

# STRALCP—structure alignment-based clustering of proteins

Adam Zemla<sup>1,\*</sup>, Brian Geisbrecht<sup>2</sup>, Jason Smith<sup>1</sup>, Marisa Lam<sup>1</sup>, Bonnie Kirkpatrick<sup>1</sup>, Mark Wagner<sup>1</sup>, Tom Slezak<sup>1</sup> and Carol Ecale Zhou<sup>1</sup>

<sup>1</sup>Computing Applications and Research, Lawrence Livermore National Laboratory, Livermore, CA 94550 and

<sup>2</sup>Division of Cell Biology and Biophysics, University of Missouri-Kansas City, Kansas City, MO 64110, USA

Received June 11, 2007; Revised October 14, 2007; Accepted November 6, 2007

## ABSTRACT

**Protein structural annotation and classification is an important and challenging problem in bioinformatics. Research towards analysis of sequence–structure correspondences is critical for better understanding of a protein’s structure, function, and its interaction with other molecules. Clustering of protein domains based on their structural similarities provides valuable information for protein classification schemes. In this article, we attempt to determine whether structure information alone is sufficient to adequately classify protein structures. We present an algorithm that identifies regions of structural similarity within a given set of protein structures, and uses those regions for clustering. In our approach, called STRALCP (STRucture ALignment-based Clustering of Proteins), we generate detailed information about global and local similarities between pairs of protein structures, identify fragments (spans) that are structurally conserved among proteins, and use these spans to group the structures accordingly. We also provide a web server at <http://as2ts.llnl.gov/AS2TS/STRALCP/> for selecting protein structures, calculating structurally conserved regions and performing automated clustering.**

## INTRODUCTION

Most protein annotation and classification approaches depend heavily on the degree of observed amino acid sequence similarity to other related proteins. But even when sequence similarity between two proteins is low, structure similarity can be high. Thus, one of the most important improvements in protein classification would be protein homology/analogy identification at very low levels of sequence similarity (1). As Redfern *et al.* (2) explain ‘despite the advances in sequence comparison methods,

remote homologs in the “Midnight Zone” of sequence similarity (<15% identity) described by Rost, can still only be identified through protein structure comparison’. Redfern *et al.* also point out that ‘structure-based classifications are becoming increasingly important resources for recognizing these distant relatives and providing datasets for more far-reaching analyses of protein family evolution’. In our research and development we follow these observations, and in order to detect the benefits and limitations of using purely structure-based approaches, we currently concentrated on structure similarity analyses only.

The Protein Data Bank (PDB) (3) already contains more than 45 000 experimentally solved protein structures, and grows at a rate of more than 500 PDB entries per month. Among current entries, approximately 40% are multi-domain proteins (2) and, thus, there have been several attempts to classify individual domains of PDB protein structures into defined clusters (e.g. classes, folds, superfamilies, families) based on structure similarity as measured by various criteria. The most commonly used protein classification databases are SCOP (4) and CATH (5). The Structural Classification of Proteins (SCOP) database, a manual classification of PDB structures, is recognized by many as the gold standard of protein classification. In SCOP, proteins are classified to reflect both structural and evolutionary relatedness. Clustering is based mainly on visual inspection of similarities between conformations of secondary structure elements and on sequence similarities. However, SCOP classification lags the insertion of new structures in PDB, and manual classification cannot scale to meet the demands of this rapidly growing dataset.

There are several algorithms already proposed to facilitate automated protein structure classification. For example, clustering can be done by selecting a single metric (e.g. Z-score (6) used in FSSP (7) Dali Fold Classification) or by combining different criteria to score the level of similarity, some examples of which are secondary structure content and orientation combined with calculated sequence similarities, and manual inspection, as

