Bayesian Modelling for Matching and Alignment of Biomolecules

P. J. Green, K. V. Mardia, V. B. Nyirongo and Y. Ruffieux

August 9, 2009

Abstract

The three-dimensional shape of a protein plays a key role in determining its function, so proteins in which particular atoms have very similar configurations in space often have similar functions. There is therefore a need for efficient methodology to identify, given two or more proteins represented by the coordinates of their atoms, subsets of those atoms which match within measurement error, after allowing for appropriate geometrical transformations to align the proteins. This chapter describes a Bayesian model-based methodology for such tasks, and presents several challenging applications.

1 Introduction

Technological advances in molecular biology over the last 15 to 20 years have generated many datasets – often consisting of large volumes of data on protein or nucleotide sequences and structure – that require new approaches to their statistical analysis. In consequence, some of the most active areas of research in statistics at present are aimed at such bioinformatics applications.

It is well known that proteins are the work-horses of all living systems. A protein is a sequence of amino acids, of which there are twenty types. The sequence folds into a 3-dimensional structure. We can describe the shape of this structure in terms of a main chain and side chains (for examples, see Branden and Tooze, 1999; Lesk, 2000). Three atoms of each amino acid, denoted N, C_α and C', are in the main chain or backbone. The other part of the amino acid that is attached to the C_α atom is called the residue, and forms a side chain (that is, there is one side chain or residue for each amino acid in the protein).

One of the major unsolved biological challenges is the protein folding problem: how does the amino acid sequence fold into a three-dimensional protein? In particular, how can the three dimensional protein structure (and its function) be predicted from the amino acid sequence? These are key questions, since both shape and chemistry are important in understanding a protein's function.

1.1 Protein data and alignment problems

One particular task where statistical modelling and inference can contribute to scientific understanding of protein structure is that of matching and alignment of two or more proteins. This chapter addresses the analysis of 3 data sets of the following nature.

The 3-dimensional structure of a protein is important for it to perform its function. In chemoinformatics, it is a common assumption that structurally similar molecules have similar activities; in consequence, protein structure similarity can be used to infer the unknown function of a candidate protein (Leach, 2003). In drug design for example, a subject of prime interest is the local interaction between a small molecule (the ligand) and a given protein receptor. If the geometrical structure of the receptor is known, then established methods such as docking can be applied in order to specify the protein–ligand interaction. However, in most cases this structure is unknown, meaning the drug designer must rely on a study of the similarity (or diversity) in available ligands.

The alignment of the molecules is an important first step towards such a study. Thus one of the problems in protein structural bioinformatics is matching and aligning 3-dimensional protein structures or related configurations e.g. active sites, ligands, substrates, steroid molecules. To consider the matching between proteins, one normally considers the C_α atoms. A related application but with different aims is the matching of 2-dimensional protein gels.

Given two or more proteins represented by configurations of points in space representing locations of particular atoms of the proteins, the generic task of matching and alignment is to discover subsets of these configurations that coincide, after allowing for measurement error and unknown geometrical transformations of the proteins. These applications require algorithms for matching, as well as statistics and distributions of measures for quantifying quality of matching and alignment.

Statistical shape analysis potentially has something to offer in solving matching and alignment problems; the field of labelled shape analysis with labelled points is well developed (see Dryden and Mardia, 1998, also Appendix A.1) but unlabelled shape analysis is still in its infancy. The methodology we have developed for matching and alignment is a contribution to shape analysis for unlabelled and partially-labelled data.

Pairwise matching of active sites data sets

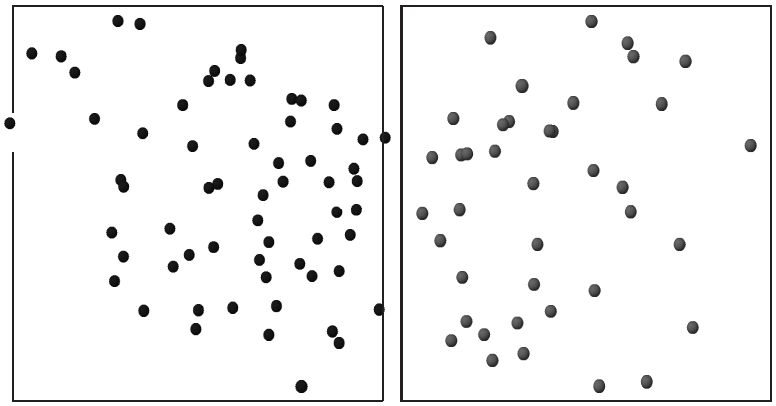
An active site is a local 3-dimensional arrangement of atoms in a protein that are involved in a specific function e.g. binding a ligand (and so known as a binding site). Atoms in active sites are from amino acids that are close to each other in 3-dimensional space but do not necessarily follow closely in sequence order.

We consider two datasets, analysed with different objectives. We have configurations of C_α atoms in two functional sites shown in Figure 1 from 17–β hydroxysteroid-dehydrogenase and carbonyl reductase proteins (two dimensional view of the data, given at `http://www.stats.bris.ac.uk/~peter/Align/index.html`). These functional sites are related but which and how many atoms correspond are unknown. Our aim would be to find matching atoms and align these configurations.

We also consider matches of protein active sites from SITESDB (cf. Gold, 2003). An alcohol dehydrogenase active site (1hdx_1) is matched against NADP-binding sites of a quinone oxidoreductase protein (1qor_0); this is a small example of a query used against a large database.

Matching multiple configurations of steroid molecules

The CoMFA (Comparative Molecular Field Analysis) database is a set of steroid molecules which has become a benchmark for testing various 3D quantitative structure-activity relationship (QSAR) methods (see Coats, 1998). This database contains the three-dimensional coordinates for the atoms in each of the 31 molecules, in addition to additional



a) 1a27_0 (63 atoms) b) 1cyd_0 (40 atoms)

Figure 1: C_α atoms in functional sites of 17–β hydroxysteroid-dehydrogenase (1a27_0) and carbonyl reductase (1cyd_0) represented as spheres in RasMol

information such as atom type and partial charge. The geometrical similarity between the molecules makes this database an ideal test-bed for our multiple alignment methods. The CoMFA steroids can be accessed from `http://www2.ccc.uni-erlangen.de/services/steroids/`.

1.2 Geometrical transformations

Matching and alignment is conducted generally assuming that a geometrical transformation has to be applied to each of the point configurations to bring them all onto a standard orientation and scale in which points to be matched are simply those closest together. These transformations are usually considered to be unknown and have to be inferred from the data. In the Bayesian formulations we follow below, modelling the transformations amounts to deciding on a space of transformations that are appropriate for the problem, and a prior distribution over that space; these choices will depend on understanding of both the physical processes that led to any variation in the geometry of the different configurations, and of the observational process by which the point coordinates are recorded. In the two applications presented in this chapter, we assume the space of "rigid-body" transformations, that is, rotation and translation, and place a uniform uninformative prior on the rotation. This would be appropriate, for example, if there was no systematic distortions between the configurations, but they are presented for measurement in an arbitrary orientation. Non-trivial priors on rigid-body or other affine transformations are handled by similar methodology. For other problems, other spaces of transformation will be necessary – for example, aligning protein gels would usually require smooth "warping" rather than any affine transformation. Methodology for such situations is still under development.

1.3 Structure and scope

Green and Mardia (2006) proposed a Bayesian hierarchical model allowing inference about the matching and alignment of two configurations of points; in Section 2 this is described as it would be used for pairwise matching of configurations of active sites. Ruffieux and Green (2008) developed the Green and Mardia (2006) model to handle the simultaneous

matching of multiple configurations (see also Marín and Nieto, 2008), and this is reviewed in Section 3. Analysis of the three data sets introduced above is presented in Section 4. We end the chapter with discussion of conclusions and future directions for research. There are various appendices including Appendix A.1 reviewing labelled shape analysis, Appendix B.1 on model formulation, Appendix B.2 on MCMC implementation, and Appendix B.3 on web data sources.

2 A Bayesian hierarchical model for pairwise matching

We have two point configurations, $X^{(1)} = \{x_j, j = 1, 2, \ldots, m\}$ and $X^{(2)} = \{y_k, k = 1, 2, \ldots, n\}$, in d-dimensional space $\mathcal{R}^d$. The points are labelled for identification, but arbitrarily. In our applications the points are C_α atoms.

A latent point process model

The key basis for our model for the configurations is that both point sets are regarded as noisy observations on subsets of a set of unobserved true locations $\{\mu_i\}$, where we do not know the mappings from j and k to i. There may be a geometrical transformation between the x-space and the y-space, which may also be unknown. The objective is to make model-based inference about these mappings, and in particular make probability statements about matching: which pairs (j, k) correspond to the same true location?

