, Lovell et al ,2003). In protein geometry, the basic shape of a protein is formed by the backbone while the side-chains are

pointing away from the backbone. Their shape can be summarized by some dihedral angles needing multivariate circular statistics. Mardia et al (2007) have used mixtures of bivariate circular distributions for modeling the dihedral angles on the backbone of proteins. However, since the dihedral angles on both the backbone and side-chain vary a lot, all of these angles should be considered for assessing the 3D structure of protein and RNA. We need to understand correlations and patterns in multivariate angular data. Various models have been proposed as extensions of the univariate von Mises in Mardia and Patrangenaru (2005) and Mardia et al (2008) but these have intractable normalizing constants. In general, these data come from mixtures of distributions so in implementation, EM and Dirichlet process mixtures, for example, are relevant.

### 3.2 Directional Models

The von Mises distribution(see, for, example,Mardia and Jupp,2000) is the basic distribution for the angular data which has the probability density function

$$f(\theta;\mu,\kappa) = \{2\pi I_0(\kappa)\}^{-1}\exp\{\kappa\cos(\theta-\mu)\}, \theta\in(0,2\pi],\ \kappa\geq 0,$$

and $\mu\in(0,2\pi]$ is the mean direction, $\kappa$ is the concentration parameter and $I_0(\kappa)$ is the modified Bessel function of the first kind and order zero. Mardia(1975) proposed the following family of bivariate von Mises distributions with the probability density function

$$\begin{aligned} f(\theta,\phi) &= c(\kappa_1,\kappa_2,A)\exp\{\kappa_1\cos(\theta-\mu)+\kappa_2\cos(\phi-\nu)+ \\ &\quad [\cos(\theta-\mu),\sin(\phi-\mu)]A[\cos(\theta-\nu),\sin(\phi-\nu)]^T\}, \end{aligned} \tag{3}$$

where the angles $\theta,\phi,\mu,\nu\in(-\pi,\pi]$ lie on torus, and A is a $2\times 2$ matrix.

This model has eight parameters and allows for varied dependence between the two angles. Various important submodels with five parameters have appeared: the sine model (Singh et al,2002) and the cosine model (Mardia et al ,2007) have been proposed to mimic the bivariate normal distribution. The *Sine model* has the density

$$f(\theta,\phi)\propto\exp\{\kappa_1\cos(\theta-\mu)+\kappa_2\cos(\phi-\nu)+\lambda\sin(\theta-\mu)\sin(\phi-\nu)\}. \tag{4}$$

Figure 2 shows a plot of a mixture of three sine distributions (with various parameter values ) on the torus ; the key point is that the model has diffused to concentrated shapes. Kent et al (2008) have given an overview of various bivariate circular submodels.

In protein bioinformatics, there has been a growing keen interest in bivariate von Mises directional distributions. For example, these distributions are used in various fundamental applications arising from the conformational angles of protein backbones (see Boomsma et al, 2008; Mardia et al, 2008; Mardia, 2010; Lennox et al 2009,2010).

A multivariate extension of the distribution relevant for variables $\theta_1,...,\theta_p$ can be written down with density proportional to

$$\exp\{\Sigma a_s\cos\theta_s+\Sigma b_s\sin\theta_s+\Sigma a_{st}\cos\theta_s\cos\theta_t+\Sigma b_{st}\cos\theta_s\sin\theta_t+\Sigma c_{st}\sin\theta_s\sin\theta_t\} \tag{5}$$

where $a_{ss}=b_{ss}=c_{ss}=0$, and $b_{st}\neq b_{ts}$. A particular case of importance is von Mises sine distribution, with probability density function (Mardia et al, 2008):

$$f(\theta;\mu,\kappa,\Lambda) = C_p^{-1}(\kappa,\Lambda)\exp\{\kappa^T c(\theta,\mu)-\frac{1}{2}s(\theta,\mu)^T\Lambda s(\theta,\mu)\}, \tag{6}$$

where $-\pi<\theta_j\leq\pi$, $-\pi<\mu_j\leq\pi$, $\kappa_j\geq 0$, $-\infty<\lambda_{jl}<\infty$,

$$\begin{aligned} c(\theta,\mu) &= (\cos(\theta_1-\mu_1),\cos(\theta_2-\mu_2),\ldots,\cos(\theta_p-\mu_p)), \\ s(\theta,\mu) &= (\sin(\theta_1-\mu_1),\sin(\theta_2-\mu_2),\ldots,\sin(\theta_p-\mu_p)), \\ \mu^T &= (\mu_1,\mu_2,\ldots,\mu_p),\quad \kappa^T=(\kappa_1,\kappa_2,\ldots,\kappa_p), \\ (\Lambda)_{jl} &= \lambda_{jl}=\lambda_{lj},\ \ j\neq l,\ \ (\Lambda)_{jj}=\lambda_{jj}=0,\ \ j,l=1,\ldots,p \end{aligned}$$