\*To whom correspondence should be addressed. Tel: +1 925 4235571; Fax: +1 925 4236437; Email: adamz@llnl.gov

used in the semiautomatic CATH database. Depending on the algorithm, classification results may differ significantly if different criteria are used to assess the level of similarity between compared structures or if the clustering criteria are focused on different structural features. The same set of proteins could be grouped differently by automatic sequence or structure comparison tools based on minor modifications to cutoffs or classification parameters. The Dali Fold Classification is based on exhaustive, all-against-all 3D structure comparisons of proteins from the PDB, and is constructed by average linkage clustering of the structural similarity score derived from calculated alignments of distance matrices. The tree (dendrogram) is cut at Dali Z-score levels 2, 4, 8, 16, 32 and 64, where the first level ( $Z > 2$ ) can be used as an operational definition of folds. A similar approach is used in CE (8) classification. After performing all-against-all comparisons of protein chains from the PDB, resulting CE Z-score values of 4.5 and above are used to discriminate at the family level, values between 4.0 and 4.5 at the superfamily and/or fold levels, and values between 3.5 and 4.0 are presumed to indicate possible biologically interesting similarities. The authors of the STRuster (9) method explore the calculation of root mean square deviations (RMSD) and use their algorithm to cluster alternative structural models from the PDB (i.e. models that correspond to different structure determination experiments). In addition to the traditional RMSD measure, the STRuster method uses two filters to define the final scoring metric called dissimilarity measure M (9). These two filters are introduced in order to identify both large and small (but significant) backbone conformational changes by reducing the influence in local large distances (only distances below 14.0 Å are considered) and also to restrict the analysis to significant structural differences (the distances above 1.0 Å). An approach for structural comparisons, fundamentally different from those using RMSD, was proposed by Rogen and Fain (10). They introduced the SGM (Scaled Gauss Metric), which is a metric derived from knot theoretical ideas to cluster proteins according to their structural topologies. They applied their method to predicting membership of proteins in CATH and achieved 95% accuracy at all levels of the classification hierarchy.

In order to achieve a high level of agreement with other clustering schemes, some algorithms that use a multi-criterion approach (weighted combination of different scoring schemes), are initially trained on labeled data from an existing structural hierarchy (SCOP or CATH) and use cross-validation (or similar methods) to select the best parameters for their classifiers. For example, ProtClass (11) uses a nearest-neighbor-based classification scheme and several structural features to classify proteins at the fold level of the SCOP hierarchy. Their features include secondary structure elements predicted by the Stride program (12), the sequence length, and the percentage of observed helices. SCOPmap (13) is an approach that achieves roughly 95% accuracy when classifying proteins into the superfamily level of the SCOP hierarchy. This approach combines many existing protein sequence and structure comparison tools, including PSI-BLAST (14),

MAMMOTH (15) and Dali (6). The classification results depend on the accuracy of the individual tools, so the authors use a variety of cutoffs and parameters optimized by training schemes to apply these tools in a specific order. In Ref. (16), the authors introduce a new structural representation of proteins to predict the family membership of proteins in the SCOP hierarchy. They define a graph theoretic representation of protein structures with nodes being residues and edges connecting residues when the distance separating them falls below a specified cutoff. Using these graphs as features, they train their Support Vector Machine (SVM) classifier with proteins from several SCOP families.

The ultimate goal of the work presented here was to define criteria and to develop an algorithm that for a given set of protein structures would automatically identify structurally conserved regions and use them to create clustering results similar to those that would be obtained by manual inspection (e.g. SCOP curators). In our novel approach, called STRALCP (STRucture ALignment-based Clustering of Proteins), for a given set of protein structures, we generate and combine detailed structural information about automatically detected global and local similarities between protein pairs, identify similar regions that are conserved within the set of proteins, report these regions, and use them to cluster the proteins according to their similarities in such identified structural frames. We use the Local-Global Alignment (LGA) algorithm (17) to perform all necessary structure comparison calculations.

## METHODS

Our algorithm starts from structure alignment calculations performed by LGA (with a default value of distance cutoff  $DIST = 5 \text{ \AA}$ ) to determine *de novo* (no sequence information is used) residue-residue correspondences between compared proteins. We use the *LGA\_S* measure as a scoring function to evaluate the overall level of structure similarity and to allow an initial grouping and structural clustering of proteins. In our STRALCP approach, an optimal number of clusters is determined by grouping models according to their overall similarity (*LGA\_S*) combined with the information about local similarities in detected structurally conserved frames (we call them 'spans').