We will assume the geometrical transformation between the x-space and the y-space to be affine, and denote it by $x = \mathcal{A}y = Ay + \tau$. Later we will restrict A to be a rotation matrix, so that this is a rigid-body transformation. We regard the true locations $\{\mu_i\}$ as being in x-space, without loss of generality.

The mappings between the indexing of the $\{\mu_i\}$ and that of the data $\{x_j\}$ and $\{y_k\}$ are captured by indexing arrays $\{\xi_j\}$ and $\{\eta_k\}$; to be specific we assume that

$$x_j = \mu_{\xi_j} + \varepsilon_{1j}, \tag{1}$$

for $j = 1, 2, \ldots, m$, where $\{\varepsilon_{1j}\}$ have probability density f_1, and

$$Ay_k + \tau = \mu_{\eta_k} + \varepsilon_{2k}, \tag{2}$$

for $k = 1, 2, \ldots, n$, where $\{\varepsilon_{2k}\}$ have density f_2. All $\{\varepsilon_{1j}\}$ and $\{\varepsilon_{2k}\}$ are independent of each other, and independent of the $\{\mu_i\}$. We take f_1 and f_2 to be normal but the method generalises to any f_1 and f_2.

Multiple matches are excluded, and thus each hidden point μ_i is observed at most once in each of the x and y configurations; equivalently, the ξ_j are distinct, as are the η_k. The label i is not used in our subsequent development, and all that is needed is the matching matrix M, defined by $M_{jk} = 1$ if $\xi_j = \eta_k$ otherwise 0. This structure is a latent variable in our model, and its distribution is derived from the latent point process model as follows.

Prior for M. Suppose that the set of true locations $\{\mu_i\}$ forms a homogeneous Poisson process with rate λ over a region $V \subset \mathcal{R}^d$ of volume v, and that N points are realised in this region. Some of these give rise to both x and y points, some to points of one kind and not the other, and some are not observed at all. We suppose that these four possibilities occur independently for each realised point, with probabilities parameterised so that with probabilities $(1 - p_{\mathrm{x}} - p_{\mathrm{y}} - \rho p_{\mathrm{x}} p_{\mathrm{y}}, p_{\mathrm{x}}, p_{\mathrm{y}}, \rho p_{\mathrm{x}} p_{\mathrm{y}})$ we observe neither, x alone, y alone, or

both x and y, respectively. The parameter ρ is a certain measure of the tendency a priori for points to be matched: the prior probability distribution of L conditional on m and n is proportional to

$$p(L) \propto \frac{(\rho/\lambda v)^L}{(m-L)!(n-L)!L!}, \tag{3}$$

for $L = 0, 1, \ldots, \min\{m, n\}$. The normalising constant here is the reciprocal of $H\{m, n, \rho/(\lambda v)\}$, where H can be written in terms of the confluent hypergeometric function

$$H(m, n, d) = \frac{d^m}{m!(n-m)!}\, {}_1F_1(-m, n-m+1, -1/d),$$

assuming without loss of generality that $n > m$; (see Abramowitz and Stegun, 1970, p. 504)

Assuming that M is a priori uniform conditional on L, we have

$$p(M) = p(L)p(M|L) = \frac{(\rho/\lambda v)^L}{\sum_{\ell=0}^{\min\{m,n\}} \ell! \binom{m}{\ell}\binom{n}{\ell}(\rho/\lambda v)^\ell}.$$

One application of this distribution of L is a similarity index (Davies et al., 2007).

The likelihood

Let

$$x_j \sim N_d(\mu_{\xi_j}, \sigma_{\mathrm{x}}^2 I) \qquad \text{and} \qquad Ay_k + \tau \sim N_d(\mu_{\eta_k}, \sigma_{\mathrm{y}}^2 I),$$

with $\sigma_{\mathrm{x}} = \sigma_{\mathrm{y}} = \sigma$, say. On integrating out the μs the joint model for parameters, latent variables and observables is (Green and Mardia, 2006)

$$p(M, A, \tau, \sigma, x, y) \propto |A|^n p(A)p(\tau)p(\sigma) \prod_{j,k:M_{jk}=1} \left(\frac{\kappa\phi\{(x_j - Ay_k - \tau)/\sigma\sqrt{2}\}}{(\sigma\sqrt{2})^d} \right), \tag{4}$$

where ϕ is the standard normal density in $\mathcal{R}^d$. Here $|A|$ is the Jacobian of transformation from x-space into y-space; $p(A)$, $p(\tau)$ and $p(\sigma)$ denote prior distributions for A, τ and σ; d is the dimension of the configurations i.e. $d = 2$ for 2-dimensional configurations e.g. protein gels and $d = 3$ for 3-dimensional configurations e.g. active sites; $\kappa = \rho/\lambda$ measures the tendency *a priori* for points to be matched and can be a function of concomitant information e.g. amino acid types in matching protein structures. See the directed acyclic graph, Figure 2a, for a graphical representation of the model. Green and Mardia (2006) give a generalisation where ϕ can be replaced by a more general density depending on f_1 and f_2. Details on priors are given below.

There is a connection between maximising the joint posterior derived from Equation 4 and minimising root mean square deviations (RMSD), defined by

$$\mathrm{RMSD}^2 = Q/L, \text{ where } Q = \sum_{j,k} M_{jk}||x_j - Ay_k - \tau||^2, \tag{5}$$

and $L = \sum_{j,k:M_{jk}=1} M_{jk}$ denotes the number of matches. RMSD is the focus of study in combinatorial algorithms for matching. In the Bayesian formulation the log likelihood (with uniform priors) is proportional to

$$const. - 2\left(\sum M_{jk}\right)\log\sigma + \left(\sum M_{jk}\right)\log\rho - \frac{1}{2}\frac{Q}{\sigma^2\sqrt{2}}.$$

The maximum likelihood estimate of σ for a given matching matrix M is the same as the $RMSD$ which is the least squares estimate. $RMSD$ is a measure commonly used in bioinformatics, although joint uncertainty in $RMSD$ and the matrix M is difficult to appreciate except in the Bayesian formulation.

Prior distributions for continuous variables

For the continuous variables τ, σ^{-2} and A we use conditionally conjugate priors so

$$\tau \sim N_d(\mu_\tau, \sigma_\tau^2 I), \qquad \sigma^{-2} \sim \Gamma(\alpha, \beta), \qquad A \sim \text{Fisher}(F)$$

Here, Fisher(F) denotes the matrix Fisher distribution; (see, for example, Mardia and Jupp, 2000, p. 289). For $d = 2$, A has a von Mises distribution. For $d = 3$, it is useful to express A in terms of the Eulerian angles. Some efficient methods to simulate A are given in Green and Mardia (2006). If the point configurations are presented in arbitrary orientations, it is appropriate to assume a uniform distribution on A, that is, $F = 0$, and this is usually adequate.

3 Alignment of multiple configurations

In this section we consider a hierarchical model for matching configurations $X^{(1)}, X^{(2)}, \ldots, X^{(C)}$ simultaneously.

Multi-Configuration Model

The pairwise model presented above can be readily extended to the multi-configuration context. Suppose we have C point configurations $X^{(1)}, X^{(2)}, \ldots, X^{(C)}$, such that $X^{(c)} = \left\{x_j^{(c)}, j = 1, 2, \ldots, n_c\right\}$, where $x_j^{(c)} \in \mathcal{R}^d$ and n_c is the number of points in configuration $X^{(c)}$. As in the pairwise case we assume the existence of a set of 'hidden' points $\mu = \{\mu_i\} \subset \mathcal{R}^d$ underlying the observations. Our multiple-configuration model is thus:

$$\mathcal{A}^{(c)} x_j^{(c)} = \mu_{\xi_j^{(c)}} + \varepsilon_j^{(c)}, \quad \text{for } j = 1, 2, \ldots, n_c, \quad c = 1, 2, \ldots, C. \tag{6}$$

The unknown transformation $\mathcal{A}^{(c)}$ brings the configuration $X^{(c)}$ back into the same frame as the μ-points, and $\xi^{(c)}$ is a labelling array linking each point in configuration $X^{(c)}$ to its underlying μ-point. As in the previous section the elements within each labelling array are assumed to be distinct. In this context a match can be seen as a set of points $\left(x_{j_1}^{(i_1)}, x_{j_2}^{(i_2)}, \ldots, x_{j_k}^{(i_k)}\right)$ such that $\xi_{j_1}^{(i_1)} = \xi_{j_2}^{(i_2)} = \ldots = \xi_{j_k}^{(i_k)}$.

