Hierarchical Bayesian Modelling of Pharmacophores in Bioinformatics

Kanti V. Mardia*

Department of Statistics, The University of Leeds, Leeds, LS2 9JT, UK

**email:* k.v.mardia@maths.leeds.ac.uk

and

Vysaul B. Nyirongo*

Department of Statistics, The University of Leeds, Leeds, LS2 9JT, UK

**email:* stavn@maths.leeds.ac.uk

and

Christopher J. Fallaize*

Department of Statistics, The University of Leeds, Leeds, LS2 9JT, UK

**email:* chrisf@maths.leeds.ac.uk

and

Stuart Barber*

Department of Statistics, The University of Leeds, Leeds, LS2 9JT, UK

**email:* stuart@maths.leeds.ac.uk

and

BIOMETRICS 00, 1–18 DOI: 10.1111/j.1541-0420.2005.00454.x
Month 2009

Richard M. Jackson*

Institute of Molecular and Cellular Biology, The University of Leeds, Leeds, LS2 9JT, UK

**email:* r.m.jackson@leeds.ac.uk

SUMMARY: One of the key ingredients in drug discovery is the derivation of conceptual templates called pharmacophores. A pharmacophore model characterises the physico-chemical properties common to all active molecules, called ligands, bound to a particular protein receptor, together with their relative spatial arrangement. Motivated by this important application, we develop a Bayesian hierarchical model for the derivation of pharmacophore templates from multiple configurations of point sets, partially labelled by the atom type of each point. The model is implemented through a multi-stage template hunting algorithm which produces a series of templates that capture the geometrical relationship of atoms matched across multiple configurations. Chemical information is incorporated by distinguishing between atoms of different elements, whereby different elements are less likely to be matched than atoms of the same element. We illustrate our method through examples of deriving templates from sets of ligands which all bind structurally related protein active sites. The resulting templates are considered to be plausible by experts with respect to the chemical affinity of the subsets of molecules used to derive them.

KEY WORDS: Alignment; ligands; MCMC; pharmacophore; shape analysis; spatial matching; template.

1. Introduction

One of the key ingredients in drug discovery is the derivation of conceptual templates called pharmacophores. A pharmacophore model is a specific three-dimensional map of chemical properties common to active conformations of a set of small molecules, known as ligands, that exhibit a particular biological activity. A pharmacophore model can be generated from three-dimensional structural data describing ligands and their interaction with a particular protein receptor site. Currently, this is often done manually by inspection and expert judgement, see for example Rella et al. (2006). We note that methods for the multiple alignment of configurations have been proposed by, for example, Ruffieux and Green (2008) and Dryden, Hirst and Melville (2007), but these methods are not specifically tailored to producing pharmacophore templates from matched points. Hence, there is a need to develop a statistical methodology which simultaneously enables the automated identification of pharmacophore models and the quantification of their plausibility. For more details on the pharmacophore concept, see Leach and Gillet (2003, Chapter 3).

It is common to represent structures of protein-ligand complexes as configurations of points in $\mathbb{R}^3$, with each point representing the location of an individual atom. Pharmacophore identification is therefore reduced directly to the problem of finding points common to a set of configurations. Motivated by this, we develop a hierarchical model for the derivation of pharmacophore templates. The method identifies common matched points from multiple configurations, or subsets of them, and builds a hierarchy of templates capturing the geometry of the matched points. We also consider the chemical plausibility of the templates, through the use of chemical information to distinguish between atoms of different elements, with the interpretation that different elements are less likely to be matched than atoms of the same element type. This ensures that the resulting templates are sensible with regards to their chemical properties, as well as their geometry.

Within our model, we require the use of a method for the pairwise alignment of two configurations. Here we use the pairwise alignment method described by Green and Mardia (2006), which provides us with many of the ingredients needed to implement our strategy. An alternative method could easily be substituted; all that we require is a method that estimates which atoms match and the corresponding probabilities, allowing a "score" rating the overall quality of matching between two ligands to be computed. We then use the output from these alignments within a multi-stage algorithm for building templates, which requires the use of a scoring function for discriminating between various pairwise alignments at each stage. The templates are formed hierarchially, successively merging configurations or previously formed templates, using only the common matched points identified from the pairwise alignments. Our proposed algorithm is capable of identifying multiple subsets of configurations and outputs templates representing the matched points in each.

An outline of this article is as follows. We describe the model behind our methodology in Section 2. In Section 3, we consider the implementation of our model and outline an example iteration of the algorithm for a fixed number of configurations. In Section 4 we consider two applications of our method to finding points common to subsets of ligands which are bound to related protein active sites. Finally, we discuss the proposed methodology in Section 5.