### *LGA\_S* structure similarity scoring function (overall structure similarity)

To perform a particular clustering for a set of protein structures, a suitable scoring function or, in general, a scoring algorithm that takes into account a number of characteristics of the compared proteins must be defined. Depending on the goal of the clustering, this can be done by selecting one measure or by combining different criteria to score the level of similarity. The *LGA\_S* scoring function has two components, *LCS* (longest continuous segments) and *GDT* (global distance test), defined for the detection of regions of local and global structure similarities between analyzed structures. In comparing two protein structures, LGA superimposes a 'model' structure

onto a 'target' structure (where the model is designated 'M' and the target is 'T'). The LCS procedure localizes and superimposes the longest segments of residues that can fit under a selected set of RMSD cutoffs. The GDT algorithm is designed to complement evaluations made with LCS searching for the largest (not necessary continuous) set of 'equivalent' residues that deviate by no more than specified distance cutoffs. Let:

- $m$ —the number of residues in  $M$ ,
- $t$ —the number of residues in  $T$ ,
- $R(r, M, T) = 100/t * L(r, M, T)$  is the percentage of the target's ( $T$ 's) residues that are involved in the maximal (longest) continuous segment that fits within an RMSD of  $r$  Å.  $L(r, M, T)$  is the length of such identified longest continuous segment of  $M:T$  residue pairs,
- $X(M, T)$ —the set of all  $M:T$  superpositions calculated by the LGA algorithm,
- $G(s, d, M, T)$ —the number of  $M:T$  residue pairs for which the distance between Ca (alpha carbon) atoms is not greater than  $d$  Å after the superposition  $s \in X(M, T)$  is applied,
- $D(d, M, T) = 100/t * \max\{G(s, d, M, T) : s \in X(M, T)\}$  is the maximal detected percentage of the Ca atoms in  $T$  structure that are within a distance threshold of  $d$  Å from  $M$  structure upon calculated  $s \in X(M, T)$  superpositions.

$LGA\_S(M, T)$  structure similarity scoring function is defined as a function of two structures  $M$  and  $T$  calculated as a combination of  $R(r, M, T)$  results from  $LCS(M, T)$  calculations, and  $D(d, M, T)$  results from  $GDT(M, T)$ :

$$LGA\_S(M, T) = (1 - w) * S(LCS(M, T)) + w * S(GDT(M, T))$$

where

$$S(LCS(M, T)) = \frac{2}{n \times (n + 1)} \sum_{j=1}^n (n - j + 1) * R(r_j, M, T),$$

$$n = 3, r_j = 1.0, 2.0, 5.0,$$

$$S(GDT(M, T)) = \frac{2}{k \times (k + 1)} \sum_{j=1}^k (k - j + 1) * D(d_j, M, T),$$

$$k = 20, d_j = 0.5, 1.0, \dots, 10.0,$$

and  $w = 0.75$  is a parameter ( $0 \leq w \leq 1$ ) representing a weighting factor between  $S(LCS)$  and  $S(GDT)$  results.  $S(LCS)$  is a weighted sum of  $R(r, M, T)$  values calculated for  $n$  different RMSD cutoffs  $r$  (e.g.  $n = 3$ ;  $r = 1.0, 2.0, 5.0$ ), and  $S(GDT)$  is a weighted sum of  $D(d, M, T)$  values calculated using  $k$  different distance cutoffs  $d$  (e.g.  $k = 20$ ;  $d = 0.5, 1.0, \dots, 10.0$ ). In the formulae  $S(LCS)$  and  $S(GDT)$ , the weighting schemes weight higher those  $R$  and  $D$  results that were calculated for smaller RMSD and distance cutoffs, respectively.