Thus matches may now involve more than two points at once. These are stored in a structure $\mathcal{M}$. This parameter plays the same role as the matrix M from Section 2, in that it contains the relevant information on the matches. We choose to categorise the matches according to their 'type'. Consider a generic set $I \subset \{1, 2, \ldots, C\}$ of configuration indices, with $I \neq \emptyset$. This set corresponds to a type of match: for example if $C = 3$, then $I = \{2, 3\}$ refers to a match involving a point from the $X^{(2)}$ configuration and a point from the $X^{(3)}$ configuration but none from the $X^{(1)}$ configuration. We call I-*match* a match involving *exactly* the configurations whose index is included in I.

The likelihood

Write S_I as the set of I-matches contained in $\mathcal{M}$. The elements of S_I are written as index arrays of the form $\left(j_1, j_2, \ldots, j_{|I|}\right)$, with the convention that $\left\{x_{j_1}^{(i_1)}, x_{j_2}^{(i_2)}, \ldots, x_{j_{|I|}}^{(i_{|I|})}\right\}$ is the corresponding set of matched points. Let

$$\mathcal{A}^{(c)} x_j^{(c)} \sim N_d\left(\mu_{\xi_j^{(c)}}, \sigma^2 I\right),$$

for $c = 1, 2 \ldots, C$ and $j = 1, 2, \ldots, n_c$. Supposing $\mathcal{A}^{(c)}$ is made up of a transformation matrix $A^{(c)}$ and a translation vector $\tau^{(c)}$, our posterior model has the form

$$\begin{aligned} p\left(\mathcal{A}, \mathcal{M} \mid X\right) &\propto \prod_{c=1}^{C} \left\{ p(A^{(c)}) p(\tau^{(c)}) \left| A^{(c)} \right|^{n_c} \right\} \\ &\times \prod_{I} \prod_{(j_1,\ldots,j_{|I|}) \in S_I} \kappa_I |I|^{-d/2} (2\pi\sigma^2)^{-d(|I|-1)/2} \times \exp\left\{ -\frac{1}{2\sigma^2} \gamma_{\mathcal{A}}\left(x_{j_1}^{(i_1)}, x_{j_2}^{(i_2)}, \ldots, x_{j_{|I|}}^{(i_{|I|})}\right) \right\}, \end{aligned} \tag{7}$$

where $\mathcal{A} = \left(\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \ldots \mathcal{A}^{(C)}\right)$ and $X = \left(X^{(1)}, X^{(2)}, \ldots, X^{(C)}\right)$, and

$$\gamma_{\mathcal{A}}\left(x_{j_1}^{(i_1)}, x_{j_2}^{(i_2)}, \ldots, x_{j_{|I|}}^{(i_{|I|})}\right) = \sum_{k=1}^{|I|} (y_k - \overline{y})^2$$

where

$$y_k = A^{(i_k)} x_{j_k}^{(i_k)} + \tau^{(i_k)} \quad \text{and} \quad \overline{y} = \sum_{k=1}^{|I|} y_k / |I|.$$

If we set $C = 2$, $A^{(1)} = I$ and $\tau^{(1)} = 0$, this posterior distribution is equivalent to the one given in the previous Section.

Note that we have implicitly defined a prior distribution for $\mathcal{M}$ in the expression for the posterior distribution: this prior is described explicitly below.

Prior distributions

Prior distributions for the $\tau^{(c)}, A^{(c)}$, and σ^2 are identical to the ones for the parameters in the pairwise model. In particular we set, for $c = 1, 2, \ldots, C$,

$$\tau^{(c)} \sim N_d\left(\mu^{(c)}, \sigma_c^2 I\right), \qquad A^{(c)} \sim \text{ Fisher}(F_c),$$

and $\sigma^{-2} \sim \Gamma(\alpha, \beta)$.

To construct a prior for the matches $\mathcal{M}$, we again assume that the μ-points follow a Poisson process with constant rate λ over a region of volume v. Each point in the process gives rise to a certain number of observations, or none at all. For I as defined previously, let q_I bet the probability that a given hidden location generates an I-match. We impose the following parametrisation:

$$q_I \;=\; \rho_I \cdot \prod_{c \in I} q_{\{c\}},$$

where $\rho_I = 1$ if $|I| = 1$. Define L_I as the number of I-matches contained in $\mathcal{M}$, and assume the conditional distribution of $\mathcal{M}$ given the L_Is is uniform. After some combinatorical work we find that the prior distribution for the matches can be expressed as

$$p(\mathcal{M}) \quad \propto \quad \prod_I \left(\frac{\kappa_I}{v^{|I|-1}}\right)^{L_I},$$

where $\kappa_I = \rho_I/\lambda^{|I|-1}$. It is easy to see that this is simply a generalisation of the prior distribution for the matching matrix M.

Identifiability issue

To preserve symmetry between configurations, we only consider the case where the $A^{(c)}$ are uniformly distributed *a priori*. It is then true that the relative rotations $\left(A^{(c_1)}\right)' \cdot A^{(c_2)}$ are uniform and independent for $c_2 \neq c_1$ and fixed c_1. So without loss of generality, we can now impose the identifying constraint that $\mathcal{A}^{(1)}$ be fixed as the identity transformation. This is the same as saying that the first data configuration lies in the same frame as the hidden point locations, as was the case in the pairwise model.

4 Data analysis

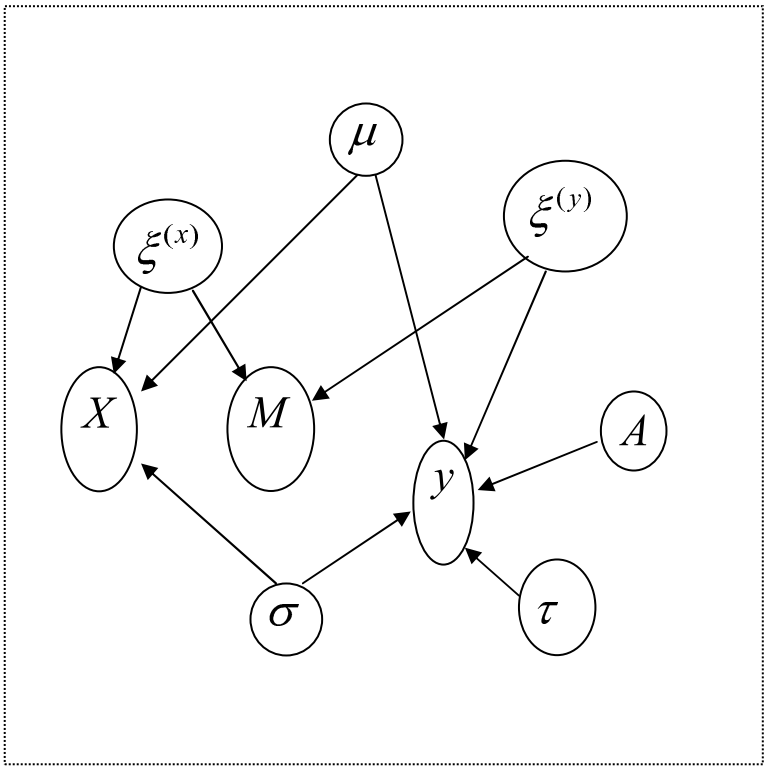
4.1 Active sites and Bayesian refinement

4.1.1 Two sites

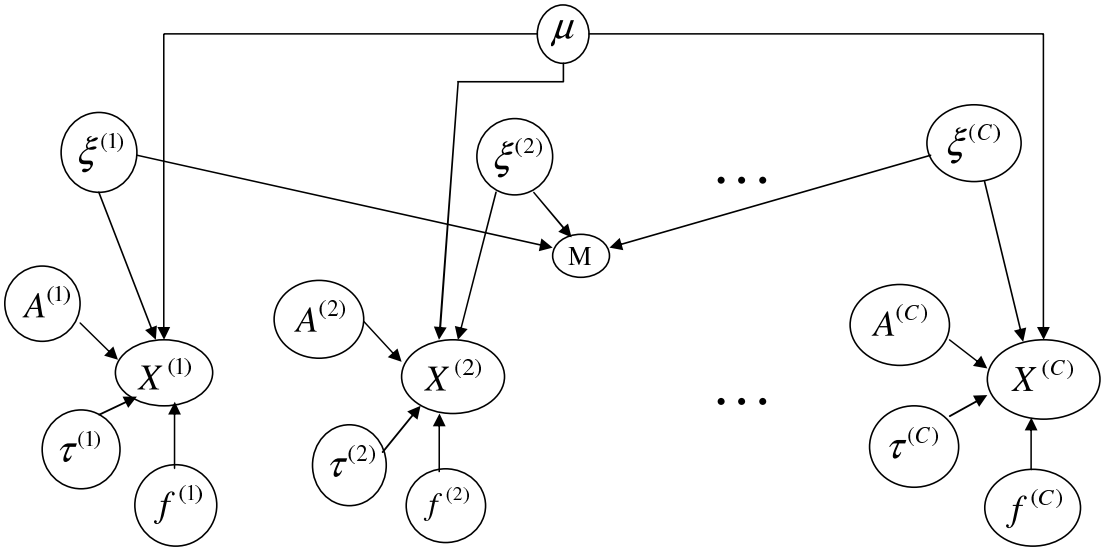
We now illustrate our method applying on 2 active sites described in Section 1. A sampler described in Appendix B.2 was run for 100,000 sweeps (including 20,000 iterations for burn-in period) to match 17–β hydroxysteroid-dehydrogenase and carbonyl reductase active sites. These sites are described in Section 1.1. Prior and hyperprior settings were $\alpha = 1$, $\beta = 2$, $\mu_\tau = (0, 0, 0)'$, $\sigma_\tau = 20$, $\lambda/\rho = 0.0005$, $F = 0$. We match 34 points as shown in Figure 3a with $RMSD = 0.949$Å. The algorithm was also used with restriction to match only points representing same type of amino acid. With the restriction on type of points (concomitant information), 15 matches shown in Figure 3b are made with $RMSD = 0.727$Å.

