Hierarchical Bayesian modelling of pharmacophores

Kanti V. Mardia[1], Vysaul B. Nyirongo[1], Christopher J. Fallaize[1*], Stuart Barber[1] and Richard M. Jackson[2]

[1] Department of Statistics, University of Leeds
[2] Institute of Molecular and Cellular Biology, University of Leeds

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. 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 and hence there is a need to develop statistical methodology for deriving pharmacophore models and quantifying their plausibility.

Since protein-ligand complexes are commonly represented as configurations of points in $\mathbb{R}^3$, with each point representing the location of an individual atom, pharmacophore identification can be reduced directly to the problem of finding points common to a set of configurations. Methods for the alignment of multiple configurations have been proposed by, for example, Dryden *et al*. (2007), and Ruffieux and Green (2008). Here we develop a hierarchical model for the derivation of pharmacophore templates from multiple configurations of point sets. Within our model, we require the use of a method for the pairwise alignment of two configurations that can estimate which points match and the corresponding probabilities. Here we use the pairwise alignment method described by Green and Mardia (2006)- an alternative choice could be to use the EM algorithm, as proposed by Kent *et al*. (2004). 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. Chemical information is used by labelling points by element type, whereby points representing points of different elements are less likely to be matched than points representing the same element type. Our method is illustrated through application to two example datasets of ligands binding structurally-related protein active sites.

References

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.

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

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

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