The range of the  $LGA\_S$  values is 0–100, and hierarchical clustering experiments performed on various

folds from SCOP database showed that  $LGA\_S$  alone can serve as a good discriminator for the initial protein structure clustering (see the results shown in Figure 5a).

### STRALCP clustering approach (similarity in the set of structurally conserved local regions)

The essence of the STRALCP algorithm is the ability to compare hundreds of protein structures in a single reference frame, identify similar fragments that are conserved within a set of analyzed proteins, and use this information to calculate the number of required clusters. Each calculated cluster is assigned with its structural fingerprint that can be defined by a representative structure and a set of spans that are shared among structures grouped together. Comparison of a new structure with a structural fingerprint determines whether the structure should be included to the particular cluster or whether it should be a member of another cluster. Our STRALCP algorithm, which automatically clusters proteins and identifies representative structures, can be described as the following list of steps:

- (i) LGA is used to perform all-against-all comparisons in which, for a given set of structures, each structure is used as a frame of reference for comparisons with others.
- (ii) Each frame of reference is assigned a set of sequential fragments, which are defined by splitting the corresponding amino acid sequence into consecutive  $n$ -residue-long sub-sequences ( $n = 10$  is used as a default parameter; e.g. a 120-residue-long protein comprises 12 fragments).
- (iii) After performing all-against-all structure comparisons (step 1) the following information is assigned to each frame of reference:
  - (a)  $LGA\_S$  values between the frame of reference and all other structures,
  - (b) the number of residue pairs that are superimposed locally within RMSD cutoff 0.5 Å (using 3-residue-long window). Continuous structural segments formed by such residue pairs that are at least five residues long are marked as candidate spans,
  - (c) the number of non-empty fragments (non-empty fragments are sequential fragments defined in step 2 that overlap by at least one residue with at least one detected span in compared structures).
- (iv) For each frame of reference, all structures having at least 80% (default parameter) of the non-empty fragments in common are identified. A list consisting of maximum number of such structures is created and assigned to each frame of reference.
- (v) An optimal number of clusters is determined based on the following criteria:
  - (a) the minimum number of clusters that yields a complete set of proteins in the combined lists from (4),

- (b) *LGA\_S* between any pair of proteins from the cluster is at least 60% (default parameter), minimum value from (iii.a).
- (vi) Within each cluster, a representative structure is selected, which in comparison with other members of the cluster has the highest values determined in steps (iii.a), (iii.b) and (iii.c).

*Note:* In step (v.a) a minimum number of clusters are defined based on local similarities in non-empty fragments along the protein sequence using initially selected representative frames of reference. Step (v.b) allows reassignment of less similar structures from one cluster to another. It also allows sub-division of clusters in order to satisfy the requirement that within each cluster any pair of proteins has at least 60% overall structure similarity. This way less similar structures are not grouped together even if they satisfy the requirement regarding a common set of non-empty fragments (step 4).

## RESULTS

A proper protein classification is critical for better understanding and prediction of a protein's structure, function and interaction with other proteins. It is known that sequence similarities nearly always correspond to structure similarities, enabling structure and function prediction for uncharacterized proteins. Structural similarity, however, does not necessarily correspond to sequence similarity (Figure 2). Through structural

comparison and classification, we identified a family of crystal structures that failed to be detected (18) by sequence-based methods like PSI-BLAST (14). Using a structure-based method (e.g. DALI, LGA) it was found that three EAP domains from *Staphylococcus aureus* (18), which could not be properly classified by sequence-based methods, shared a previously unrecognized similarity to another class of bacterial toxins.