4.1.2 Queries in a database

We now discuss how we could use an initial solution from other software, e.g. the graph theoretic BK technique (Gold, 2003), derived from a physico-chemistry view point, and use our algorithm to refine the solution. In this context, the data comparisons are substantial since they involve comparing a query with other family members. We compare here with the graph theoretic approach that requires adjusting the matching distance threshold *apriori* according to noise in atomic positions, which is difficult to pre-determine in bioinformatics applications involving matching configurations in a database with varying crystallographic precision. Furthermore, the graph method is unable to identify alternative but sometimes important solutions in the neighbourhood of the distance based solution because of strict distance thresholds. On the other hand, the graph theoretic approach is very fast, robust and can quickly give corresponding points for small configurations from which we can get initial estimates for rotation and translation. We illustrate here how our approach finds more biologically interesting and statistically significant matches between functional sites (Mardia et al., 2007*a*).

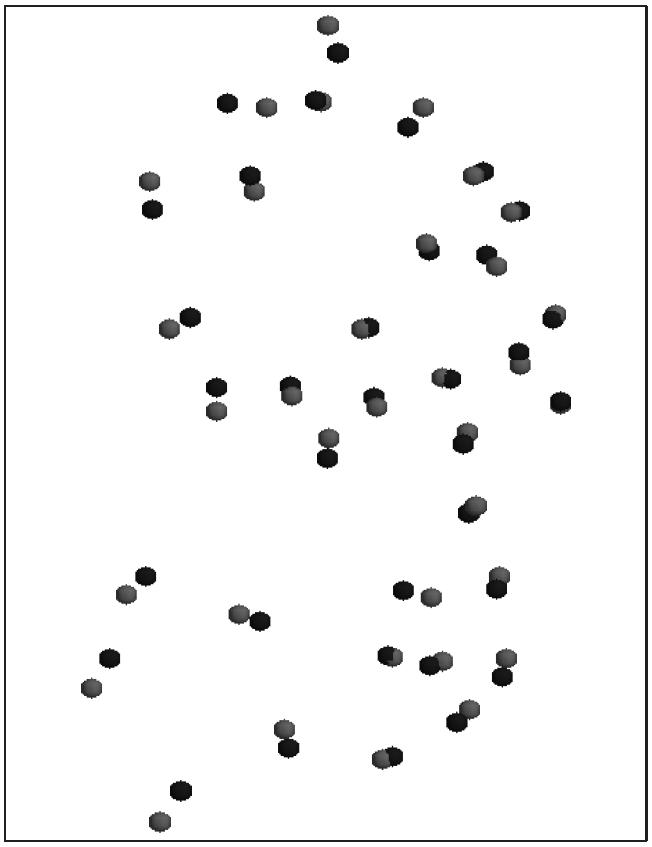


a)



b)

Figure 2: Directed acyclic graph representing the model for matching a)two configurations, b) multiple configurations. The graph shows all data and parameters treated as variables.



a) Without restriction on matches

	1a27_0	1cyd_0
1	G 9	G 14
2	G 13	G 18
3	I 14	I 19
4	G 15	G 20
5	R 37	R 39
6	D 65	D 60
7	N 90	N 83
8	A 91	A 84
9	L 93	L 86
10	V 113	V 106
11	S 142	S 136
12	Y 155	Y 149
13	K 159	K 153
14	V 188	V 182
15	T 190	T 184

b)With restriction on matches

Figure 3: Matched points (C_α atoms): a) without restriction on matches according to amino acid type; b) matching rotation and same types of amino acids.

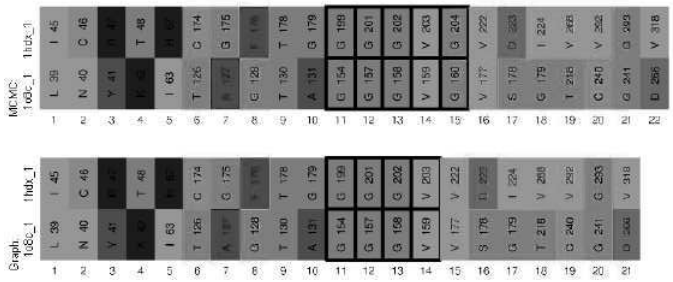


Figure 4: Corresponding amino acids between the NAD-binding site of alcohol dehydrogenase (1hdx_1) and NADP-binding site of quinone oxidoreductase (1qor_0) before and after MCMC refinement step. Amino acids with bold borders are part of the dinucleotide binding motif GL-GGVG.

We model graph theoretic matches of protein active sites from SITESDB (cf. Gold, 2003). An alcohol dehydrogenase active site (1hdx_1) is matched against NADP-binding sites of a quinone oxidoreductase protein (1qor_0). Figure 4 gives matched amino acids by the graph method and refined by the Bayesian algorithm (see Appendix B.2).

The Markov chain Monte Carlo (MCMC) refinement step produced improvements with obvious biochemical relevance. These proteins share a well known glycine rich motif (GXGXXG) in the binding site. For 1qor_0, before the MCMC refinement step, only 2 glycines in dinucleotide binding motif GLGGVG were matched by the graph theoretic approach and this increased to 3 glycines after MCMC refinement.

In this kind of context, only short MCMC runs may be possible, and we cannot have full confidence that the whole posterior space is being sampled. However, we should explore the mode containing the graph-theoretic initial solution, possibly refine that solution, and get an idea of uncertainty.

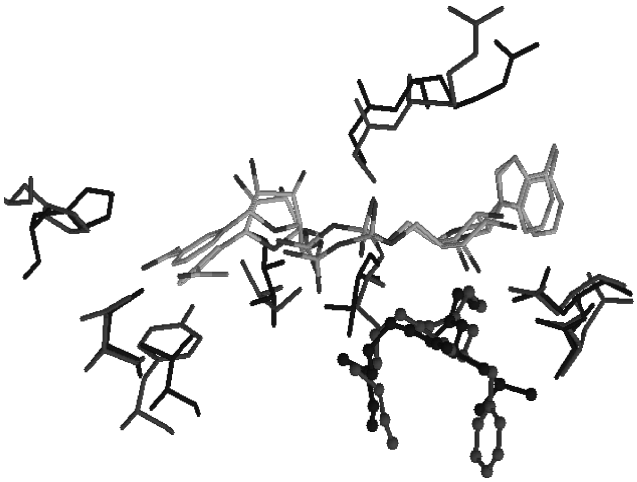


Figure 5: Superposition of matching amino acids between alcohol dehydrogenase (1hdx_1; blue) and glyceraldehyde-3-phosphate dehydrogenase (3dbv_3; red) binding sites after MCMC refinement (RMSD = 0.672; number of corresponding amino acids = 12; p-value = 3.68e-05). The matched dinucleotide binding motif is shown in ball-and-stick representation. Ligands are coloured in CPK colours.

4.2 Aligning multiple steroid molecules

Here we attempt to align $C = 5$ configurations simultaneously, using the method described in Section 3. An MCMC algorithm (see Appendix B.2) is used to simulate a random sample from the distribution (7); this sample is then used as a basis for inference.

We select 5 molecules from the CoMFA database (see Section 1.1); these are aldosterone, cortisone, prednisolone, 11-deoxycorticosterone, and 17a-hydroxyprogesterone. All of these molecules contain 54 atoms. We set the following hyperparmeter values: $\alpha = 1, \beta = 0.1$ and $\mu^{(c)} = 0, \sigma_c^2 = 10$, for $c = 2, 3, 4, 5$, and $F_c = 0$ for all c. For the match prior parameters we set $\kappa_I = 1$ for $|I| = 1$, $\kappa_I = 14$ for $|I| = 2$, $\kappa_I = 289$ for $|I| = 3$, $\kappa_I = 12056$ for $|I| = 4$ and $\kappa_I = 15070409$ for $|I| = 5$. These values were determined by making initial 'guesses' on the number of matches of each type, and adapting the prior distribution (8) accordingly. As in the pairwise case we obtain estimates for the rotations, translations, and matches. In particular, the matches are ranked according to their frequency in the posterior distribution, and we choose to select the k most frequent, say.