2. Methodology

We consider using data obtained from the multiple alignment of protein binding sites to produce pharmacophore templates from a set of ligands. The data is in the form of ligands reduced to point configurations in three-dimensional space, with each point partially labelled by element type of the atom at that point. In this paper, we assume we have three-dimensional data but the method can easily be extended to $d \neq 3$ dimensions. We have ligand configurations x_i: $i = 1, \ldots, C$ of sizes n_i. That is, configuration i contains n_i points (atoms). The aim is to construct a template, μ_0 say, comprising of n_0 atoms, where n_0 is unknown. We set

out to identify n_0 common points in a set of configurations I, where $I \subseteq \{x_i\}$ and $\{x_i\}$ is the set of all configurations. Candidate templates are formed from pairwise alignments between individual configurations and/or previously constructed templates, and these are evaluated with a scoring function which we use to select the best candidate at each stage. Hence, we require the use of a method for the pairwise alignment of two configurations. Here we use the method described by Green and Mardia (2006), the output from which we can use to build candidate templates at each stage and evaluate them according to our scoring function. We will henceforth refer to this method as the Green-Mardia (GM) algorithm. Below, we first briefly describe the GM algorithm, before describing our proposal for a hierarchical template model, which we will refer to as the HT algorithm.

2.1 *The GM pairwise alignment algorithm*

Green and Mardia (2006) describe the pairwise alignment of two configurations using a fully Bayesian aproach. Consider aligning a pair of configurations x and y under rigid body transformations. Denote the j^{th} atom in the x configuration by x_j where $j = 1, \ldots, m$. Similarly, y_k denotes the k^{th} atom in the y configuration where $k = 1, \ldots, n$. Let A and τ denote the rotation matrix and translation vector to bring y into alignment with x. Furthermore denote prior distributions on these parameters by $p(A)$ and $p(\tau)$. We denote the prior for σ, parameterising noise in atomic positions for x and y coordinates, by $p(\sigma)$. The joint posterior distribution for the model is

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

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, where

$$M_{jk} = \begin{cases} 1 & \text{if } x_j \text{ corresponds to } y_k \\ 0 & \text{otherwise.} \end{cases} \qquad (2)$$

We also use the priors

$$A \sim \text{uniform}, \quad \tau \sim N(0, \sigma_\tau^2 I_3), \quad \sigma^{-2} \sim \Gamma(\alpha, \beta). \tag{3}$$

Euler angles, θ_{12}, θ_{13} and θ_{23} say, are used to parameterise the rotation matrix as a product of elementary rotations, such that

$$A = A_{12}(\theta_{12})A_{13}(\theta_{13})A_{23}(\theta_{23}),$$

where $-\pi < \theta_{12}, \theta_{23} < \pi$ and $-\pi/2 < \theta_{13} < \pi/2$ (see Green and Mardia, 2006). The uniform measure is then $\cos\theta_{13} d\theta_{12} d\theta_{13} d\theta_{23}$.

A point estimate of M, $\widehat{M}$, is found by minimising the point-wise error rates $P(\widehat{M}_{jk} = 1|M_{jk} = 0)$ and $P(\widehat{M}_{jk} = 0|M_{jk} = 1)$ and is controlled by the cost ratio, K, of falsely matching points. The posterior probability that the pair of points (j, 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 $\widehat{M}$ is a solution to a "linear assignment" problem with cost matrix $(p_{jk} - K)$. A standard linear assignment program (lpsolve, Berkelaar, 1996) is then used to find $\widehat{M}$, with the cost matrix $(p_{jk} - K)_+$.

2.2 *The Hierarchical Templates (HT) model*

We denote the j^{th} atom in the i^{th} configuration by x_{ij}, $j = 1, \ldots, n_i$. Similarly, the j'^{th} atom in the template is denoted by $\mu_{0j'}$. Let $M_i = (M_{ijj'})$ be the matching matrix for the i^{th} ligand and μ_0, where

$$M_{ijj'} = \begin{cases} 1 & \text{if } x_{ij} \text{ corresponds to } \mu_{0j'} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Also let A_i and τ_i, $i = 1, \ldots, |I|$ be rotation matrices and translation vectors to register x_i with respect to μ_0. For given $\mu_{0j'}$, and for each j' such that $M_{ijj'} = 1$, we assume the likelihood

$$A_i x_{ij} + \tau_i | \mu_{0j'} \sim N(\mu_{0j'}, \sigma^2 I_3). \tag{5}$$

We use a Bayesian formulation where the priors for A_i, τ_i and σ^{-2} are as in (3) and μ_0 has a uniform prior. We assume an exchangeable prior for $M_{ijj'}$ with the geometric probability distribution

$$p(M_{ijj'}) \propto (\kappa)^{n_0}, \tag{6}$$

where κ is a matching propensity parameter and n_0 is the number of points in the template configuration μ_0.