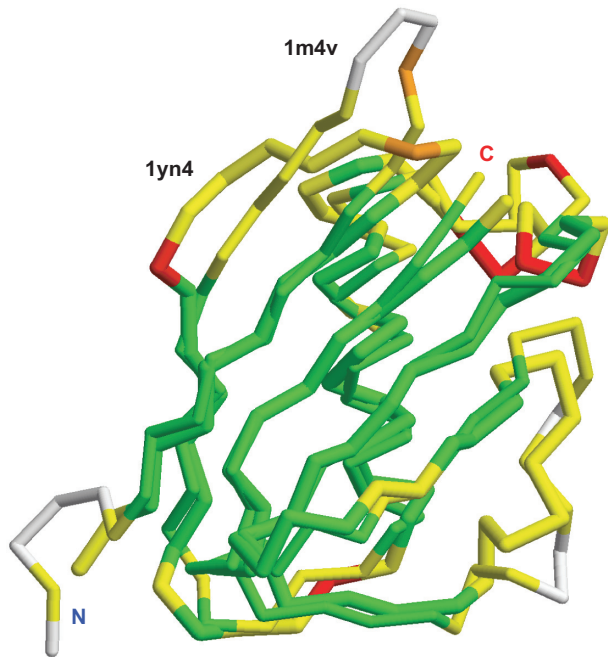
Here we present our structure classification approach, STRALCP, applied to these domains. For each of the EAP domains [Eap2 (PDB entry: 1yn3), EapH1 (1yn4), EapH2 (1yn5)], we have performed structural PDB searches using our LGA server (19). As a result, 134 domains from the SCOP superfamily d.15.6 (Superantigen toxins, C-terminal domain) were identified as most similar to EAP structures (only 20 structures are shown in provided Figures 1, 3, and 4; 3 EAP domains and 17 domains from SCOP). Figure 1 shows that all 20 proteins are very similar in detected structurally conserved frames formed by 4 strands and 1 helix (Figure 2). The superposition of 1yn4\_A and d1m4va2 (1m4v\_A in Figure 2) corresponds to the fourth bar in Figure 1 and shows that these two structures differ in several loop regions only (structural deviations above 2 Å are colored in yellow-red). Note that the level of sequence identity between these two proteins is only ~14% (*Seq\_ID*), whereas the level of structure similarity is ~75% (*LGA\_S*).

In general, all EAP domains have a high level of overall structure similarity (*LGA\_S* over 60%) to most of the other analyzed structures, whereas the level of sequence identity is very low (below 20%). In Figure 1, we show



**Figure 1.** Structure similarities between EAP domains from *S. aureus* (PDB: 1yn3, 4 and 5) and 17 protein domains from the SCOP superfamily comprising superantigen toxins. All proteins were compared to the structure of EapH1 (1yn4\_A), which serves as a frame of reference. Colored bars represent *Calpha-Calpha* distance deviation between 1yn4\_A [99 residues; from the left (N-terminal) to the right (C-terminal)] superimposed with 20 structures from PDB (first bar represents a 1yn4\_A–1yn4\_A self-comparison). Colors represent distances between aligned residues and range from green (below 2 Å) to red (above 6 Å). The columns at the right contain information about the level of sequence identity (*Seq\_ID*) and structure similarity (*LGA\_S*).

the results from the structure comparisons of the set of selected 20 proteins when structure 1yn4\_A was chosen as a frame of reference and in Figure 3 the structure SEH (PDB: 1f77; SCOP domain d1f77a2) (20) was selected as a frame of reference. From the comparison of these two plots we can conclude that d1f77a2 may serve as a better representation (average structure) for the analyzed set



**Figure 2.** A 3D plot of structural superposition between 1yn4\_A and 1m4v\_A (SCOP domain: d1m4va2) that corresponds to the fourth colored bar in Figure 1. The level of sequence identity between proteins *Seq\_ID*: ~14%, and the level of structure similarity *LGA\_S*: ~75%.

of 20 proteins (at least for the top 13 of them) than the structure 1yn4\_A.

The obtained results suggest that a given set of 20 structures can be structurally divided into at least two clusters. Our STRALCP system creates such a clustering automatically (Figure 4). By this clustering the EAP structures: 1yn3-5 are grouped together (Cluster2) with four other protein structures: SET1 (PDB: 1vlp) (21), SET3 (PDB: 1m4v) (22), and TSST1 (PDB: 1aw7, 2tss) (23,24). Additional tests showed that if we had introduced more strict structure similarity requirements [e.g. *LGA\_S* cutoff 80% (see step 5.b in STRALCP algorithm)], then Cluster2 would have been split into two additional clusters (data not shown) where all three EAP domains (1yn3-5) were separated from the SET1, SET3 and TSST1 structures.

