2003. In Mathematical Methods for Protein Structure Analysis and Design. Advanced Lectures. (C. Guerra, S. Istrail., eds.). Springer Verlag Lecture Notes in Bioinformatics. 2666:109-115.

A Comparison of Methods for Assessing the Structural Similarity of Proteins

Dean C. Adams and Gavin J. P. Naylor *

Dept. Zoology and Genetics, Iowa State University, Ames, IA 50011, U.S.A.

1 Introduction

The link between biological form and function is well known, and is assumed to hold true at the molecular level. Since identifying similar protein structures is the first step in identifying similar functions, much effort has been placed in developing methods to detect structural similarity. Several methods exist, including: SCOP [8], the DALI algorithm (from the FSSP Database [6]), the VAST algorithm (from the MMDB database [5]), and Root Mean Square (RMS) superimposition [9]. The latter three provide quantitative metrics describing protein similarity on an objective, continuous scale. Statistical analyses can then be performed on similarity scores for a set of proteins, to obtain a plot of 'protein structure space' [7]. Before such analyses are done however, one must be sure that the metric used accurately represents similarity.

In this paper, we describe the DALI Z-score and RMS-distance (D_{RMS}) metrics, and discuss their shortcomings. We then present a novel means of comparing protein structures using Geometric Morphometric (GM) methods: statistical shape methods borrowed from anatomy. Finally, we compare results from these three methods for a data set of globin structures, and show that the more intuitive GM method markedly outperforms existing techniques.

2 The DALI Algorithm

The DALI algorithm [6],[7] compares protein structures using two-dimensional matrices, where each element in the matrix (d_{ij}) is the Euclidean distance between the i^{th} and j^{th} residues for that protein. Distance matrices are aligned in pairwise fashion, and n homologous residues are identified. The structural similarity for the two proteins (A and B) is then calculated as:

$$S = \sum_i \sum_j \left(0.2 - \frac{|d_{ij}^A - d_{ij}^B|}{d_{ij}^*}\right) e^{-\left(\frac{d_{ij}^*}{20A}\right)^2} \qquad (2.1)$$

where d_{ij}^* is the mean distance for those residues (a standardized version of S, the Z-score, is also calculated). Z-scores are calculated for all protein pairs, and

* Work supported in part by a National Science Foundation Postdoctoral Fellowship in Biological Informatics: DBI-9974207 to DCA.

the best three-dimensional ordination of the structure space is found through an eigen-decomposition (correspondence analysis) of the Z-score matrix, where similar proteins are close together, and dissimilar proteins are far apart.

Though Z-scores quantify some aspects of structural similarity, details of this metric warrant careful scrutiny. First, Z-scores are generated from pairwise alignments, so different residues can be used for each pair. Thus, values in the Z-score matrix represent different aspects of structural similarity, and are not directly comparable. Second, the metric contains a dissimilarity cut-off (0.2) to eliminate protein comparisons $> 20\%$. However, most protein comparisons in a large database are $> 20\%$, yielding negative scores, which are arbitrarily truncated to zero. An eigen-analysis of such data will explain little of the variation in few dimensions, and a low dimensional ordination from this analysis will fail to capture the essence of 'structure space.' Finally, the exponential term in the metric downweights contributions from residues far from one another. This results in Z-scores for self-comparisons that are not the same for each protein, implying that some proteins are more 'perfectly' similar to themselves than others, which is nonsensical. Thus, DALI Z-scores are not a true similarity metric, and statistical analyses of them are unpredictable.

3 The Root Mean Square Algorithm

Root Mean Square (RMS) methods assess structural similarity using a least squares (LS) criterion. First, two proteins (X & Y) are structurally aligned to identify the set of n homologous residues [4]. Next, they are translated to a common location, and are rotated so that homologous residues line up as closely as possible [9]. Finally, the Euclidean distance (D_{RMS}) between them is calculated:

$$D_{RMS} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{3}(X_{ij} - Y_{ij})^2} \tag{3.2}$$

where X_{ij} and Y_{ij} are the coordinate sets for the i^{th} residue. D_{RMS} is calculated for all protein pairs, and the best three-dimensional ordination of the structure space is found through an eigen-decomposition (principal coordinates analysis) of the D_{RMS} matrix.

RMS methods are appealing because D_{RMS} makes intuitive sense: unlike proteins have a large D_{RMS}, while similar proteins align quite well and have a small D_{RMS}. It is also a true distance measure, because all self-comparisons of proteins yield an identical value of zero (no structural differences). Like Z-scores however, D_{RMS} is calculated in pairwise fashion, so different residues can be used for each protein pair, rendering D_{RMS} scores incomparable.