The joint log posterior for the model is

$$\begin{aligned} -\tfrac{1}{2\sigma^2} \sum_{i=1}^{|I|} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_0} M_{ijj'} ||A_i x_{ij} + \tau_i - \mu_{0j'}||^2 - 3|I| \log(\sqrt{2}\sigma) \\ -|I| \log(\kappa) - \beta/\sigma - (\alpha - 1) \log \sigma + \log(\cos \theta_{13}). \end{aligned} \tag{7}$$

For given $|I|$ there can be multiple configuration sets maximising the joint log posterior. It should be noted that the model is identifiable only up to the equivalence class of the form of μ_0 i.e. $A\mu_0 + \tau$, where A is a rotation matrix and τ a translation vector. Given μ_0 and I, the conditional distribution of all the parameters (A_i, τ_i, σ and M) can be obtained in a hierarchical pairwise method, using MCMC as follows:

We have explicit full conditionals for τ and σ and these parameters are updated using Gibbs sampling; $M_{ijj'}$ is updated using Metropolis-Hastings (Green and Mardia, 2006). With Euler angles θ_{12}, θ_{13} and θ_{23} parameterising the rotation A, θ_{12} and θ_{23} are updated using Gibbs sampling from their full conditional von Mises distributions. Metropolis-Hastings is used to update θ_{13}.

2.3 *Estimating* μ_0

We note from equation (7) that $\log p(\mu_0|\text{rest})$ is (except for a constant)

$$-\tfrac{1}{2\sigma^2} \sum_{i=1}^{|I|} \sum_{j=1}^{n_i} \sum_{j'=1}^{n_0} M_{ijj'} ||A_i x_{ij} + \tau_i - \mu_{0j'}||^2. \tag{8}$$

Thus a point estimate $\hat{\mu}_0$ of μ_0, where $\mu_0 = (\mu_{01}, \ldots, \mu_{0n_0})^T$, is given by

$$\hat{\mu}_{0j'} = \left(\sum_{i=1}^{|I|} \sum_{j=1}^{n_i} M_{ijj'} ||A_i x_{ij} + \tau_i|| \right) \left(\sum_{i=1}^{|I|} \sum_{j=1}^{n_i} M_{ijj'} \right)^{-1} \tag{9}$$

$j' = 1, \ldots, n_0$, up to equivalence class of form.

Hence we can estimate $\hat{\mu}_0$ given point estimates of all other parameters. We now assume that we can compute goodness of fit statistics, S, for selecting "optimal" estimates of $\hat{\mu}_0$ using only these pairwise estimates of the other parameters from any pairwise alignment method giving matching matrices M and the corresponding posterior probability matrix of matches, P, as well as estimates of the transformation parameters A and τ, such as GM and EM (Kent, Mardia and Taylor, 2004) algorithms. Our goodness of fit statistics S then depend on M and P in building μ_0.

We first consider a simple example with 3 configurations to construct stage by stage estimates of μ_0. Suppose now that for pairs of configurations $(1, 2)$, $(1, 3)$ and $(2, 3)$ we obtain matching matrices M and posterior probability matrices $P(1, 2)$, $P(1, 3)$, $P(2, 3)$, where the $(i, j)^{th}$ element of $P(1, 2)$ is the estimated posterior probability of matching the i^{th} point of x_1 to the j^{th} point of x_2 and so on. Now say S selects the subset of x_1 and x_2 to be the "best". Then we can obtain the estimate $\hat{\mu}_0^{(1)}$ from equation (9), as all the parameters are now known. We now apply pairwise comparison of $\hat{\mu}_0^{(1)}$ and x_3 and calculate the statistic S. If S rejects this new subset, we keep $\hat{\mu}_0^{(1)}$ as our estimate, otherwise we calculate $\hat{\mu}_0^{(2)}$ from equation (10) as the new estimate of μ_0 using the matching subset of $\hat{\mu}_0^{(1)}$ and x_3. In this case these are the only options but even for 4 configurations, there can be many options, including the following multiple estimates for μ_0.

(1) $\hat{\mu}_0$ may be obtained using all 4 configurations.

(2) $\hat{\mu}_0$ may include only 3 configurations such as $(1, 2, 3)$ and $(1, 3, 4)$.

(3) $\hat{\mu}_0$ may include only 2 configurations such as $(1, 2)$ or $(3, 4)$.