**Performance**

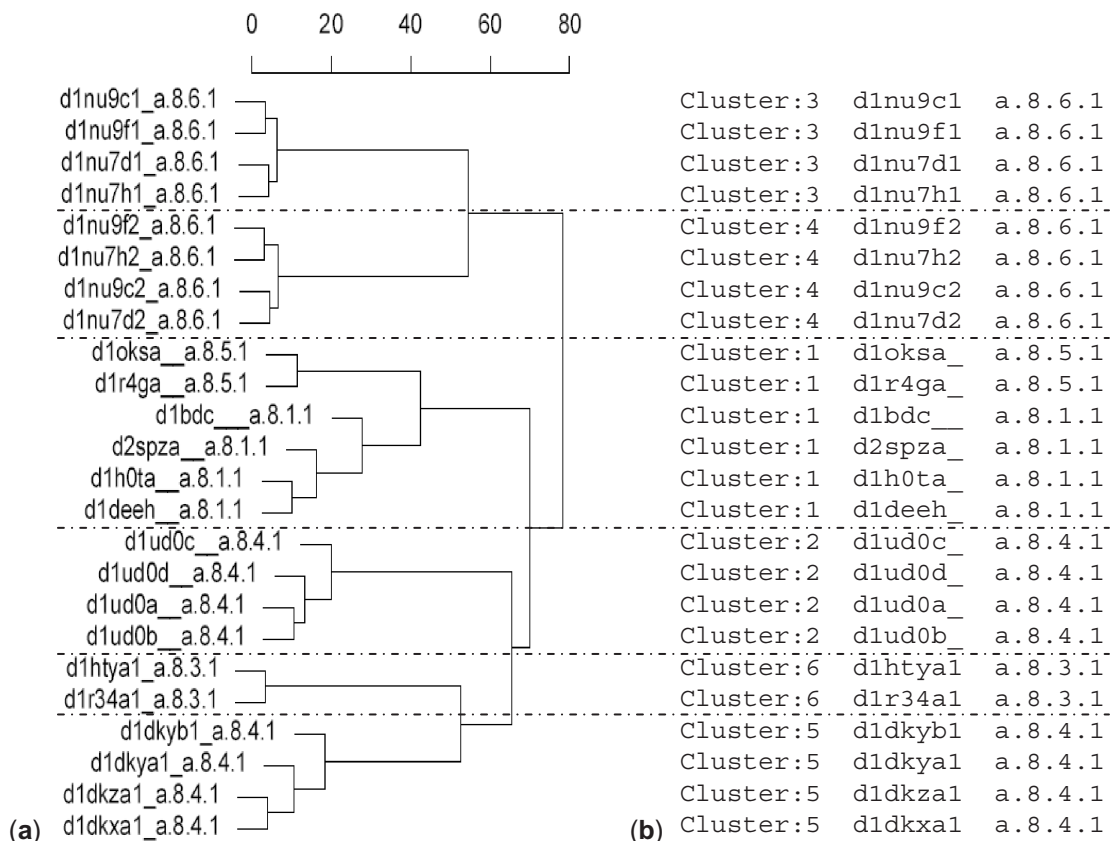
As described in the Methods section the STRALCP clustering approach consists of two steps: (i) calculating all-against-all structural alignments using LGA program, and (ii) extracting from calculated alignments structurally conserved regions and using them to group proteins accordingly. The most cpu expensive is the first step. For example, on a single linux workstation equipped with AMD-64 5000 dual core processor, the calculations of all-against-all pairwise structural alignments for 20 discussed above structures (3 EAP domains and 17 domains from SCOP) lasted about 10min, while the clustering (step 2) was completed in less than 10s.

