UCRL-TR-218924

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Protein Classification Based on Analysis of Local Sequence-Structure Correspondence

A. T. Zemla

February 13, 2006

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

Auspices Statement

This work was performed under the auspices of the U.S. Department of Energy by University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48.

Protein Classification Based on Analysis of Local Sequence-Structure Correspondence

Final Report Authors: Adam Zemla

Principal Investigator: Adam Zemla
Co-investigators: Carol Zhou, Jason Smith, Marisa Lam
Tracking Code: 04-ERD-068
Primary Category of Work: Biological Sciences
Type of Research: Basic

Summer students working on the project:
Summer 2004, Davinder Rama (student at California State University)
Summer 2005, Bonnie Kirkpatrick (Ph.D. student at UC Berkeley)

Project Description

The goal of this project was to develop an algorithm to detect and calculate common structural motifs in compared structures, and define a set of numerical criteria to be used for fully automated motif based protein structure classification. The Protein Data Bank (PDB) contains more than 33,000 experimentally solved protein structures, and the Structural Classification of Proteins (SCOP) database, a manual classification of these structures, cannot keep pace with the rapid growth of the PDB. In our approach called STRALCP (STRucture Alignment based Clustering of Proteins), we generate detailed information about global and local similarities between given set of structures, identify similar fragments that are conserved within analyzed proteins, and use these conserved regions (detected structural motifs) to classify proteins. Our developed algorithm for automatic classification of proteins reflects the manual classification.

Expected Results

The software and database resulting from this project demonstrate how the problem of automation of the protein structure classification can be solved. The ability to verify sequence-based alignments by comparison to the correctly calculated structural alignments significantly improves the quality of protein modeling, function recognition, and identification of regions on protein surfaces as candidates for ligand binding sites. Because accurate structural analysis is requisite to computational protein-based detection schemes, this work will improve the success rate and reduce the cost for choosing regions in proteins for antibody or high-affinity ligand recognition, and will improve our ability to identify possibly cross-reactive proteins related to the protein targeted for detection. The part of our protein structure comparison and analysis system is already made accessible to the scientists through the web based interface at http://as2ts.llnl.gov/AS2TS bringing positive recognition and visibility to LLNL and DOE.

Mission Relevance

The proposed protein structure comparison system will enhance the accuracy of protein classification and the quality of modeled protein structures. It implies the applicability for research related to the Laboratory's biodefense mission. This project leverages LLNL's capabilities in bioinformatics and high-speed computing; and the mission relevance to biodefense has been shown numerous times in FY04-FY05. Our developing STRALCP database of protein sequence-structure motifs used for protein structure classification improves protein modeling capabilities that enable us to predict more high-quality protein signature targets for pathogens of interest. The achieved capabilities of our system have been applied to improve structural models of critical proteins of causative agents of smallpox, ricin, plague, foot and mouth disease (FMD), monkeypox, and others. The generated models were used to predict the regions in protein structures upon which DNA and protein signatures designed at LLNL land, to determine potential unique protein signature candidates, to identify promising vaccine targets, and to suggest probable functions of unknown proteins. Exploitation of these models is ongoing at LLNL and by collaborators.

Accomplishments and Results

This project builds on LLNL's capabilities in bioinformatics and high-speed computing, and enhances biodefense capabilities at the Laboratory by providing automated system of protein structure classification. We have developed a protein structure comparison algorithm to generate information about sequence-structure correspondence between related proteins. Our STRALCP system is capable to evaluate the level of overall structure similarity, and also to generate detailed information about the regions of local similarities between compared structures. We have designed a set of numerical criteria that use detected structurally conserved regions for automated protein structure classification. We have developed a prototype of protein structure database where proteins are clustered based on their similarity in identified structural motifs. Our automated clustering method detects relationships between proteins on the level of structural families with a very good agreement with manual SCOP classification. The developed automated structure classification capabilities will allow for better protein annotation of many microbes being studied in collaborative work with scientific groups from LLNL and other laboratories.