4 Geometric Morphometrics

Both DALI Z-scores and D_{RMS} can be used to generate a map of protein structure space. However, both have methodological difficulties which limit their

utility. Interestingly, these same difficulties have already been addressed in a completely different discipline: Geometric Morphometrics (GM). GM methods were originally developed to analyze anatomical structures (e.g., skulls), but may easily be adapted to compare macromolecular structures. First, a set of homologous points recorded on all specimens are superimposed using generalized Procrustes analysis (GPA), which translates specimens to a common location, scales them to unit size, and optimally rotates them (in a LS sense) [3],[10]. Shape variables are then generated for each specimen, which may be used in statistical analyses [2]. Additionally, Procrustes distance (D_{PROC}) between two specimens (X & Y) can be calculated as:

$$D_{PROC} = 2\sin^{-1}\left(\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{3}(X_{ij}-Y_{ij})^2/2}\right) \tag{4.3}$$

where X_{ij} and Y_{ij} are the aligned coordinates for the i^{th} residue. D_{PROC} is calculated for all protein pairs, and the best three-dimensional plot of protein shape space is found through an eigen-decomposition (principal coordinates analysis) of this data. Although GM and RMS are quite similar, they differ in two important respects. First, size is mathematically held constant in GPA (not in RMS), and second, GPA superimposes all specimens simultaneously.

5 Comparison of Methods

To compare the three methods described above we used a representative set of protein structures. We extracted all globin sequences (as of 12/10/1999) from the Protein Data Bank, and separated them into their individual chains, so that monomeric and non-monomeric globins could be used. Structural similarity among the 560 chains was then assessed using each of the three methods (Z-scores, D_{RMS}, D_{PROC}). Pairwise structural alignments were calculated in the DALI domain dictionary [7] (http://www2.ebi.ac.uk/dali) and both DALI Z-scores and D_{RMS} scores were obtained for each protein pair. For GM, we aligned the amino acid sequences with Clustal W [11] and deleted all gaps, yielding 96 homologous residues [1]. We then superimposed the structural data for these residues with GPA, and generated D_{PROC} for each protein pair.

The ability of each metric to capture structural variation was assessed using multivariate ordination methods. The DALI Z-score matrix was summarized using correspondence analysis (as per [7]), and the D_{RMS} and D_{PROC} matrices were summarized using principal coordinates analysis. The percentage of variation explained by the first three dimensions from the ordination analysis was compared for each method, and their ability to identify biologically meaningful clusters was assessed through a visual inspection of the ordination plots.

Using D_{RMS}, the 1^{st} three dimensions of structure space explained 76.1% of globin chain variability. Inspection of this ordination plot revealed separation of a few individual chains (mostly hemoglobin chains), but no obvious

groups were identifiable (Fig. 1). Thus, although D_{RMS} explained much of the variation, it was unable to identify any biologically meaningful globin clusters.

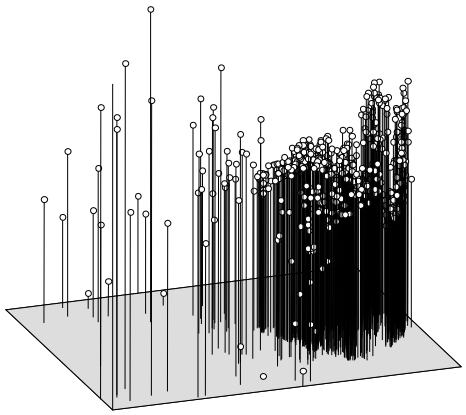


Fig. 1. Three-dimensional view of globin structure space from D_{RMS}.

Using DALI Z-scores, the 1^{st} three dimensions of structure space explained 33.5% of globin chain variation, and to describe an equivalent amount of variation to D_{RMS} (76%), 56 dimensions of the ordination were needed. Further, the ordination plot revealed no obvious clusters of globin chains (Fig. 2). Thus, DALI Z-scores were much less effective at summarizing structural variability, and were unable to reveal biologically interpretable clusters of proteins.

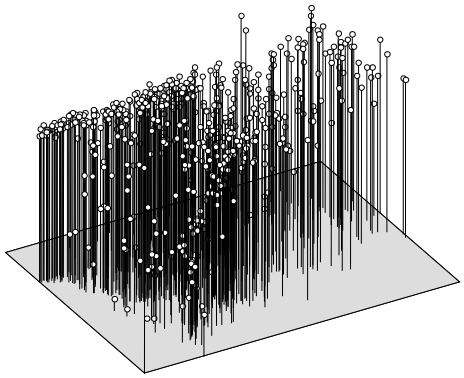


Fig. 2. Three-dimensional view of globin structure space from DALI Z-scores.

Using D_{PROC}, the 1^{st} three dimensions of the GM shape space explained 76.6% of the variation, which was similar to that found with D_{RMS}. However, the ordination plot revealed remarkable separation of globin chains into identifiable groups. These groups corresponded to meaningful biological partitions of the data set, including: bacterial hemoglobins, clam hemoglobins, ferric hemoglobins, hemoglobin α (& β)-chains, lamprey hemoglobins, leghemoglobins, and myoglobins (Fig. 3). Thus, much more biological information is obtained using GM, as compared to either DALI Z-scores or D_{RMS}.