We give full details of a hypothetical example using six configurations in the supplementary material.

3. Implementation of the HT Model

We consider a point estimate for μ_0 within the equivalence class of form $[\mu_0]$, i.e. $A\hat{\mu}_0 + \tau$ is equivalent to $\hat{\mu}_0$. We propose removing this non-identifiability for μ_0 by taking μ_0 to be in the configuration space of one of the observed configurations and use a MAP estimator. Since our priors for A and τ in (3) are symmetric, it does not matter which configuration is used as reference. We estimate model parameters for hierarchical templates conditional on μ_0, which requires pairwise alignment involving μ_0. That is we need to embed a pairwise alignment algorithm such as the EM algorithm (Hancock and Cross, 1998; Luo and Hancock, 2001; Kent et al., 2004) or Bayesian alignment (Green and Mardia, 2006; Dryden et al., 2007; Schmidler, 2007) within a hierarchical structure. Here we use the GM algorithm which is a fully Bayesian pairwise alignment methodology giving all the ingredients for comprehensive Bayesian inference including the log-posterior, matching probabilities and point estimates that can be used to summarise pairwise alignments.

3.1 *Mechanism of the HT algorithm*

The HT algorithm starts by considering each single configuration as a template and, taking a bottom-up approach, successively merges pairs of templates to form new templates. At each iteration, all pairwise alignments of items in the list of templates are evaluated and the best pair, according to some criteria, are merged to form a new template. The new template is added to the list, the two merged items are removed and the process is repeated. The algorithm continues until no pairwise alignment satisfies the merging criteria. For computational efficiency, details of pairwise alignments involving items which are not merged are kept, so they do not have to be re-evaluated at the next iteration.

Here we have used a template merging criterion based on the geometric mean of matching probabilities for declared matches, $\mathcal{G} = \prod p_{jk}^{(1/n_0)}$, where n_0 is the number of matches given by the estimate $\hat{M}$ of the matching matrix M, which on the log scale is equal to $n_0^{-1} \sum \log p_{jk}$. The merging criteria we have used is to select the pair of configurations (a, b) such that

$$(a, b) = \arg\max_{a,b} \mathcal{G}_{a,b}, \tag{10}$$

where $\mathcal{G}_{a,b}$ is the geometric mean of the pairwise alignment between templates a and b. Additionally, we impose the threshold values g_{min} and n_{min}, so that the conditions $\mathcal{G} \geqslant g_{min}$ and $n_0 \geqslant n_{min}$ must be satisfied for a merge to be accepted. Hence, if no pair satisifes this criteria then the algorithm will terminate. An outline of an example iteration of the algorithm is given below. In the supplementary material, we give full details of a possible algorithm flow for a hypothetical example involving six configurations.

3.2 *Outline Iteration*

We now outline an example iteration for $C = 11$ configurations, $x_1 \dots x_{11}$. We denote our list of templates by $\mathcal{X}$. Our initial list of templates is just the 11 configurations, so $\mathcal{X} = \{x_1 \dots x_{11}\}$. The choice of $C = 11$ is motivated by the real examples we consider in section 4.

Let r denote the iteration number for the algorithm. At the r^{th} iteration we denote a vector of matching probabilities for the optimal pairwise alignment between the i^{th} and i'^{th} configurations or templates by $P^r_{(i,i')} = (p^r_1, \dots, p^r_{n_0})$, where $n_0 = n^r_{0(i,i')}$ is the number of matched points from the optimal pairwise alignment at iteration r. Let $\mathcal{G}^r_{(i,i')} = \prod_{l=1}^{n_0} (p^r_l)^{1/n_0}$, the geometric mean of matching probabilities. Example iterations would proceed as follows:

(1) Consider all pairs of configurations. There are ${}^{11}C_2 = 55$ pairwise matches. For each pairwise match, between configurations x_i and $x_{i'}$ say, obtain the number of matching atoms $n^1_{0(i,i')}$, the corresponding matching probabilities $P^1_{(i,i')}$ and geometric mean $\mathcal{G}^1_{(i,i')}$.

(2) Merge matching configurations with the highest $\mathcal{G}^1_{(i,i')}$ meeting the merge criteria. Say

this pair is $(1, 2)$ with $j = 1, \ldots, n_0$ corresponding points. Form a template denoted by $T_{ii'}$ with the coordinates for point j given by mean coordinates of the j^{th} corresponding points in the registered configurations. Note that the new template consists of n_0 points, formed using only the n_0 matched points from the pairwise alignment.