Introduction:

There are many ways how a given set of protein structures could be clustered. Depends on the applied algorithm the results of the classification may differ significantly if different numerical criteria are used to assess the level of similarity between compared structures or if applied clustering criteria are focused on different features in protein structures. To perform particular clustering a suitable scoring function (or, in general, a scoring algorithm that takes into account a number of different features from compared proteins) has to be defined. Depending on the goal of the clustering it can be done by selecting one measure or by combining different criteria to assess (score) the level of similarity between analyzed proteins. The goal of our research was to define criteria and to develop an algorithm for automatic classification of proteins that reflects the manual SCOP classification. In our approach, called STRALCP, we generate detailed information about global and local similarities between any pair of analyzed protein structures, identify similar fragments that are conserved within analyzed proteins, and use such conserved regions to classify proteins according to their similarities in the detected structural motifs (spans). Our approach also allows automated detection of structural and sequence deviations within analyzed family or set of proteins.

Software Development:

(1) LGA_S structure similarity scoring function (overall similarity). LGA (Local and Global Alignment) program enables searching for the regions of local and global similarities and the "best" structure superposition between two protein structures. In order to consider local regions of the proteins in assessing their similarity a new scoring function has been implemented in LGA program. 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 (e.g. M-model and T-target). In comparing two protein structures, the LCS procedure is able to localize and superimpose the longest segments of residues that can fit under a selected RMSD (root mean square deviation) cutoff. 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 a specified *distance* cutoff. Let:

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

LGA_S structure similarity scoring function is defined as a function of two structures M and T calculated as a combination of R(v) results from LCS calculations using set of n RMSD cutoffs v (e.g. n=3; v = 1.0, 2.0, 5.0), and D(v) results from GDT calculations using the set of k thresholds v (e.g. k=20; v = 0.5, 1.0, …, 10.0):

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

where:

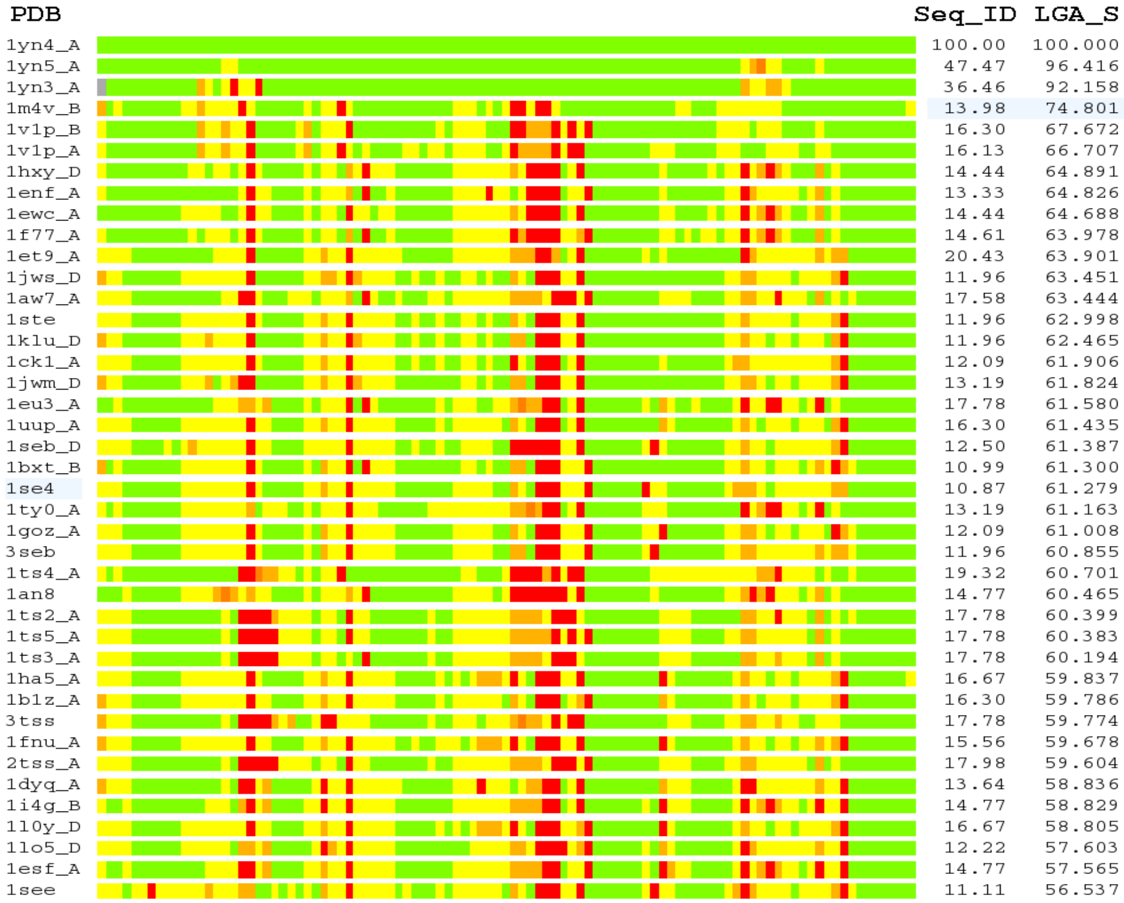
$$\mathbf{S(LCS) = \frac{2}{n \cdot (n+1)} \sum_{j=1}^{n} (n - j + 1) * R(v_j), n = 3, v_j = 1.0, 2.0, 5.0,}$$