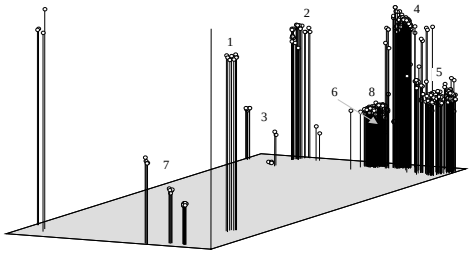


Fig. 3. Three-dimensional view of globin shape space from D_{PROC}. Labels correspond to the following groups: 1: bacterial hemoglobins, 2: clam hemoglobins, 3: ferric globins, 4: hemoglobin α-chains, 5: hemoglobin β-chains, 6: lamprey hemoglobins, 7: leghemoglobins, and 8: myoglobins.

6 Discussion

In this paper we described two metrics used for assessing structural similarity (DALI Z-scores and D_{RMS}), and described how geometric morphometric (GM) methods, commonly used in anatomical studies, may also be employed to compare protein structures. We then compared the ability of three metrics to summarize structural variation in a set of globin structures. DALI Z-scores explained very little (33.5%) of the total variation in three dimensions, and were unable to identify any globin clusters. D_{RMS} explained significantly more variation (76.1%) in three dimensions, but it too was unable to identify clusters of globins. On the other hand, D_{PROC} explained a large proportion of the variation (76.6%) in three dimensions, *and* was able to identify biological clusters of globins (e.g., bacterial hemoglobins, leghemoglobins, myoglobins, etc.). Further, all but 1 globin chain was correctly classified to its biological group (1 hemoglobin α-chain was classified as hemoglobin β-chain).

These results suggest that GM methods may be more useful for extracting meaningful biological information from protein structures than are either the DALI or RMS methods. Why might this be the case? It seems that DALI Z-scores are predisposed *not* to identify meaningful structural variation: they

are calculated in pairwise fashion, and are not a true similarity measure. Further, the arbitrary similarity cut-off predisposes them to identify many protein comparisons as 'maximally' different. Explaining the performance of D_{RMS} however, is more difficult. D_{RMS} is a true distance measure, so it does not suffer the same problem as Z-scores. Further, D_{RMS} and D_{PROC} differ algorithmically in only two respects: D_{RMS} is a linear distance (where D_{PROC} is curve-linear), and the D_{PROC} protocol standardizes the size of each specimen (but for globins, size accounts for a tiny portion of variation). The only other difference between the RMS and GM methods is that D_{RMS} is calculated from homologous residues from *pairwise* structural alignments, whereas D_{PROC} is calculated from homologous residues common to *all* structures (found from a multiple alignment). This assures that the information contained in D_{PROC} is consistent among comparisons, which is an explicit requirement of any statistical analysis. It appears then, that the GM protocol, using homology defined for *all* proteins simulataneously, provides the best chance for identifying natural clusters of structurally-similar proteins.

References

1. D. C. Adams and G. J. P. Naylor: A new method for evaluating the structural similarity of proteins using geometric morphometrics. in S. Miyano, R. Shamir, and T. Takagi (eds.) *(*Currents in computational molecular biology). Universal Academy Press, Tokyo 2000.
2. D. C. Adams and F. J. Rohlf: Ecological character displacement in *Plethodon*: biomechanical differences found from a geometric morphometric study. Proc. Natl. Acad. Sci. U.S.A. **97** (2000) 4106–4111.
3. F. L. Bookstein: *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press, Cambridge 1991.
4. M. Gerstain and M. Levitt: Comprehensive assessment of automatic structural alignment against a manual standard, the scop classification of proteins. Protein Sci. **7** (1998) 445–456.
5. J.-F. Gibrat, T. Madej, and S. H. Bryant: Surprising similarities in structure comparison. Curr. Opin. Struct. Biol. **6** (1996) 377–385.
6. L. Holm and C. Sander: Protein structure comparison by alignment of distance matrices. J. Mol. Biol. **233** (1993) 123–138.
7. L. Holm and C. Sander: Mapping the protein universe. Science **273** (1996) 595–602.
8. A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia: SCOP: A structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247** (1995) 536–540.
9. S. T. Rao and M. G. Rossman: Comparison of super-secondary structures in proteins. J. Mol. Biol. **76** (1973) 241–246.
10. F. J. Rohlf and D. E. Slice: Extensions of the Procrustes method for the optimal superimposition of landmarks. Syst. Zool. **39** (1990) 40–59.
11. J. D. Thompson, D. G. Higgins, and T. J. Gibson: Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucl. Acids Res. **22** (1994) 4673–4680.