(3) Configurations 1 and 2 are removed from the set of configurations as they are a subset of a newly formed template; the new template $T_{ii'}$ is added to the list of configurations as x_{12}, so $\mathcal{X} = \{x_3, x_4, \ldots, x_{12}\}$.

(4) Match the newly formed template, T_{12}, against configurations $x_3, \ldots, x_{11}$. Obtain number of matching atoms $n^2_{0(12,i')}$, matching probabilities $P^2_{(12,i')}$ and $\mathcal{G}^2_{(12,i')}$, $i' = 3, \ldots, 11$.

(5) The second iteration considers a set of pairwise alignments among configurations $x_3, \ldots, x_{11}$ as well as new pairwise alignments involving T_{12} and configurations $x_3, \ldots, x_{11}$ evaluated in step (4), i.e we begin the cycle again in step (2). Note that for previously considered pairs (i, i') that have not been merged, $P^r_{(i,i')} = P^{r-1}_{(i,i')}$ and $\mathcal{G}^r_{(i,i')} = \mathcal{G}^{r-1}_{(i,i')}$. These are being tracked together with the number of matches, $n^r_{0(i,i')}$, so we do not need to recompute them.

For following iterations only pairs of configurations or templates that are not merged in previous iterations are considered to be merged to form a new template. Successive templates are formed hierarchically whereby the coordinates for the template involving a set of configurations, $I \subseteq \{x_1, x_2, \ldots, x_{11}\}$ are $\hat{\mu}_I = \frac{1}{|I|} \sum_{i' \in I} x_{i'}$.

Plausible pharmacophores are templates consisting of say $q \geqslant 2$ configurations and $n_0 \geqslant 3$ atoms.

The algorithm may output one template, containing matched points across some or all configurations, or multiple templates derived from matched points from different subsets of configurations. We do not allow overlap between subsets, so each configuration can only contribute to at most one template. Part of our strategy is analogous to that of an agglom-

erative clustering algorithm, (see, for example, Mardia, Kent and Bibby, 1979, pp. 371-373), but with some important differences. Our objects are point configurations rather than single points, and we require a similarity measure between pairs of configurations, or templates. Additionally, our similarity measure is dynamic, in the sense that new similarity measures must be calculated from pairwise alignments between a newly formed template and all existing templates. Therefore, at each stage we have an updated list of similarity measures, consisting of all measures previously calculated as well as the new measures obtained from pairwise alignments involving the most recently formed template.

The algorithm has computational complexity $\mathcal{O}(C^2)$ for the number of pairwise alignments performed. Recall that we denote the iteration number of the algorithm by r, where $r \geqslant 1$. We begin with a finite number C of configurations. Note that for the first step, $r = 1$, we need to perform $\frac{C(C-1)}{2}$ pairwise alignments. For each subsequent step $r \geqslant 2$, we have $C - r + 1$ configurations, or templates, and we need only perform $C - r$ pairwise alignments between the template formed at step $r - 1$ and the other templates in our list. The total number of iterations is at most C, since after C iterations we would have only one template remaining. Thus the algorithm has the polynomial cost complexity of $\mathcal{O}(C^2)$. Additionally, an extra layer of cost complexity is added for each pairwise alignment performed, which depends on the sizes of the configurations.

3.3 *Restricted Transformations*

There are situations where we may wish impose restrictions on transformation parameters (rotation matrices A and translation vectors τ) when identifying the template from pre-aligned ligands. For example we might

(1) allow only small deviations of A_i from identity matrix I_3 and small deviations for τ_i from the zero vector;

(2) set $A_i = I_3$ and $\tau_i = 0$ to prohibit any degree of transformation.

These restrictions would be important in situations where alternative geometrical alignments are to be avoided, such as in the presence of ring-like structures or when the ligands have already been aligned. In our application described in Section 4 we do not consider transformation in aligning the ligands when identifying plausible pharmacophores as they are pre-aligned in some meaningful sense. However, there may be occasions when one would not wish to restrict transformations, such as when searching a database of compounds for matches to a pharmacophore template. In this case, the compounds would not necessarily have any meaningful pre-alignment.

3.4 *Atom type information*

We can consider that points are "coloured" with the interpretation that like-coloured points are more likely to be matched than unlike-coloured ones. In the context of searching for commonality between ligands, one might take the atom elements (carbon, nitrogen, etc) as the colour information. We can specify the matching propensity parameter κ as a function of concomitant information like atom types to parameterise the tendency *a priori* for points to be matched. In pairwise alignment, we have either a colour match or mismatch. With multiple alignment there can be many sophisticated ways of scoring colour-mismatching, as the number of different colours in a match can range from one to say $S > 2$, the total number of colours. Here we consider a simple way whereby we have binary categorization as follows:

- matches with all atoms having the same type,
- matches with at least one atom type different.