In Figure 6 we display the time series traces of $L_{\{1,2,3,4,5\}}$ and $L_{\{2,3,4,5\}}$ in the MCMC output. Here we find that 56 matches have sample probability higher than 0.5. In Figure 7 we align the five molecules by applying the estimated transformations to each configuration. It is interesting to note that in the latter figure, the top right portion of the first molecule is slightly detached from the other four, and indeed the MCMC output suggests that those points from the first configuration should not be matched to the other four. However when aligning the molecules in pairs using the method from Section 2, the inference tends to favour matching these points, even if we set the match hyperparameters to be biased against matching. This confirms that inclusion of two or three additional configurations may have a positive impact on the alignment inference. One might understand this as a 'borrowing of strength' of sorts: further configurations provide further information on the number and location of implied μ-points, information which can in turn be exploited in the alignment of the initial configurations. Clearly there is no way to take advantage of this information if the molecules are aligned independently in pairs.

5 Further discussion

5.1 Advantages of the Bayesian modelling approach

Some advantages of our Bayesian approach to these problems are:

1. Simultaneous inference about both discrete and continuous variables.
2. The full Bayesian posterior "tool kit" is available for inference.
3. It allows in a natural way any prior information
4. The MCMC may be too slow in some application but it has a role to play as a gold standard against heuristic approaches.
5. The MCMC implementation provides a greater chance to escape local modes compared to optimisation methods.

Wilkinson (2007*b*) has given a review of Bayesian methods in bioinformatics and computational system biology more generally, citing some of these points; in particular he has pointed out why bioinformatics and computational system biology are undergoing a Bayesian revolution similar to that already seen in genetics.

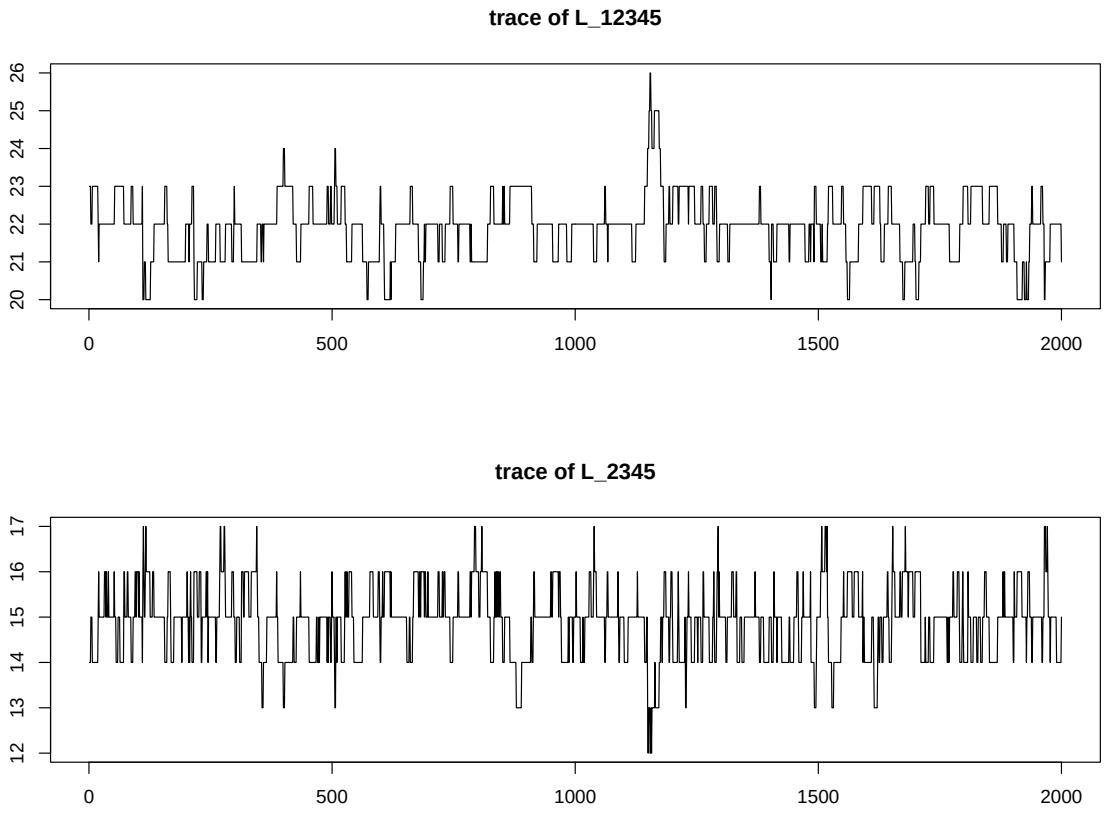


Figure 6: Time series traces of the number of matches involving all configurations (top), and involving all configurations except the first (bottom). Taken from a thinned sample of 2000 after burn-in.

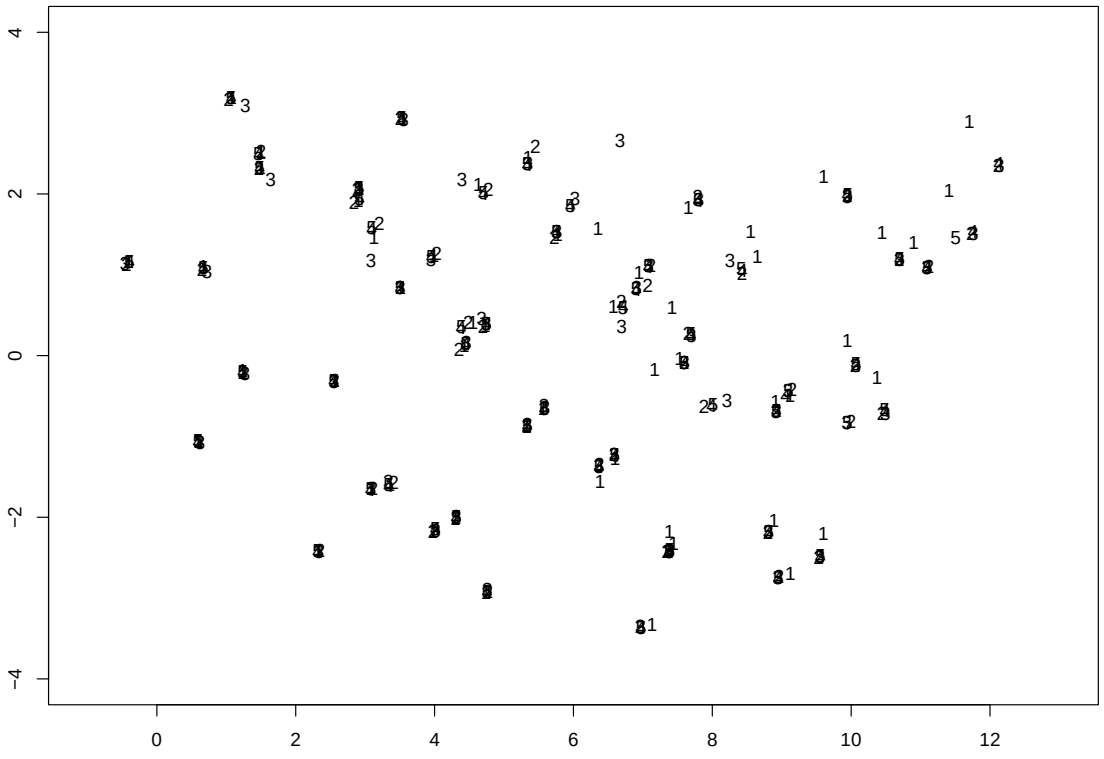


Figure 7: Multiple alignment of the five steroid molecules: the full transformations are estimated from a MCMC subsample of size 2000, and are filtered out from the data. The points are projected onto the first two canonical axes, and are labelled according to the number of the configuration they belong to.

5.2 Outstanding issues

We expect that further work, by the authors and others, will be directed to refinement of the methodology described here.

5.2.1 Modelling

Spherical normality of the errors ε in (1), (2) and (6) was assumed for simplicity, and it may be necessary to relax this assumption. These errors represent both measurement error in recording the data, and 'model errors', small variations between the molecules in the locations of the atoms. There is an interplay between the sphericity assumption and the modelling of the geometrical transformations $\mathcal{A}$, so care is needed here, but it is straightforward to replace normality by a heavy-tailed alternative.

