EM algorithm, Bayesian and distance approaches to matching functional sites

Vysaul Nyirongo*[1], Kanti V. Mardia[1] and David R. Westhead[2]

[1]Department of Statistics, University of Leeds
[2]School of Biochemistry and Microbiology, University of Leeds.

The explosion in volume of protein structural information prior to any knowledge of protein biochemical function has made the characterisation of protein functional sites to be an area of huge interest. Structural similarity of functional sites from proteins with unknown function to those with known functions can be used to infer on the function of the former.

Structural similarity has been formulated as a graph theoretic problem (Gold, 2003), but can as well be cast in missing data framework in statistics. In this poster we compare some of the solutions to the problem from graph theoretic and missing data perspectives. In particular we consider the EM algorithm for the mixture model formulation of the problem (Kent *et al.*, 2003) , MCMC algorithm and the graph theoretic approach. The MCMC algorithm is used for a Bayesian approach in a hierarchical model formulation (Green and Mardia, 2004).

We discuss advantages and disadvantages of these approaches and an application to the representatives of tyrosine dependent oxidoreductase family is given. Illustrated in Figure 3 is a superposition of the matched atoms of functional sites from 17-beta-hydroxysteroid dehydrogenase (1a27) and carbonyl reductase (1cyd) proteins.

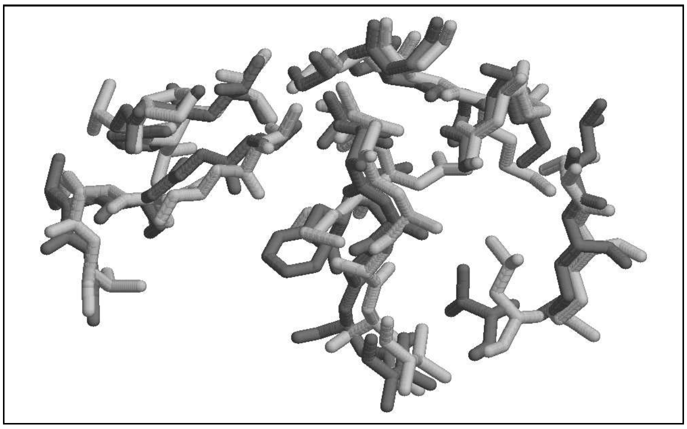

Figure 1: Full MCMC Matching of active sites of two proteins: 17-beta-hydroxysteroid dehydrogenase and carbonyl reductase.

The figure shows that the matching is good even at the level of amino acid though only C^α atoms were used in finding the correspondence and alignment.

References

Gold, N.D., Pickering, S.J. and Westhead, D.R. (2003). Predicting protein function from structure using SITEDB: evaluation of a method based on functional site-similarity. Preprint.

Green, P.J. and Mardia, K.V. (2004). Bayesian alignment using hierarchical models with applications in protein bioinformatics. Preprint, available from `http://www.stats.bris.ac.uk/∼peter/papers/align.pdf`.

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

Taylor, C.C., Mardia, K.V. and Kent, J.T. (2003). Matching Unlabelled Configurations Using the EM Algorithm. *Proceedings in Stochastic Geometry, Biological Structure and Images*, 19-21. Edited by R.G. Aykroyd, K.V. Mardia and M.J. Langdon. Leeds University Press.