A priori, matching probabilities are proportional to $\exp(\gamma)$ for same colour matches and $\exp(\delta)$ for different colour matches, where γ and δ are specified "award/penalty" parameters for matching or mismatching colours, where $\gamma > \delta$. For example, setting $\gamma = 1.0$ and $\delta = -0.5$ awards colour matching twice as much as penalizing colour mismatching. Note that $\delta \longrightarrow -\infty$ prohibits matching different colours, for instance element types.

4. Application

We have two sets of eleven ligand configurations from multiple structural alignments of ligand binding sites in SitesBase (Gold and Jackson, 2006), one for a protein kinase and one for trypsin. SitesBase entries were automatically formed from the protein data bank, or PDB (Berman et al., 2000), by locating the local protein environment (amino acids within 5Å) around bound ligands (identified by PDB HETATM records, as described by Gold and Jackson (2006)). For simplicity we will refer to the set of ligands from a series of protein kinases as 1ATP and label these $1-11$. Similarly, we will refer to the set of ligands from a series of trypsin-ligand bound structures as 3PTB and label these $12-22$. These refer to two sets of ligand binding sites which in each case are superimposed on a single site. In the case of 1ATP, ten more distantly related protein kinase binding sites were superimposed on a subunit of protein kinase 1ATP. They contain a diverse set of kinase inhibitors. In the case of 3PTB, ten more trypsin-ligand bound complexes were superimposed on the site of trypsin bound to benzamidine, 3PTB. The ligands are shown in Figure 1 and the sizes (number of atoms) of each ligand are given below in Tables 1 and 2.

[Table 1 about here.]

[Table 2 about here.]

Finding a pharmacophore model for the 1ATP ligands by manual inspection is considered difficult even for an experienced biologist, while finding a pharmacophore model for the 3PTB ligands is an easier task. Using our HT algorithm we find three subsets of conformations for the 1ATP data and two subsets of configurations for the 3PTB data. From these we can obtain templates capturing the geometry of the matched points in each subset. Since the ligands are pre-aligned by their binding to other proteins, we set $A = I_3$ and $\tau = 0$ to prohibit spatial transformation. Allowing unrestricted transformation gave either the same (1ATP) or very similar (3PTB) results. With slight tuning of the hyperparameters for σ^{-2}

we obtain the same results for the 3PTB case as well. In the following two applications, we keep the hyperparameters for σ^{-2} fixed throughout, with $\alpha = 1$ and $\beta = 5$.

[Figure 1 about here.]

4.1 *1ATP ligands*

We denote 1ATP ligands by numbers $1, 2, \ldots, 11$. Using the HT algorithm we identify common atoms in three different subsets, each consisting of three configurations. Here we have used the geometric mean threshold $g_{min} = 0.5$ and the cost ratio $K = 0.1$, as experience showed these values provide a good balance between allowing templates to merge but preventing the final templates becoming too general, in agreement with expert opinion. We also impose the restriction $n_{min} \geqslant 3$, meaning acceptable templates must consist of at least 3 atoms.

Table 3 shows the configurations contributing to each of the templates found and the number of atoms they have in common. Figure 2 shows the geometry of the configurations. We use colouring information in order to distinguish between atoms of different elements, as described in section 3.4. It is more sensible from a chemical viewpoint to match points of the same element, hence we discourage matches between different element types. It is still possible to match elements of different types if their interpoint distances are relatively small. Here we use the values $\gamma = 0$ and $\delta = -20$ to discourage matches between elements of different types; no matches between different elements were found in this case. If we do not use colouring information, the three subsets of configurations found are the same as those given in Table 3, but there are differences in a small number of the individual matches between points. We find one match between different elements in template A and three matches between different elements if templates B and C. There are two fewer points in template C when colouring information is used, and the same number of points in templates

A and B. We give details of the results obtained without using colouring information in the supplementary material.

[Table 3 about here.]

[Figure 2 about here.]

4.2 *3PTB ligands*

We denote the eleven 3PTB ligands by numbers $12, 13, \ldots, 22$. The HT algorithm identifies 2 subsets of configurations. We have used the values $g_{min} = 0.5$, $K = 0.1$ and $n_{min} = 3$ as before. Once again, we use colouring information to distinguish between different element types, with $\delta = -20$. The configurations contributing to each template are shown in Table 4 and their geometry in Figure 3. The details of the results obtained when colouring information is not used are given in the supplementary material. We find no matches between atoms of different elements, regardless of whether we use colouring information or not.

[Table 4 about here.]

