

# Bayesian alignment of continuous molecular shapes using random fields

Irina Czogiel\*, Ian L. Dryden, & Christopher J. Brignell

School of Mathematical Sciences, University of Nottingham

## 1 Introduction

In drug design one attempts to correlate the three-dimensional structure of drug molecules with their biochemical activity. A frequent objective is to find molecules with a high binding affinity towards a certain target protein of therapeutic importance. If no structural information about the target protein is available, putative ligands are often superimposed with the structure of a reference ligand which is known to bind to the target under consideration. If a ligand can be closely aligned to the reference structure, it is likely to exhibit a similar biochemical activity and hence drug potency.

We propose a statistical model for evaluating and comparing molecular shapes which can be used for the alignment procedure. Methods from the field of statistical shape analysis serve as basis for this model. In order to account for the rather continuous nature of molecules, we combine these methods with techniques used for predicting random fields in spatial statistics. Within a Bayesian framework using Markov chain Monte Carlo (MCMC), the resulting molecular fields are aligned with respect to rotation and translation. This procedure can be viewed as a continuous extension of the partial Procrustes analysis. Using a similar concept, we also propose an adaption of the generalised Procrustes analysis algorithm for the simultaneous alignment of molecular fields.

Our methods work well on a data set comprising 31 steroid molecules which has been used as a test bed for various alignment techniques. Two example steroids are shown in Figure 1.

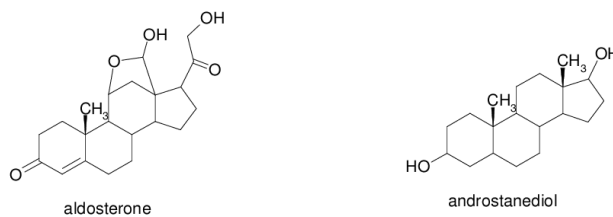


Figure 1: Two-dimensional representations of two steroid molecules from the data set. The molecules are structurally similar in that their core structure consists of four carbon rings.

## 2 Evaluating Molecular Shapes

In most data sets for molecular alignment, a molecule  $M$  is associated with its conformation matrix  $\mathbf{X} = (\mathbf{x}_1 \dots \mathbf{x}_k)^T \in \mathbb{R}^{k \times 3}$  and a matrix of marks  $\mathbf{Z} \in \mathbb{R}^{k \times p}$ , where  $k$  denotes the number of atoms in  $M$ ,  $\mathbf{x}_i \in \mathbb{R}^3$  is the  $xyz$ -coordinate vector of the position of the  $i$ th atom, and  $\mathbf{Z}$  row-wise contains  $p$ -dimensional vectors of molecular properties (e.g. partial charge, van der Waals radius, hydrophobicity, ...) observed at the atom positions.

In order to obtain a descriptor of molecular shape which captures the continuous nature of a molecule, we interpolate the values in  $\mathbf{Z}$  into  $\mathbb{R}^3$  using kriging (e.g. Cressie, 1993, Chapter 3). Let us restrict the

consideration to a single molecular property, i.e.  $\mathbf{Z} = \mathbf{z} = (z(\mathbf{x}_1), \dots, z(\mathbf{x}_k))^T$ . In the geostatistical context,  $z$  is viewed as a sample of one realisation  $z(\mathbf{x})$  of the random field  $\{Z(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^3\}$  which in the following is assumed to be second-order stationary and isotropic with a constant mean of zero and a positive definite covariance function  $\sigma(\|\mathbf{h}\|) = \text{Cov}(Z(\mathbf{x}), Z(\mathbf{x} + \mathbf{h})) \forall \mathbf{x}, \mathbf{h} \in \mathbb{R}^3$ . With the above assumptions, simple kriging is to be applied, where the predicted value of the random field at a location of interest  $\mathbf{x}_0$  is calculated as the weighted average of the  $z(\mathbf{x}_i)$  which minimises the prediction mean squared error. Generalising this procedure to the entire  $\mathbb{R}^3$  then yields the predicted field

$$\hat{Z}(\mathbf{x}) = \mathbf{z}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\sigma}(\mathbf{x}) = \sum_{i=1}^k w_i \sigma(\mathbf{x}_i - \mathbf{x}) \quad \forall \mathbf{x} \in \mathbb{R}^3, \quad (1)$$

where  $\boldsymbol{\sigma}(\mathbf{x}) = (\sigma(\mathbf{x}_1 - \mathbf{x}), \dots, \sigma(\mathbf{x}_n - \mathbf{x}))^T$  and  $(\boldsymbol{\Sigma})_{ij} = \sigma(\mathbf{x}_i - \mathbf{x}_j)$ . The vector of weights  $\mathbf{w} = \boldsymbol{\Sigma}^{-1} \mathbf{z}$  thereby combines the information about the geometry of the molecule and the observed values of the quantity  $Z$  and has a well-defined optimality property.