$$\mathbf{S(GDT) = \frac{2}{k \cdot (k+1)} \sum_{i=1}^{k} (k - i + 1) * D(v_i), k = 20, v_i = 0.5, 1.0, \ldots, 10.0,}$$

and w=0.75 is a parameter (0<=w<=1) representing a weighting factor between LCS and GDT results.

The initial version of LGA program has been published in Nucleic Acids Research (A. Zemla: "LGA - a method for finding 3D similarities in protein structures", Nucleic Acids Research, Vol. 31, No. 13, 2003, pp. 3370-3374), and the Record of the Invention and the Patent Application has been submitted to Intellectual Property Law Group at LLNL.

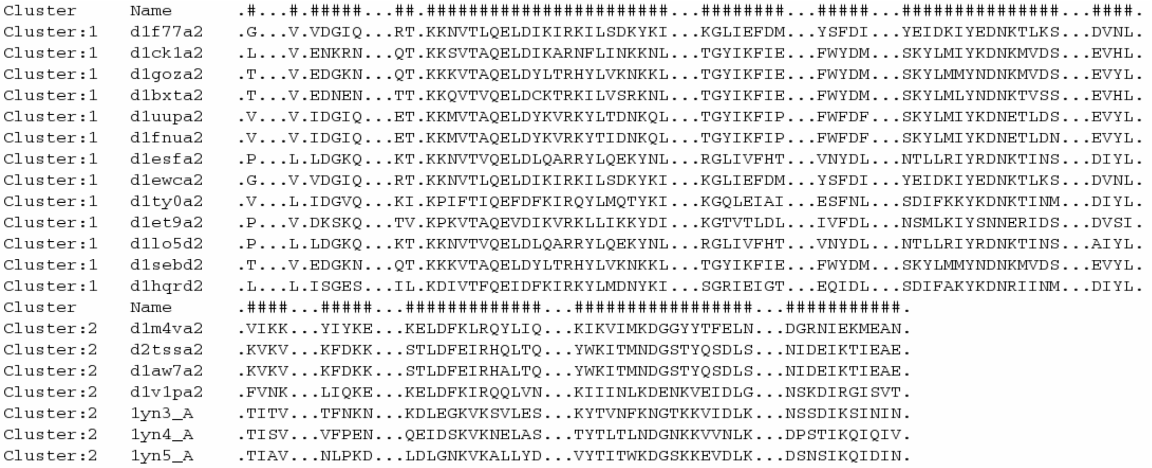
(2) Detection of structurally conserved regions (similarity in the set of local regions). The cornerstone of our STRALCP algorithm is the ability to compare hundreds of protein structures in a single reference frame (target protein) and identify similar fragments that are conserved within a set of analyzed proteins. As an example of using our algorithm we show the results from the analysis of structure similarities between EAP domains from Staphylococcus Aureus (Eap2 (PDB entry: 1yn3), EapH1 (1yn4), EapH2 (1yn5)) and other proteins from PDB. On the plot below we show the set of PDB structures that were detected as most similar to EAP by our system. All identified by our algorithm proteins belong to one SCOP's superfamily called “Superantigen toxins, C-terminal domain”.



