A New Method for Evaluating the Structural Similarity of Proteins Using Geometric Morphometrics

Dean C. Adams [1] **Gavin J. P. Naylor** [1]
dcadams@iastate.edu gnaylor@iastate.edu

[1] Department of Zoology and Genetics, Iowa State University, Ames, Iowa 50011-3223, USA

1 Introduction

Recently, Holm and Sander [2, 3] described a method to quantify structural similarity among proteins. They used this method to generate a 'protein structure space' where proteins with similar structure cluster together. Proteins are compared with the similarity measure:

$$S = \sum_i \sum_j \left(0.2 - \frac{|d_{ij}^A - d_{ij}^B|}{d_{ij}^*}\right) e^{-\left(\frac{d_{ij}^*}{20A}\right)^2}$$

where d_{ij} is the distance between amino acid residues, and d_{ij}^* is the mean distance for those residues (a standardized Z-score of S is typically used [2, 3]). A matrix of Z-scores is then built up in pairwise fashion and used to generate a protein structure space. While this protocol is useful, we feel it can be improved upon. Three aspects of their protocol are particularly problematic: 1) different pairwise contrasts use different sets of amino acids, so the resulting Z-scores are not comparable; 2) a structural similarity threshold of 20% is arbitrarily imposed and unfortunately results in mathematical inconsistencies of their space; 3) a weight function in their equation, the exponential term, confounds shape similarity with size similarity. Tools currently used for the analysis of anatomical shape do not have these difficulties. These geometric morphometric methods [1, 4], utilize the three-dimensional coordinates of homologous landmarks (residues) as the starting point of shape comparisons. Non-shape variation is removed through a least-squares superimposition, and a mathematically consistent multi-dimensional space results in which proteins with similar shape cluster together. Here we explore these methods for protein structure comparisons using globins as an example data set.

2 Methods and Results

We extracted 560 globin sequences from the PDB and aligned them with Clustal W [5]. All gaps were deleted, yielding a data set of 96 homologous residues per sequence. The corresponding x,y,z coordinates extracted from the PDB were then used in geometric morphometric (GM) shape analyses. We first eliminated non-shape variation (position, orientation, and size) from the data using the generalized least squares superimposition [4] equation:

$$\mathbf{X_i} = \rho \mathbf{X} \mathbf{H} + \mathbf{1}\tau$$

where ρ describes scaling, $\mathbf{H}$ is the rotation matrix $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$, $\mathbf{1}\tau$ is translation to a common location, and $\mathbf{X}$ is the original configuration of 3-dimensional landmarks. A protein shape space was then generated through a PCA of the superimposed protein structures, and similar proteins were identified through their clustering patterns.

The first 3 dimensions of GM space explain 76% of the variation among globins (Fig. 1). By contrast, only 34% of the variation is explained using Holm and Sander's approach. Furthermore, 56 dimensions are required to explain 76% of the variation. Thus the GM method more succinctly describes the shape variation among proteins. The GM method also identifies distinct clusters of proteins known to have similar shapes (e.g., cyanomet haemoglobins, leghaemoglobins, apo-ovotransferrins).

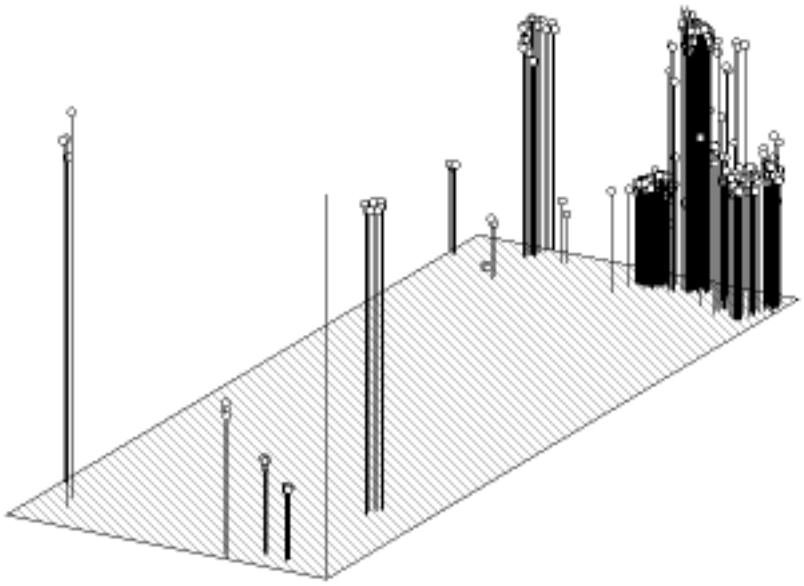


Figure 1: Three dimensional representation of protein shape space using GM methods.

3 Conclusions

Identifying structurally similar proteins is useful because structural similarity is believed to reflect functional similarity. In order to identify similar proteins it is important to use methods that are mathematically consistent and efficient in their representation. We have presented a simple method commonly used in anatomy that compares shapes using the coordinates of biologically homologous landmarks. When applied to globin proteins these methods describe a significant percentage of the variation and yield a clustering pattern consistent with biological expectations. Other methods currently used [2, 3] do not perform as well.

References

[1] Bookstein, F. L. *Geometric Morphometrics: Geometry and Biology*, Cambridge Univ. Press, 1991.

[2] Holm, L. and Sander, C. Mapping the protein universe, *Science*, 273:595–602, 1996.

[3] Holm, L. and Sander, C. Dictionary of recurrent domains in protein structures, *Proteins: Struct. Funct. Gen.*, 33:88–96, 1998.

[4] Rohlf, F. J. and Slice, D. E. Extensions of the Procrustes method for the optimal superimposition of landmarks, *Syst. Zool.*, 39:40–59, 1990.

[5] Thompson, J. D., Higgins, D. G., and Gibson, D. G. Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice, *Nuc. Acids Res.*, 22:4673–4680, 1994.