Further study is needed on setting hyperparameters and sensitivity to these choices. The analysis seems to be most sensitive to the parameter κ but otherwise rather robust.

As mentioned before, particular applications of matching and alignment demand more subtle modelling of the geometrical transformations $\mathcal{A}$, with the extension to non-parametric warping being most pressing.

Finally, there are interesting modelling issues concerned with using sequence information to influence the inference on matching and alignment. The most promising direction within our modelling paradigm involves the use of non-uniform 'gap' priors on the matching matrix M, encouraging or requiring matchings that respect the sequence order.

5.2.2 Computation

Design of MCMC samplers to deal comprehensively with problems of multimodality in the posterior distribution is the major challenge here. One can expect that generic techniques such as simulated tempering will have a part to play. We also expect further work on inferential methods that are perhaps not fully Bayesian, but are computationally faster, including the use of fast initialisation methods, such as

(a) Starting from the solution from 3-dimensional deterministic methods such as graph theoretic, CE, geometric hashing and others.

(b) For full protein alignment of structure, using well established sequence alignment software e.g. BLAST as the starting point.

5.3 Alternative approaches

5.3.1 EM approach

The interplay between matching (that is, allocation), and parameter uncertainty has something in common with mixture estimation. This might suggest considering maximisation of the posterior by using the EM algorithm, which could of course in principle be applied either to maximum likelihood estimation or to maximum a posteriori estimation. For the EM formulation, the "missing data" are the matches.

The "expectations of missing values" are just probabilities of matching. These are only tractable if we drop the assumption that a point can only be matched with at most one other point; that is, that $\sum_j M_{jk} \leq 1$ for all k, $\sum_k M_{jk} \leq 1$ for all j. We then get "soft matching".

EM allows us to study only certain aspects of an approximate version of our model, and is not trivial numerically. Obtaining the complete posterior by Markov chain Monte

Carlo sampling gives much greater freedom in inference. For a direct EM approach see Kent et al. (2004, 2008).

5.3.2 Procrustes type approaches

Dryden et al. (2007) and Schmidler (2007) use a MAP estimator for M after estimating "nuisance parameters" (A, τ, σ^2) from Procrustes registration. Schmidler (2007) has provided a fast algorithm using geometric hashing. Thus these borrow strength from labelled shape analysis. The Green and Mardia (2006) procedure also allows informative priors so the procedure is very general. Dryden (2007) has given some initial comparisons and in particular their MAP approach often get stuck in local modes. For small variability, both approaches lead to similar results. Wilkinson (2007*a*) has touched many important problems in Bayesian alignment in particular the uniform prior would be strongly biased towards larger values of $L = |M|$. Schmidler (2007) and Dryden et al. (2007) effectively assume that Procrustes alignment is "correct" and does not reflect uncertainty in geometric alignment. By integrating out the geometrical transformation as in Green and Mardia (2006) then alignment uncertainty will be propagated correctly without suffering significant computation penalty. Mardia (2007*a*) has raised a few general issues about matching in the discussion.

5.4 Future directions

Bayesian approaches are particularly promising in tackling problems in bioinformatics, with their inherent statistical problems of multiple testing, large parameter spaces for optimisation and model or parameter non-identifiability due to high dimensionality. One of the major statistical tasks is to build simulation models of realistic proteins by incorporating local and long-range interactions between amino acids. A protein can be uniquely determined by a set of conformational angles so directional statistics (see Mardia and Jupp, 2000) plays a key role as well as shape analysis. Boomsma et al. (2008) has given a dynamical Bayesian network with angular distributions and amino acid sequences as its nodes for protein (local) structure prediction. This solves one of the two major bottle necks for protein based nanotechnology namely generating native(natural) protein-like structures. Another is an appropriate energy function to make it compact. The angular distributions used a priori adequately describe Ramachandran plots of the dihedral angles of the backbone (see Mardia, Taylor and Subramaniam, 2007*b*; Mardia et al., 2008). Hamelryck et al. (2006) have given a method of simulating realistic protein conformation for C_α trace focusing on mimicking secondary structure while Mardia and Nyirongo (2008) focus on global properties e.g. compactness and globularity.

The word homology is used in a technical sense in biology, especially in discussion of protein sequences, which are said to be homologous if they have been derived from a common ancestor. Homology implies an evolutionary relationship and is distinct from similarity. In this chapter, alignment focuses only on similarity.

To sum up, there are real challenges for statisticians in the understanding of protein structure. Similar remarks apply to understand the RNA structure (see, for example Frellsen et al., 2008). All this might need is holistic statistics which implies a shift of paradigm by statisticians (Green, 2003; Mardia and Gilks, 2005; Mardia, 2007*b*, 2008). However, protein bioinformatics is a subset of very large area of bioinformatics which has many challenging problems (see Gilks, 2004; Mardia, 2005; Wilkinson, 2007*b*).

Acknowledgements

V.B. Nyirongo acknowledges funding from the School of Mathematics, University of Leeds as a visiting research fellow during the period in which part of this chapter was drafted.

References

Abramowitz, M. and Stegun, I.A. (1970), *Handbook of Mathematical Functions*, Dover, New York.

Artymiuk, P.J., Poirrette, A.R., Grindley, H.M., Rice, D.W. and Willett, P. (1994), A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures, *J. Mol. Biol.* **243**, 327–44.

Berkelaar, M. (1996), lpsolve - simplex-based code for linear and integer programming, `http://www.cs.sunysb.edu/~algorith/implement/lpsolve/implement.shtml`.

Bookstein, F.L. (1986). Size and shape spaces for landmark data in two dimensions. *Statistical Science*, **1**, 181-242.

Boomsma, W., Mardia, K.V., Taylor, C.C., Perkinghoff-Borg, J., Krogh, A. and Hamelryck, T. (2008), A generative, probabilistic model of local protein structure, *Proceedings of the National Academy of Science* **105**, 8932–8937.

Branden, C. and Tooze, J. (1999), *Introduction to Protein Structure*, 2nd edition, Garland Publishing Inc, New York.

Burkard, R.E. and Cela, E. (1999), Linear assignment problems and extensions, *in* P. Pardalos and D.-Z. Du, eds, Handbook of Combinatorial Optimization, Vol. 4, 75–149, Kluwer Academic Press, Boston.

Coats, E.A. (1998), The CoMFA steroid database as a benchmark dataset for development of 3D QSAR methods, *Perspectives in Drug Discovery and Design* **12-14**, 199–213.

Davies, J.R., Jackson, R.M., Mardia, K.V. and Taylor, C.C. (2007), The Poisson index: A new probabilistic model for protein-ligand binding site similarity, *Bioinformatics* **23**, 3001–3008.

Dryden, I.L. (2007), Discussion to Schmidler, 17–18. Listed here.

Dryden, I.L., Hirst, J.D. and Melville, J.L. (2007), Statistical analysis of unlabelled point sets: comparing molecules in chemoinformatics, *Biometrics* **63**, 237251.

Dryden, I.L. and Mardia, K.V. (1998), *Statistical Shape Analysis*, John Wiley, Chichester.

Frellsen, J., Moltke, I., Thiim, M., Mardia, K.V., Ferkinghoff-Borg, J. and Hamelryck, T. (2008), A probabilistic model of local RNA 3-d structure, *Submitted*.

Gilks, W. (2004), Bioinformatics: new science- new statistics., *Significance* **1**, 7–9.

Gold, N.D. (2003), Computational approaches to similarity searching in a functional site database for protein function prediction, Ph.D thesis, Leeds University, School of Biochemistry and Microbiology.

Green, P.J. (2003), Diversities of gifts, but the same spirit., *The Statistician* **52**, 423-438.

Green, P.J. and Mardia, K.V. (2006), Bayesian alignment using hierarchical models, with applications in protein bioinformatics, *Biometrika* **93**(2), 235–254.

Hamelryck, T., Kent, J.T. and Krogh, A. (2006), Sampling realistic protein conformations using local structural bias, *Computational Biology* **2**(9), 1121–1133.

Holm, L. and Sander, C. (1993), Protein structure comparison by alignment of distance matrices, *J. Mol. Biol.* **233**, 123–138.

Horgan, G.W., Creasey, A. and Fenton, B. (1992), Superimposing two dimensional gels to study genetic variation in malaria parasites, *Electrophoresis* **13**, 871–875.

Jonker, R. and Volgenant, A.A. (1987), Shortest augmenting path algorithm for dense and spare-linear assignment problems, *Computing* **38**, 325–340.

Kendall, D.G. (1984). Shape manifolds, Procrustean metrics and complex projective shapes. *Bulletin of London Mathematical Society*, **16**, 81–121.