Figure 2 displays  $xy$ -cross-sections of the three-dimensional steric field of aldosterone which results when the Gaussian covariance function and the van der Waals radii of the atoms are used in (1). The ring structure shown in Figure 1 is clearly visible.



Figure 2:  $xy$ -cross-sections of the steric field for aldosterone at different  $z$ -values.

### 3 Comparing Molecular Shapes

A similarity index for field-based representations of molecular shapes which is well-established in the literature on molecular alignment is the Carbo index (Carbo *et al.*, 1980). In terms of the Carbo index, the similarity between the kriged fields of two molecules  $A$  and  $B$  in a certain relative position is given by

$$C_{AB}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}) = \frac{\int \hat{Z}_A(\mathbf{x}) \hat{Z}_B(\mathbf{x}) d\mathbf{x}}{(\int \hat{Z}_A^2(\mathbf{x}) d\mathbf{x})^{1/2} (\int \hat{Z}_B^2(\mathbf{x}) d\mathbf{x})^{1/2}} \in [-1, 1], \quad (2)$$

where  $\boldsymbol{\Gamma} \in SO(3)$  and  $\boldsymbol{\gamma} \in \mathbb{R}^3$  denote the rotation matrix and the translation vector which define the relative position of  $A$  and  $B$ , respectively. The numerator term in (2) measures the ‘‘overlap’’ of the molecular fields whereas the denominator acts as a normalising constant.

In some cases, it may be of interest to compare only parts of two molecular structures, e.g. the parts which bind to a common receptor protein. This can be achieved by introducing two mask vectors  $\boldsymbol{\lambda}_A \in \Lambda_{k_A}$  and  $\boldsymbol{\lambda}_B \in \Lambda_{k_B}$ . The space  $\Lambda_{k_M}$  ( $M = A, B$ ) thereby contains all  $k_M$ -vectors with entries  $\in \{0, 1\}$ , and the entry  $\lambda_i^M$  determines whether or not the  $i$ th atom of molecule  $M$  and the corresponding terms  $w_i$  and  $\sigma(\mathbf{x}_i^M - \mathbf{x})$  are included when calculating the individual fields using (1).

In case a distance rather than a similarity between two molecules is required, the Carbo index of the masked fields can be mapped into the appropriate codomain using the transformation

$$D_{AB}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}, \boldsymbol{\lambda}_A, \boldsymbol{\lambda}_B) = \frac{1 - C_{AB}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}, \boldsymbol{\lambda}_A, \boldsymbol{\lambda}_B)}{1 + C_{AB}(\boldsymbol{\Gamma}, \boldsymbol{\gamma}, \boldsymbol{\lambda}_A, \boldsymbol{\lambda}_B)} \in [0, \infty).$$

The drawback of this ‘‘partial Carbo distance’’ is that it depends on the relative position of the molecules and the applied mask vectors. To overcome this problem, we define a Bayesian framework within which the rigid-body parameters and the masks can be estimated using posterior analysis.

## 4 Bayesian Alignment of Two Molecular Fields

Like Dryden *et al.* (2007), we use an MCMC scheme for the pairwise alignment of two molecules in which one molecule is viewed as random while the other one serves as a fixed reference molecule: Let  $A$  be the random molecule with an estimated field  $\hat{Z}_A(\mathbf{x})$  and  $B$  the fixed molecule represented by  $\hat{Z}_B(\mathbf{x})$ . Under the premise that  $A$  and  $B$  bind to the same protein (and that parts of their structures are therefore similar), we define the likelihood for the random molecule as

$$L(\hat{Z}_A(\mathbf{x})|\mathbf{\Gamma}, \gamma, \boldsymbol{\lambda}_A, \boldsymbol{\lambda}_B, \tau, \xi, \hat{Z}_B(\mathbf{x})) \propto \tau^{\xi-1} \exp(-\tau D_{AB}(\mathbf{\Gamma}, \gamma, \boldsymbol{\lambda}_A, \boldsymbol{\lambda}_B)),$$

where  $\tau \in \mathbb{R}^+$  is a precision parameter which determines mean and variance of the likelihood. This parameter as well as the rotation and translation parameters and the mask vectors are considered as random whereas  $\xi$  is a fixed additional parameter which can be estimated from the data.