PDB	Seq_ID	LGA_S
1yn4_A	100.00	100.000
1yn5_A	47.47	96.416
1yn3_A	36.46	92.158
1m4v_B	13.98	74.801
1v1p_B	16.30	67.672
1v1p_A	16.13	66.707
1hxy_D	14.44	64.891
1enf_A	13.33	64.826
1ewc_A	14.44	64.688
1f77_A	14.61	63.978
1et9_A	20.43	63.901
1jws_D	11.96	63.451
1aw7_A	17.58	63.444
1ste	11.96	62.998
1klu_D	11.96	62.465
1ck1_A	12.09	61.906
1jwm_D	13.19	61.824
1eu3_A	17.78	61.580
1uup_A	16.30	61.435
1seb_D	12.50	61.387
1bxt_B	10.99	61.300
1se4	10.87	61.279
1ty0_A	13.19	61.163
1goz_A	12.09	61.008
3seb	11.96	60.855
1ts4_A	19.32	60.701
1an8	14.77	60.465
1ts2_A	17.78	60.399
1ts5_A	17.78	60.383
1ts3_A	17.78	60.194
1ha5_A	16.67	59.837
1b1z_A	16.30	59.786
3tss	17.78	59.774
1fnu_A	15.56	59.678
2tss_A	17.98	59.604
1dyq_A	13.64	58.836
1i4g_B	14.77	58.829
1l0y_D	16.67	58.805
1lo5_D	12.22	57.603
1esf_A	14.77	57.565
1see	11.11	56.537

The plot above shows that all analyzed proteins are very similar (structurally conserved) in detected core regions (green). Colored bars represent *Calpha* - *Calpha* distance deviation between superimposed PDB structures and 1yn4_A (99 residues; from the left (N terminal) to the right (C terminal)). The distances between aligned residues are represented using different colors (from green – distances below 2Å to red – above 6Å). The columns on the right side of the colored bars contain information about the level of sequence identity (Seq_ID), and level of calculated structure similarity (LGA_S). It is important to notice that the level of sequence identity (Seq_ID column) between compared proteins is very low (below 20%), so this classification was not possible to achieve by using sequence-based only techniques (e.g. PSI-BLAST analysis).