Kent, J.T. (1994). The complex Bingham distribution and shape analysis. *Journal of the Royal Statistical Society, Series B*, **56**, 285-299.

Kent, J.T., Mardia, K.V. and Taylor, C.C. (2004), Matching problems for unlabelled configurations, *in* R. Aykroyd, S. Barber and K. Mardia, eds, Bioinformatics, Images, and Wavelets, Leeds University Press, 33–36.

Kent, J.T., Mardia, K.V. and Taylor, C.C. (2008), Bioinformatics and the problem of matching unlabelled configurations, *Submitted*.

Le, H.L. (1988). *Shape theory in flat and curved spaces, and shape densities with uniform generators*. Ph.D. thesis, University of Cambridge.

Leach, A.R. and Gillet, V.J. (2003), *An Introduction to Chemoinformatics*, Kluwer Academic Press, London.

Lesk, A.M. (2000), *Introduction to protein architecture*, Oxford University Press, Oxford.

Mardia, K.V. (2005), A vision of statistical bioinformatics, *in* S. Barber, P.D. Baxter, K.V. Mardia and R.E. Walls, eds, LASR2005 Proceedings, Leeds University Press, 9–20.

Mardia, K.V. (2007*a*), Discussion to Schmidler, 18. Listed here.

Mardia, K.V. (2007*b*), On some recent advancements in applied shape analysis and directional statistics, *in* S. Barber, P.D. Baxter and K.V. Mardia, eds, Systems Biology & Statistical Bioinformatics, Leeds University Press, 9–17.

Mardia, K.V. (2008), Holistic statistics and contemporary life sciences, *in* LASR Proceedings, Leeds University Press, 9–17.

Mardia, K.V. and Gilks, W. (2005), Meeting the statistical needs of 21st-century science, *Significance* **2**, 162–165.

Mardia, K.V., Hughes, G., Taylor, C.C. and Singh, H. (2008), Multivariate von mises distribution with applications to bioinformatics, *Canadian Journal of Statistics* **36**, 99–109.

Mardia, K.V. and Nyirongo, V.B. (2008), Simulating virtual protein C_α traces with applications, *J. Comp. Biology* **15**(9), 1221–1236.

Mardia, K.V. and Jupp, P.E. (2000), *Directional Statistics*, John Wiley and Sons Ltd, Chichester.

Mardia, K.V., Nyirongo, V.B., Green, P.J., Gold, N.D. and Westhead, D.R. (2007*a*), Bayesian refinement of protein functional site matching, *BMC Bioinformatics* **257**.

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

Marín, J.M. and Nieto, C. (2008), Spatial Matching of Multiple Configurations of Points with a Bioinformatics Application, *Communications in Statistics - Theory and Methods* **37**(12), 1977–1995.

Ruffieux, Y. and Green, P.J. (2008), Alignment of multiple configurations using hierarchical models. To appear in *Journal of Computational and Graphical Statistics*. Available at `http://www.stats.bris.ac.uk/~peter/papers/MAlign.pdf`.

Schmidler, S.C. (2007), Fast Bayesian shape matching using geometric algorithms, *in* J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith and W. M., eds, Bayesian Statistics, Oxford University Press, 1–20.

Wilkinson, D.J. (2007*a*), Discussion to Schmidler (2007), 13–17.

Wilkinson, D.J. (2007*b*), Bayesian methods in bioinformatics and computational systems biology, *Briefings in Bioinformatics* **8**(2), 109–116.

Appendix A.1 Broader context and background

Shape Analysis

Advances in data acquisition technology have led to the routine collection of geometrical information, and the study of the shape of objects has been increasingly important. With modern technology, locating points on objects is now often straightforward. Such points, typically on the outline or surface of the objects, can be loosely described as landmarks, and in this discussion, and indeed throughout the chapter, we treat objects as being represented by their landmarks, regarded as points in a euclidean space, usually $\mathcal{R}^2$ or $\mathcal{R}^3$.

What do we mean by 'shape'? The word is very commonly used in everyday language, usually referring to the appearance of an object. Mathematically, shape is all the geometrical information that remains when certain transformations are filtered out, that is, a point in shape space represents an equivalence class of objects, equivalent under transformations of the given kind. We are typically concerned with transformations such as translation and rotation, sometimes with uniform scale change and/or reflection, and less commonly with unequal scale change, and hence affine transformation, or even non-affine transformations such as non-parametric warping.

When we use the term *rigid shape analysis* we refer to the most important case in applications to bioinformatics, where the transformations in question are translations and rotations, that is, rigid-body motions. This case might more formally be termed *size and shape analysis*, or *the analysis of form*. The more common notion of shape in morphometrics is *similarity shape*, where uniform scale change is also allowed. In *reflection shape analysis*, the equivalence class also allows reflections.

To visualise the distinctions, consider right-angled triangles $\triangle ABC$ with sides in the ratio $AB : BC : CA = 3 : 4 : 5$. In reflection shape space, all such triangles are equivalent. For equivalence in similarity shape space, their vertices must be ordered in the same sense (clockwise or anticlockwise), and in rigid shape space, they must also have the same size.

Most theory and practice to date is concerned with the case of *labelled* shape analysis, where the landmarks defining an object are uniquely identified, so are regarded mathematically as an ordered set (or stacked as a matrix). But increasingly we see applications, including that dealt with in the present chapter, in which the landmarks are either not identified at all (the *unlabelled* case) or identification is incomplete, so that two different landmarks can have the same label (which we might call the *partially labelled* case). In the unlabelled case, the landmarks form an *unordered* set. To return to our 3:4:5 triangle, in unlabelled rigid shape analysis, the triangles with $AB = 3, BC = 4, CA = 5$ and with $AB = 4, BC = 5, CA = 3$ are equivalent, because for example the vertices identified with A in the two figures are not associated.

The foundation for similarity space analysis was laid by Kendall (1984). For mathematical representation, we can construct a shape space with an appropriate metric. The metric is a Procrustes distance for the Kendall shape space. For the form space (see Le, 1988) the appropriate distance is the RMSD (equation (4)). Note that in practice specific coordinate representations of shape have been useful, namely Bookstein coordinates (Bookstein, 1986) and Procrustes tangent coordinates (Kent, 1994). For further details, see for example, Dryden and Mardia (1998).

Alignment and matching problems such as those considered in this chapter extend unlabelled and partially-labelled shape analysis to embrace settings where the point configurations are supersets of those that can be matched, so that the analysis includes an element of selection of points to be matched as well as inference about the geometrical

transforations involved.

Appendix B.1 Model Formulation and Inference

Pairwise Model

Using concomitant information

When the points in each configuration are 'coloured', with the interpretation that like-coloured points are more likely to be matched than unlike-coloured ones, it is appropriate to use a modified likelihood that allows us to exploit such information. Let the colours for the x and y points be $\{r_j^{\mathrm{x}}, j = 1, 2, \ldots, m\}$ and $\{r_k^{\mathrm{y}}, k = 1, 2, \ldots, n\}$ respectively. The hidden-point model is augmented to generate the point colours, as follows. Independently for each hidden point, with probability $(1 - p_{\mathrm{x}} - p_{\mathrm{y}} - \rho p_{\mathrm{x}} p_{\mathrm{y}})$ we observe neither x nor y point, as before. With probabilities $p_{\mathrm{x}}\pi_r^{\mathrm{x}}$ and $p_{\mathrm{y}}\pi_r^{\mathrm{y}}$, respectively, we observe only an x or y point, with colour r from an appropriate finite set. With probability

$$\rho p_{\mathrm{x}} p_{\mathrm{y}} \pi_r^{\mathrm{x}} \pi_s^{\mathrm{y}} \exp(\gamma I[r = s] + \delta I[r \neq s]),$$

where $I[\cdot]$ is an indicator function, we observe an x point coloured r and a y point coloured s. Our original likelihood is equivalent to the case $\gamma = \delta = 0$, where colours are independent and so carry no information about matching. If γ and δ increase, then matches are more probable, a posteriori, and, if $\gamma > \delta$, matches between like-coloured points are more likely than those between unlike-coloured ones. The case $\delta \to -\infty$ allows the prohibition of matches between unlike-coloured points, a feature that might be adapted to other contexts such as the matching of shapes with given landmarks.

In implementation of this modified likelihood, the Markov chain Monte Carlo acceptance ratios derived for M can be easily modified.

Other, more complicated, colouring distributions where the log probability can be expressed linearly in entries of M can be handled similarly.

Continuous concomitant information can be incorporated in our statistical models such as incorporating van der Waal radii. Such models will be very similar in character as above with obvious modifications e.g. by using pairwise interaction potentials instead of indicator functions.