[Figure 3 about here.]

5. Discussion

In this paper we have proposed a fast method for aligning multiple configurations of unlabelled point sets and identifying common matched points across all configurations, or subsets of them, in order to derive templates capturing the geometry of the matched points. Our method is able to identify multiple subsets of configurations. In this sense, part of the implementation strategy is analogous to an agglomerative clustering algorithm, but with some important differences. The algorithm is implemented via a multi-stage pairwise alignment approach, so at each stage we have an updated "similarity measure" based on a criteria calculated from the pairwise matching probabilities given by the probability matrices P, the

set of which is updated at each stage to include alignments involving newly-formed templates. The algorithm continues to merge templates until no further acceptable merges meeting the criteria can be formed. An important advantage of our method is the ability to identify multiple subsets of configurations to derive templates representing the common points in each. From the perspective of our current application, this is important in pharmacophore modelling, where experts would expect more than one plausible pharmacophore as ligands may bind active sites in more than one way. Note that we have concentrated here on rigid body transformations (form analysis). This choice is motivated by our application described in this paper, but the methodology is applicable to other transformations (see, for example, Dryden and Mardia, 1998) used in shape analysis.

In the implementation we have described in this paper, we remove items merged to form new templates at each stage, so any given ligand can contribute to at most one template. An alternative approach allowing for "overlapping" templates, would be to not remove items from the list of available objects in step (3) of the HT algorithm. The resulting templates would overlap in the sense that an individual ligand could feature in more than one template. We could also consider the number of ligands contributing to a template when discriminating between pairwise alignments to decide on a merge at each stage. For example, a scoring function of the form

$$(N^w n_0)^{-1} \sum \log p_{jk},$$

where N is the number of configurations contributing to a template, would weight templates with larger values of N more favourably for $w > 1$. This parameter could be adjusted depending on how much weight one wishes to place on more inclusive templates. The case where $w = 0$ reduces to the situation we have in the examples in this paper, where the number of configurations contributing to a template has not been considered as part of the process of evaluating them.

The algorithm is computationally simple, with complexity $\mathcal{O}(C^2)$, and fast to implement. The tracking of all acceptable templates, beyond those chosen as optimal at each stage, prevents the need to reevaluate pairwise alignments, saving considerable computation. It should be noted that the complexity also depends on the choice of pairwise alignment method, in this case that given by Green and Mardia (2006).

The method we have developed here aims to construct templates representing common points in multiple configurations. Each template is essentially an object which captures the average geometrical information of common points in an optimal way. In the context of the applications described here, it should be noted that a template does not represent a chemical entity in itself. The geometrical information contained in a template could however be used to identify the key features common to each configuration and their relative spatial orientation, so that a plausible pharmacophore model could be constructed. Further work could require templates to satisfy chemical constraints, in order to determine a representative molecule directly.

Supplementary Materials

Web Appendices, Tables, and Figures referred to in Sections 3 and 4 are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

Acknowledgments

CJF acknowledges funding from EPSRC for his research studies during which this work was done.

References

Berkelaar, M. (1996). lpsolve - Simplex-based code for linear and integer programming. *http://www.cs.sunysb.edu/ãlgorith/implement/lpsolve/implement.shtml*.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, N.E. (2000). The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242.

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

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

Gold, N.D. and Jackson, R.M. (2006). Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. *Journal of Molecular Biology* **355**, 1112–1124.

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

Hancock, E.R. and Cross, A.D.J. (1998). Graph matching with a dual-step EM algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, 1236–1253.

Kent, J.T., Mardia, K.V. and Taylor, C.C. (2004). Matching problems for unlabelled configurations. In *LASR2004 Proceedings: Bioinformatics, Images, and Wavelets*, R. Aykroyd, S. Barber and K.V. Mardia (eds), 33–36, Leeds University Press.

Leach, A.R. and Gillet, V.J. (2003). *An Introduction to Chemoinformatics.* The Netherlands: Kluwer Academic Publishers.

Luo, B. and Hancock, E.R. (2001). Structural graph matching using the EM algorithm and singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23**, 1120-1136.