In order to estimate the accuracy of a STRALCP clustering approach, we have performed comparisons with the SCOP (ver. 1.71) classification. We tested STRALCP calculations on domains from 25 different SCOP folds: a.5, a.7, a.8, a.24, a.29, a.137, b.2, b.42, b.43, b.68, b.80,

PD B	Seq_ID	LGA_S
d1f77a2	100.00	100.00
d1ewca2	99.12	99.86
d1esfa2	40.74	92.06
d1lo5d2	39.81	91.93
d1hqrd2	30.91	90.73
d1et9a2	34.26	90.51
d1bxta2	32.73	88.99
d1goza2	33.03	88.88
d1fnua2	36.94	88.08
d1uupa2	39.09	87.87
d1ty0a2	40.00	87.59
d1ck1a2	29.09	87.30
d1aw7a2	20.20	80.56
d1sebd2	32.32	79.15
d1v1pa2	24.49	78.39
d2tssa2	20.41	74.43
d1m4va2	19.59	74.17
1yn5_A	18.56	60.91
1yn3_A	17.78	57.20
1yn4_A	13.33	57.17

**Figure 3.** The results from the analysis of structure similarities between EAP domains from *S. aureus* and proteins from the SCOP superfamily of superantigen toxins (same domains as in Figure 1). SCOP domain d1f77a2 serves as a frame of reference for this comparison. The coloring scheme is the same as in Figures 1 and 2.





**Figure 5.** (a) Dendrogram showing the results of an *LGA\_S*-based (single measure) clustering of 24 SCOP domains from fold a.8. Each code (entry\_family) represents one protein from the SCOP classification: entry and family number. We used the R package (version 2.1.1; <http://www.r-project.org/>) for the hierarchical clustering and visualization of calculated *LGA\_S* results from all-against-all structure comparisons. (b) Clustering created using STRALCP algorithm with default cutoff *LGA\_S* = 60%.

even if it is based purely on structure comparisons, exhibits a low (on average ~3%) misclustering effect: domains from different SCOP families were clustered separately. It is important to keep in mind that a purely structure-based approach to clustering may result in two proteins that are identical in sequence being clustered separately if the two structures differ in conformation; we observed that the STRALCP algorithm is able to detect the structural differences between domains from the same SCOP family and cluster them separately. It is for this reason that our clustering approach may produce more clusters than the number of SCOP families. For example the family a.8.6.1 (Figure 5b) was separated by STRALCP into two clusters: cluster3 (*Staphylocoagulase* first domain) and cluster4 (*Staphylocoagulase* second domain), and the family a.8.4.1 was divided into two clusters: cluster5 (DnaK domain from *Escherichia coli*) and cluster2 (DnaK domain from Rat). The STRALCP algorithm will also group proteins in different clusters if they significantly differ in length or if multi-domain structures are in different conformations (e.g. ‘open’ and ‘closed’ versions of the same protein). We also can observe additional sub-clustering of protein families when criteria for structure comparison are sufficiently stringent (e.g. a higher *LGA\_S* cutoff is introduced). We consider this ability a beneficial

one to the developed STRALCP approach. It provides valuable information about the regions that are structurally in the same conformation, which could be useful in various studies and classification schemes. The separation of similar or identical proteins, but in different structural conformations, could be reduced by introducing a sequence similarity analysis into the STRALCP algorithm. However, in this study, in order to detect the limits of purely structure-based approaches we do not include sequence information to the scoring and clustering algorithm. The sequence-based analysis may be considered as an option in future development efforts.

## ACKNOWLEDGEMENTS

This work was performed under the auspices of the US Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48. B.K. was supported by DOE CSGF fellowship under grant number DE-FG02-97ER25308. The design and development of the described system was supported by LLNL LDRD grant 04-ERD-068 to A.Z. Funding to pay the Open Access publication charges for this article was provided by US Department of Energy.

*Conflict of interest statement.* None declared.