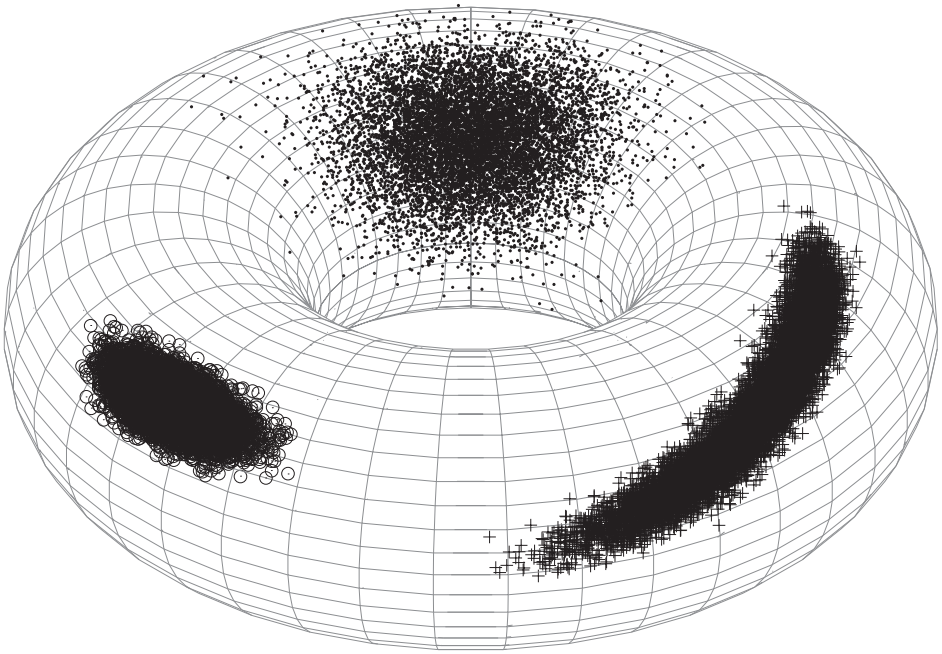

Figure 2: A mixture of three bivariate sine distributions showing a wide variety of correlation structure (Figure by Jes Frellsen ).

and $C_p^{-1}(\kappa, \Lambda)$ is a normalizing constant. Note that for $p = 1$, it becomes a univariate von Mises density, and for $p = 2$, it is the bivariate sine distribution . This distribution has a great potential for exploring RNA structure ( see, Frellsen et al, 2009). It should be noted that the distributions as those in Figure 2 give rise to what are called Ramachandran plots (Ramachandran et al 1963), which represent the distribution of the two dihedral angles of a protein backbone; the distributions are mixtures with finite unknown support as there are area which are forbidden by the underlying chemistry. We conclude with the question which multivariate circular distributions help in understanding protein and RNA data; the subject is still in its infancy but some real progress has been made.

## 4 Acknowledgments

I wish to express my thanks to Chris Fallaize, Jes Frellsen and Zhengzheng Zhang for their help.

## References

[1] Boomsma, W., Mardia, K.V., Taylor, C.C., Ferkinghoff-Borg, J., Krogh, A. and Hamelryck, T. (2008). A generative, probabilistic model of local protein structure, *PNAS*, **105**, pp.8932–8937.

[2] Branden, C. and Tooze, J.(1998) *Introduction to Protein Structure*, Garland, New York.

[3] Dryden, I. L. (2007). Discussion to Schmidler (2007).

[4] Dryden, I. L., Hirst, J. D. and Melville, J. L. (2007). Statistical analysis of unlabeled point sets: comparing molecules in chemoinformatics, *Biometrics*, **63**, pp.237-251.

[5] Frellsen, J.A., Moltke, I., Thiim. M., Mardia, K.V., Ferkinghoff-Borg, J. and Hamelryck, T. (2009). A probabilistic model of local RNA 3-D structure. *PLoS Comput. Biol.*, **5**, pp.1–11.

[6] Green, P.J. and Mardia, K.V. (2006) Bayesian Alignment Using Hierarchical Models, with Applications in Protein Bioinformatics. *Biometrika*, **93**, pp. 235-254.

[7] Green, P. J., Mardia, K. V., Nyirongo, V. B. and Ruffieux, Y. (2010). Bayesian modelling for matching and alignment of biomolecules, in *A. OHagan and M. West (eds), The Oxford Handbook of Applied Bayesian Analysis*, Oxford University Press, Oxford,pp. 27-50.