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

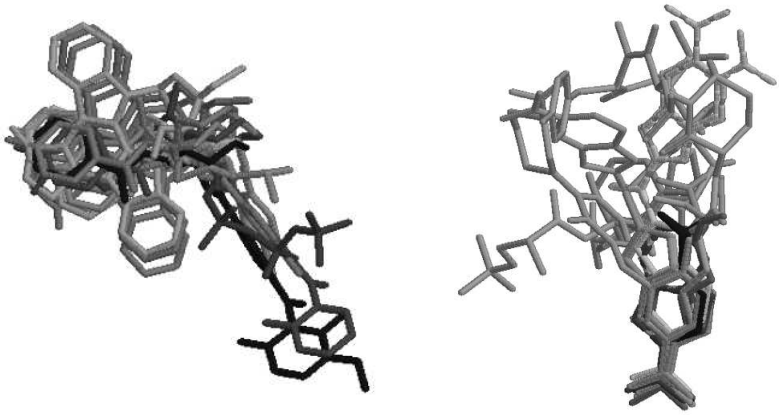
Rella, M., Rushworth, C.A., Guy, J.L., Turner, A.J., Langer, T. and Jackson, R.M. (2006). Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors. *Journal of Chemical Information and Modeling* **46**, 708–716.

Ruffieux, Y. and Green, P. J. (2008). Alignment of multiple configurations using hierarchical models. *Journal of Computational and Graphical Statistics* (To appear).

Schmidler, S.C. (2007). Fast Bayesian shape matching using geometric algorithms. In *Bayesian Statistics*, Bernardo et al. (eds),1–20, Oxford University Press.

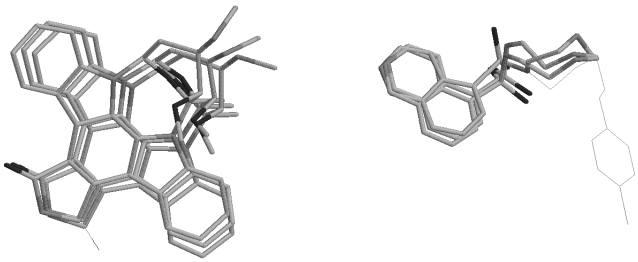
Received May 2008. *Revised MMM* 2009.

Accepted MMM 2009.

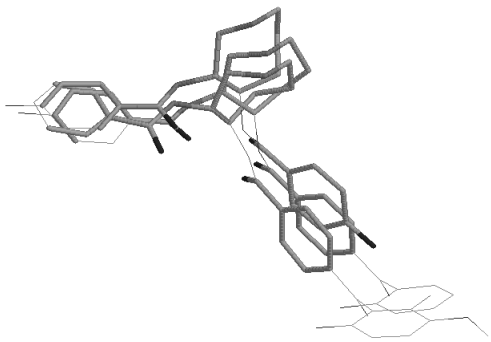


a) 1ATP and b) 3PTB ligands

Figure 1. Sets of 1ATP and 3PTB ligands.

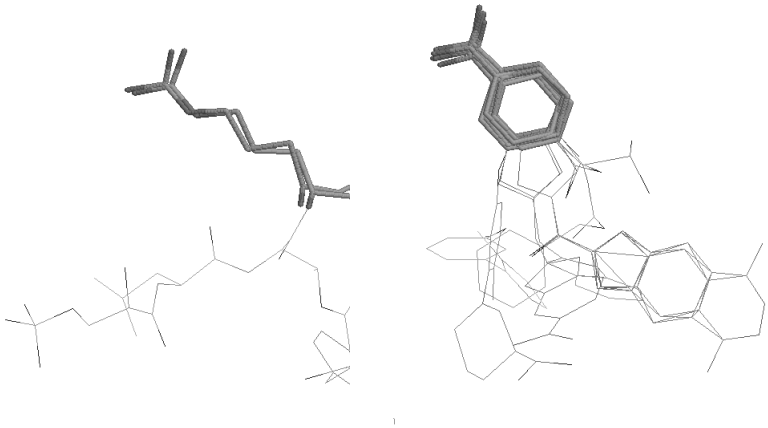


Template A and Template B



Template C

Figure 2. 1ATP ligands contributing to the templates found by the HT algorithm. The template geometry is defined by the mean position of common matching atoms.



Template A and Template B

Figure 3. 3PTB ligands contributing to the templates found by the HT algorithm. The template geometry is defined by the mean position of common matching atoms.

Table 1
Sizes of 1ATP Ligands

Ligand No.	1	2	3	4	5	6	7	8	9	10	11
No. of Atoms	31	35	35	18	27	36	20	18	27	35	37

Table 2
Sizes of 3PTB Ligands

Ligand No.	1	2	3	4	5	6	7	8	9	10	11
No. of Atoms	9	20	22	40	41	25	24	72	9	11	14

Table 3
Probable templates for 1ATP

Template #	Configurations	# of common atoms
A	2 3 6	35
B	4 5 7	18
C	9 10 11	24

Table 4
Probable templates for 3PTB

Template #	Configurations	# of common atoms
A	19 21	11
B	12 14 15 16 17 18 20 22	9