## REFERENCES

- Dietmann,S. and Holm,L. (2001) Identification of homology in protein structure classification. *Nat. Struct. Biol.*, **8**, 953.
- Redfern,O., Grant,A., Maibaum,M. and Orengo,C. (2005) Survey of current protein family databases and their application in comparative, structural and functional genomics. *J. Chromatogr. B*, **815**, 97–107.
- Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The protein data bank. *Nucleic Acids Res.*, **8**, 235–242.
- Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
- Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
- Holm,L. and Sander,C. (1993) Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.*, **233**, 123–138.
- Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.
- Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.*, **11**, 739–747.
- Domingues,F.S., Rahnenfuhrer,J. and Lengauer,T. (2004) Automated clustering of ensembles of alternative models in protein structure databases. *Protein Eng.*, **17**, 537–543.
- Rogen,P. and Fain,B. (2003) Automatic classification of protein structure by using Gauss integrals. *Proc. Natl Acad. Sci. USA*, **100**, 119–124.
- Aung,Z. and Tan,K.L. (2005) Automatic 3D protein structure classification without structural alignment. *J. Comp. Biol.*, **12**, 1221–1241.
- Frishman,D. and Argos,P. (1995) Knowledge-based secondary structure assignment. *Proteins Struct. Funct. Genet.*, **23**, 566–579.
- Cheek,S., Qi,Y., Krishna,S., Kinch,L. and Grishin,N. (2004) SCOPmap: automated assignment of protein structures to evolutionary superfamilies. *BMC Bioinformatics*, **5**, 197.
- Altschul,S.F., Madden,T.L., Scaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.
- Huan,J., Wang,W., Washington,A., Prins,J., Shah,R. and Tropsha,A. (2004) Accurate classification of protein structural families using coherent subgraph analysis. *Pacific Symp. Biocomput.*, **9**, 411–422.
- Zemla,A. (2003) LGA – a method for finding 3D similarities in protein structures. *Nucleic Acids Res.*, **31**, 3370–3374.
- Geisbrecht,B.V., Hamaoka,B.Y., Perman,B., Zemla,A. and Leahy,D.J. (2005) The crystal structures of EAP domains from *Staphylococcus aureus* reveal an unexpected homology to bacterial superantigens. *J. Biol. Chem.*, **280**, 17243–17250.
- Zemla,A., Ecale Zhou,C., Slezak,T., Kuczmarowski,T., Rama, D., Torres,C., Sawicka,D. and Barsky,D. (2005) AS2TS system for protein structure modeling and analysis. *Nucleic Acids Res.*, **33**, W111–W115.
- Hakansson,M., Petersson,K., Nilsson,H., Forsberg,G., Bjork,P., Antonsson,P. and Svensson,L.A. (2000) The crystal structure of staphylococcal enterotoxin H: implications for binding properties to MHC class II and TcR molecules. *J. Mol. Biol.*, **302**, 527–537.
- Al-Shangiti,A., Naylor,C., Nair,S., Briggs,D., Henderson,B. and Chain,B. (2004) Structural relationships and cellular tropism of staphylococcal superantigen-like proteins. *Infect. Immun.*, **72**, 4261–4270.
- Arcus,V.L., Langley,R., Proft,T., Fraser,J.D. and Baker,E.N. (2002) The three-dimensional structure of a superantigen-like protein, SET3, from a pathogenicity island of the *Staphylococcus aureus* genome. *J. Biol. Chem.*, **277**, 32274–32281.
- Earhart,C.A., Mitchell,D.T., Murray,D.L., Pinheiro,D.M., Matsumura,M., Schlievert,P.M. and Ohlendorf,D.H. (1998) Structures of five mutants of toxic shock syndrome toxin-1 with reduced biological activity. *Biochemistry*, **37**, 7194–7202.
- Prasad,G.S., Radhakrishnan,R., Mitchell,D.T., Earhart,C.A., Dinges,M.M., Cook,W.J., Schlievert,P.M. and Ohlendorf,D.H. (1997) Refined structures of three crystal forms of toxic shock syndrome toxin-1 and of a tetramer with reduced activity. *Protein Sci.*, **6**, 1220–1227.