[8] Kenobi, K. and Dryden, I. L. (2010). Bayesian matching of unlabelled point sets using Procrustes and configuration models. Available at http://arxiv.org/pdf/1009.3072.

[9] Kent, J. T., Mardia, K. V. and Taylor, C. C. (2004). Matching problems for unlabelled configurations, in *Proceedings of the Leeds Annual Statistical Research Conference, Bioinformatics, Images, and Wavelets*, Leeds University Press, Leeds,pp.33-36.

[10] Kent, J.T., Mardia, K.V. and Taylor, C.C. (2008). Modelling strategies for bivariate circular data, in *Proceedings of the Leeds Annual Statistical Research Conference, The Art and Science of Statistical Bioinformatics*, Leeds University Press, Leeds, pp.70-73.

[11] Kent, J.T., Mardia, K.V. and Taylor, C.C. (2010). Matching unlabelled configurations and protein bioinformatics, Technical report, University of Leeds. URL:http://www.maths.leeds.ac.uk/statistics/research/reports/2010.html # STAT10–01

[12] Lennox, K.P., Dahl, D.B., Vannucci, D.B. and Tsai, J.W. (2009) Density estimation for protein conformation angles using a bivariate von Mises distribution and Bayesian nonparametrics. *J. Amer. Statist. Soc.*, **104**, pp.586-596. Correction *J. Amer. Statist. Soc.*, **104**, p.1728.

[13] Lennox, K.P., Dahl, D.B., Vannucci, D.B. and Tsai, J.W. (2010) A Dirichlet process mixture of hidden Markov models for protein structure prediction. *Ann. Appl. Stat.* **4**, pp.916–942.

[14] Lovell, S.C., Davis, I.W., Arendall III, B. and others,(2003). Structure validation by $C_\alpha$ geometry: $\phi, \psi$ and $C_\beta$ deviation *Proteins: Structure, Function and Genetics*, **50** , pp.437–450.

[15] Mardia, K.V., (1975). Statistics of directional data (with discussion) *J. Royal Statist. Soc. Series B*, **37**, pp.349–393.

[16] Mardia, K.V. (2007). On some recent advancements in applied shape analysis and directional statistics, in *Proceedings of the Leeds Annual Statistical Research Conference,Systems Biology & Statistical Bioinformatics*, Leeds University Press, Leeds, pp.9–17.

[17] Mardia, K.V. (2008). Holistic statistics and contemporary life sciences, in *Proceedings of the Leeds Annual Statistical Research Conference,The Art and Science of Statistical Bioinformatics*, Leeds University Press, Leeds, pp.9–17.

[18] Mardia, K.V. (2010) Bayesian analysis for bivariate von Mises distributions. *J. Applied Statist.*, **37**, pp.515–528.

[19] Mardia, K.V., Hughes, G., Taylor, C.C. and Singh, H. (2008) A multivariate von Mises distribution with applications to bioinformatics. *Can. J. Statist.*, **36**, pp.99-109.

[20] Mardia, K.V. and Jupp, P.E. (2000) *Directional Statistics*. Wiley, Chichester, 2nd ed.

[21] Mardia, K. V., Nyirongo, V. B., Fallaize, C. J., Barber, S. and Jackson, R. M. (2010). Hierarchical Bayesian modeling of pharmacophores in bioinformatics, *Biometrics*.doi:10.1111/j.1541–0420.2010.01460.x.

[22] Mardia, K.V. and Patrangenaru, V. (2005) Directions and projective shapes, *Ann. Statist.* **33** , pp.1666–1699.

[23] Mardia, K.V., Taylor, C.C., and Subramaniam, G.K. (2007). Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics*, **63**, pp.505–512.

[24] Ramachandran, G.N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations, *Molecular Biology*, **7**, pp.95–99.

[25] Singh, H., Hnizdo, V., and Demchuk, E. (2002). Probabilistic model for two dependent circular variables, *Biometrika*, **89** , pp.719–723.

[26] Schmidler, S. C. (2007). Fast Bayesian shape matching using geometric algorithms, in J. M. Bernardo, J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. Smith and M. West (eds), Bayesian Statistics 8, Oxford University Press, Oxford, pp. 471-490.

[27] Schmidler, S. C. (2010). Bayesian flexible shape matching with applications to structural bioinformatics. URL: http://www.stat.duke.edu/ scs/PubsByTopic.shtml # Shape

## RÉSUMÉ (ABSTRACT)

*If the last century in Science belongs to Physical Sciences then this century must belong to Life Sciences with many breakthroughs starting from DNA and proteins! We highlight what new tools are required in Shape Analysis and Directional Statistics to solve some major problems in structural bioinformatics.*