Loss functions and a point estimate of M

The output from the Markov chain Monte Carlo sampler derived above, once equilibrated, is a sample from the posterior distribution. As always with sample-based computation, this provides an extremely flexible basis for reporting aspects of the full joint posterior that are of interest.

We consider loss functions $L(M, \widehat{M})$ that penalise different kinds of error and do so cumulatively. The simplest of these are additive over pairs (j, k). Suppose that the loss when $M_{jk} = a$ and $\widehat{M}_{jk} = b$, for $a, b = 0, 1$, is ℓ_{ab}; we set $\ell_{00} = \ell_{11} = 0$. For example, ℓ_{01} is the loss associated with declaring a match between x_j and y_k when there is really none, that is, a 'false positive'. Then

$$E\{L(M, \widehat{M})|x, y\} = -(\ell_{10} + \ell_{01}) \sum_{j,k:\widehat{M}_{jk}=1} (p_{jk} - K),$$

where

$$K = \ell_{01}/(\ell_{10} + \ell_{01}),$$

and $p_{jk} = \mathrm{pr}(M_{jk} = 1|x, y)$ is the posterior probability that (j, k) is a match, which is estimated from a Markov chain Monte Carlo run by the empirical frequency of this match. Thus, provided that $\ell_{10} + \ell_{01} > 0$ and $\ell_{01} > 0$, as is natural, the optimal estimate is that maximising the sum of marginal posterior probabilities of the declared matches $\sum_{j,k:\widehat{M}_{jk}=1} p_{jk}$, penalised by a multiple K times the number of matches. The optimal match therefore depends only through the cost ratio K. If false positive and false negative matches are equally undesirable, one can simply choose $K = 0.5$.

Computation of the optimal match $\widehat{M}$ would be trivial but for the constraint that there can be at most one positive entry in each row and column of the array. This weighted bipartite matching problem is equivalent to a mathematical programming assignment problem, and can be solved by special-purpose or general LP methods; (see Burkard and Cela, 1999).

For problems of modest size, the optimal match can be found by informal heuristic methods. These may not even be necessary, especially if K is not too small. In particular, it is immediate that, if the set of all (j, k) pairs for which $p_{jk} > K$ includes no duplicated j or k value, the optimal $\widehat{M}$ consists of precisely these pairs. For aligning large size of protein chains lpSolve needs to be replaced by linearass (Jonker and Volgenant, 1987).

Appendix B.2 Model Implementation

Sampling the posterior distribution for pairwise alignment

It is straightforward to update conditionally continuous variables. For updating M conditionally, we need some new ways.

The matching matrix M is updated in detailed balance using Metropolis-Hastings moves that only propose changes to a few entries: the number of matches $L = \sum_{j,k} M_{jk}$ can only increase or decrease by 1 at a time, or stay the same. The possible changes are as follows:

(a) adding a match, which changes one entry M_{jk} from 0 to 1;

(b) deleting a match, which changes one entry M_{jk} from 1 to 0;

(c) switching a match, which simultaneously changes one entry from 0 to 1 and another in the same row or column from 1 to 0.

The proposal proceeds as follows. First a uniform random choice is made from all the $m+n$ data points $x_1, x_2, \ldots, x_m, y_1, y_2, \ldots, y_n$. Suppose without loss of generality, by the symmetry of the set-up, that an x is chosen, say x_j. There are two possibilities: either x_j is currently matched, in that there is some k such that $M_{jk} = 1$, or not, in that there is no such k. This depends on $p^\star$; if x_j is matched to y_k, with probability $p^\star$ it is proposed deleting the match, and with probability $1 - p^\star$ we propose switching it from y_k to $y_{k'}$, where k' is drawn uniformly at random from the currently unmatched y points. On the other hand, if x_j is not currently matched, it is proposed adding a match between x_j and a y_k, where again k is drawn uniformly at random from the currently unmatched y points.

The acceptance probabilities for these three possibilities are easily derived (see Green and Mardia, 2006)

Note that this procedure bypasses the reversible jump.

Multimodality

The issue of multimodality is a challenging issue, as we point out in Green and Mardia (2006). The MCMC samplers used here are very simple (but adequate for the presented

examples), and there is a vast literature on more powerful methods that we have not yet brought to bear; this is one of the areas that needs exploring. See also discussion by Wilkinson (2007*a*).

Sampling the posterior distribution for multiple alignment

With our conditionally conjugate priors, we can update the parameters $\tau^{(c)}, A^{(c)}$, and σ^2 using a Gibbs move, as in the pairwise implementation. Generalising the updating of the matches to the multi-configuration context is less obvious. Write

$$\mathcal{M} = \left\{ \left(t_1^1, t_2^1, \ldots, t_C^1\right), \left(t_1^2, t_2^2, \ldots, t_C^2\right), \ldots, \left(t_1^K, t_2^K, \ldots, t_C^K\right) \right\}.$$

Each C-tuple $(t_1^k, t_2^k \ldots, t_C^k)$ represents a match, t_c^k being the index of the point from the $x^{(c)}$ configuration involved in the match. If a given configuration is not involved in the match, a '$-$' flag is inserted at the appropriate position. For instance, if $C = 3$ the 3-tuple $(2, 4, 1)$ refers to a match between $x_2^{(1)}, x_4^{(2)}$ and $x_1^{(3)}$, while $(-, 2, 1)$ is a match between $x_2^{(2)}$ and $x_1^{(3)}$, with no $x^{(1)}$–point involved. We also include unmatched points in this list: $(1, -, -)$ indicates that $x_1^{(1)}$ is unmatched, for example.

Suppose that $\mathcal{M}$ is the current list of matches in the MCMC algorithm. We define a jump proposal proceeds as follows:

- with probability q we choose to *split* a C-tuple; in this case we draw an element uniformly at random in the list $\mathcal{M}$.
  - If the C-tuple drawn corresponds to an unmatched point, we do nothing;
  - otherwise we split it into two C-tuples at random; for instance $(2, 3, 1)$ can be split into $(2, -, -)$ and $(-, 3, 1)$.
- With probability $1 - q$ we choose to *merge* two C-tuples; in this case we select two distinct elements uniformly at random from $\mathcal{M}$.
  - If the two C-tuples drawn contain a common configuration, e.g. $(j_1, k, -)$ and $(j_2, -, -)$, then we do nothing;
  - otherwise we merge the C-tuples, for example $(j, k, -)$ and $(-, -, l)$ become (j, k, l), while $(-, k, -)$ and $(-, -, l)$ become $(-, k, l)$.

This defines a Metropolis-Hastings jump, and its acceptance probability can be readily worked out from (7).

Appendix B.3 Other Data Sources

- PDB databank: This is a web resource for protein structure data (`http://www.rcsb.org/pdb/`). The RCSB PDB also provides a variety of tools and resources for studying structures of biological macromolecules and their relationship to sequence, function and disease.
  - PDBsum is a value added tool for understanding protein structure. Another important tool is SWISS-MODEL server which uses *HMM* for homology modelling to identify structural homologs of a protein sequence.
  - There are also tools for displaying 3-dimensional structures e.g. RasMol, JMol, KiNG, WebMol, MBT SimpleView, MBT Protein Workshop and QuickPDB (JMol is a new version of RasMol but is JAVA based).
- SitesBase: Resource for data on active sites. The database holds pre-compiled information about structural similarities between known ligand binding sites found in the Protein Data Bank. These similarities can be analysed in molecular recognition applications and protein structure-function relationships.
  `http://www.bioinformatics.leeds.ac.uk/sb`.
- NPACI/NBCR resource: A database and tools for 3-dimensional protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm.
  `http://cl.sdsc.edu/ce.html`.
- The Dali Database: This database is based on exhaustive, all-against-all 3-dimensional structure comparison of protein structures in the Protein Data Bank. Alignments are automatically maintained and regularly updated using the Dali search engine.
  `http://ekhidna.biocenter.helsinki.fi/dali_server/`
- SCOP: Aims to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. As such, it provides a broad survey of all known protein folds, detailed information about the close relatives of any particular protein, and a framework for future research and classification.
  `http://scop.mrc-lmb.cam.ac.uk/scop/`
- CATH: Describes the gross orientation of secondary structures, independent of connectivities and is assigned manually. The topology level clusters structures into fold groups according to their topological connections and numbers of secondary structures. The homologous superfamilies cluster proteins with highly similar structures and functions. The assignments of structures to fold groups and homologous superfamilies are made by sequence and structure comparisons. `http://www.cathdb.info/`
- BLAST: Basic Local Alignment Search Tool, or BLAST, is an algorithm for comparing primary biological sequence information, such as the amino-acid sequences of different proteins. A BLAST search enables a researcher to compare a query sequence with a library or database of sequences, and identify library sequences that resemble the query sequence above a certain threshold. `http://blast.ncbi.nlm.nih.gov/Blast.cgi`