Following Green and Mardia (2006), we want to perform simultaneous posterior inference about the precision parameter and the masks by integrating out the rigid body parameters which depend on the usually arbitrarily recorded coordinate frames of the molecules. To facilitate this, we need to specify prior distributions for the random likelihood parameters. As we do not have any prior information about the rigid body parameters, they are treated as uniformly distributed in their respective domains (which involves the Haar measure for the rotation parameters). The joint prior distribution for the mask vectors is defined as

$$\pi(\boldsymbol{\lambda}_A, \boldsymbol{\lambda}_B|\zeta) \propto \zeta^{\sum_i \lambda_i^A + \sum_i \lambda_i^B},$$

where the penalty parameter  $\zeta \in \mathbb{R}^+$  is introduced to prevent the MCMC algorithm from converging to a solution where very few atoms are used in the distance calculation. With the further assumptions that the precision parameter is Gamma distributed *a priori* with shape parameter  $\alpha$  and scale parameter  $\beta$ , and that all unknown parameters are independent *a priori*, their joint posterior density conditioned on the given data and the covariance function which is applied when calculating the molecular fields has the property

$$\begin{aligned} \pi(\boldsymbol{\theta}, \gamma, \boldsymbol{\lambda}_A, \boldsymbol{\lambda}_B, \tau|\hat{Z}_A(\mathbf{x}), \hat{Z}_B(\mathbf{x}), \alpha, \beta, \xi, \zeta) \\ \propto \tau^{\xi+\alpha-2} \exp\{-\tau (D_{AB}(\mathbf{\Gamma}, \gamma, \boldsymbol{\lambda}_A, \boldsymbol{\lambda}_B) + \beta)\} \cdot \zeta^{\sum_i \lambda_i^A + \sum_i \lambda_i^B} \cos(\theta_2). \end{aligned}$$

Bayesian inference can now be carried out in order to obtain a rotation/translation invariant notion of the (dis)similarity between the molecular fields  $\hat{Z}_A(\mathbf{x})$  and  $\hat{Z}_B(\mathbf{x})$ . In particular, we use MCMC to sample from the posterior distribution and obtain point estimates for the random model parameters. Within the MCMC scheme,  $\tau$  is updated with a Gibbs step using its full conditional distribution. Updated versions of the other parameters are obtained in four blocks, each using a Metropolis–Hastings step with a random walk proposal density.

## 5 Results

The 31 steroid molecules fall into three classes with respect to their binding activity towards their common receptor protein (1=high, 2=medium, 3=low). We perform all 930 possible pairwise superposition and use the geometric mean of the MAP partial Carbo distances for each pair of molecules as a symmetric distance measure for a cluster analysis. Figure 3 shows the resulting dendrogram of the steroids which is labelled with their activity classes. As steroids within an activity class can overall be aligned more closely than steroids between activity classes, the Bayesian alignment of molecular fields can be shown to produce chemically meaningful results.

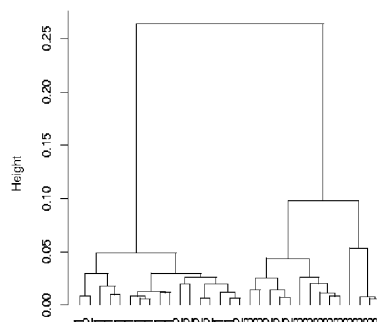


Figure 3: Separation of the 31 steroid molecules using a symmetrical partial Carbo distance.

## 6 Extension to Multiple Comparisons

The above methodology can be viewed as an extension of the partial Procrustes registration for discrete sets of landmarks (e.g. Dryden & Mardia, p.94) to the continuous case. Similarly, superimposing each molecular field in turn onto a weighted average of the remaining fields in the data set provides an extension of the generalised Procrustes analysis (GPA) (Gower, 1975) to the field context. After applying this continuous GPA to the multiple alignment of the steroid molecules, significant differences between the mean steric fields of the three activity classes can be detected which could explain the different biochemical behaviour.

## References

- Carbo, R., Leyda, L. and Arnau, M. (1980). An electron density measure of the similarity between two compounds. *International Journal of Quantum Chemistry*, **17**, 1185–1189.
- Cressie, N.A.C. (1993). *Statistics for Spatial Data*. Chichester, Wiley.
- Dryden, I.L. and Mardia, K.V. (1998). *Statistical Shape Analysis*. Chichester, Wiley.
- Dryden, I.L., Hirst, J.D. and Melville, J.M. (2007). Statistical analysis of unlabelled point sets: comparing molecules in chemoinformatics. *Biometrics*, **63**, 237–251
- Gower, J.C. (1975). Generalized Procrustes analysis. *Psychometrika*, **40**, 33–50.
- Green, P.J. and Mardia, K.V. (2006). Bayesian alignment using hierarchical models, with application to bioinformatics. *Biometrika*, **93**, 235–254.