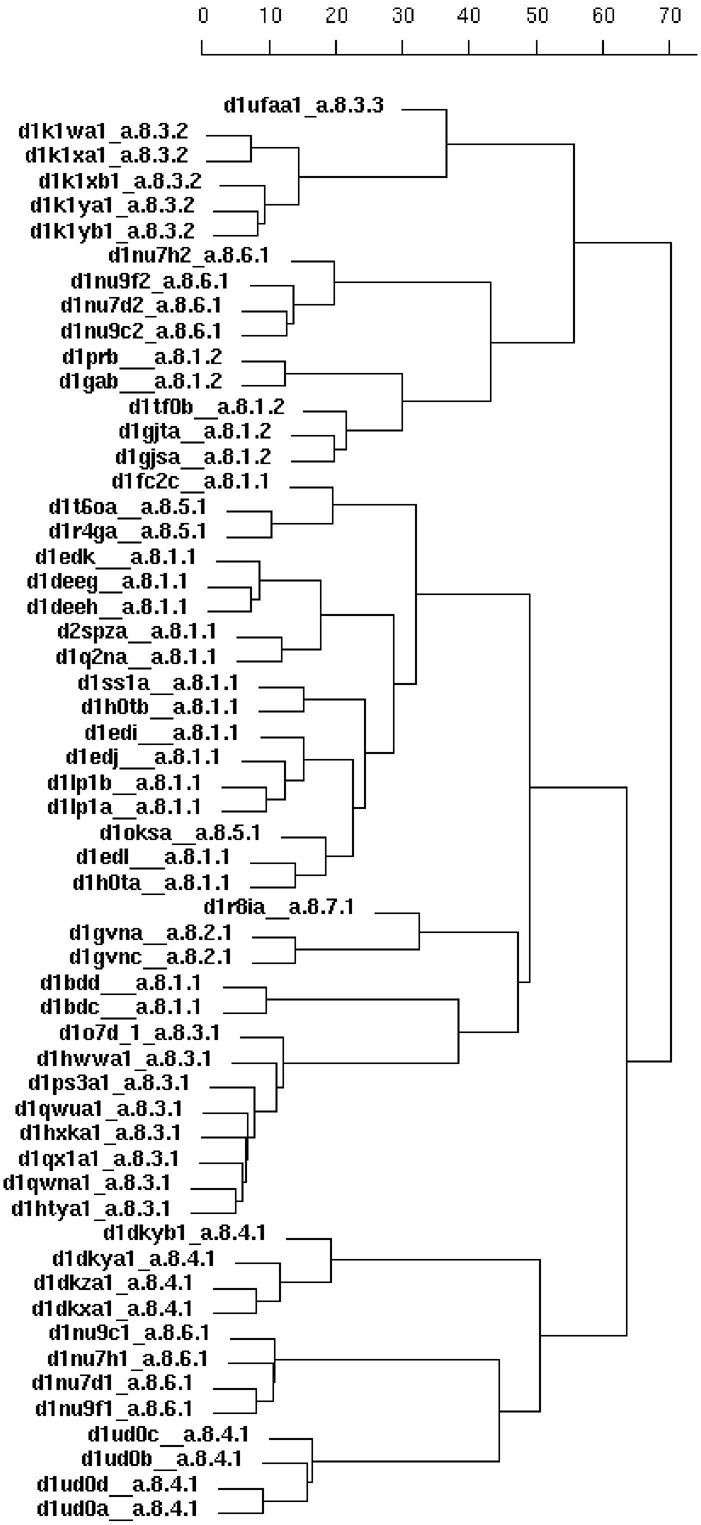
(3) STRALCP clustering algorithm. In our STRALCP approach we identify conserved regions and use them to classify proteins and protein domains according to their similarities in the detected structural motifs. For a given group of proteins and a target protein, the LGA algorithm creates one structural alignment of each protein to the target protein (see (2)). We cluster the proteins according to the structural regions they share with the target. Taking each of the proteins in turn as the target yields an ensemble of clusters, multiple partitions on the same set of proteins. Discrepancies are resolved by grouping together proteins that clustered together across many of the partitions. Each cluster of structures is defined by the representative structure and the set of shared structural features that we call a fingerprint. Comparison with the structural fingerprint determines whether a given structure belongs to the cluster. The clusters of proteins and the representative structures are created automatically. Below we show an example of using our algorithm where 1yn3-5 structures are clustered together (Cluster2) with four other protein structures: SET3 (PDB: 1m4v), SET1 (PDB: 1v1p), and TSST1 (PDB: 1aw7, 2tss).



```
Cluster     Name       .#...#.#####...##.#######################...########...#####...###############...####.
Cluster:1   d1f77a2    .G...V.VDGIQ...RT.KKNVTLQELDIKIRKILSDKYKI...KGLIEFDM...YSFDI...YEIDKIYEDNKTLKS...DVNL.
Cluster:1   d1ck1a2    .L...V.ENKRN...QT.KKSVTAQELDIKARNFLINKKNL...TGYIKFIE...FWYDM...SKYLMIYKDNKMVDS...EVHL.
Cluster:1   d1goza2    .T...V.EDGKN...QT.KKKVTAQELDYLTRHYLVKNKKL...TGYIKFIE...FWYDM...SKYLMMYNDNKMVDS...EVYL.
Cluster:1   d1bxta2    .T...V.EDNEN...TT.KKQVTVQELDCKTRKILVSRKNL...TGYIKFIE...FWYDM...SKYLMLYNDNKTVSS...EVHL.
Cluster:1   d1uupa2    .V...V.IDGIQ...ET.KKMVTAQELDYKVRKYLTDNKQL...TGYIKFIP...FWFDF...SKYLMIYKDNETLDS...EVYL.
Cluster:1   d1fnua2    .V...V.IDGIQ...ET.KKMVTAQELDYKVRKYTIDNKQL...TGYIKFIP...FWFDF...SKYLMIYKDNETLDN...EVYL.
Cluster:1   d1esfa2    .P...L.LDGKQ...KT.KKNVTVQELDLQARRYLQEKYNL...RGLIVFHT...VNYDL...NTLLRIYRDNKTINS...DIYL.
Cluster:1   d1ewca2    .G...V.VDGIQ...RT.KKNVTLQELDIKIRKILSDKYKI...KGLIEFDM...YSFDI...YEIDKIYEDNKTLKS...DVNL.
Cluster:1   d1ty0a2    .V...L.IDGVQ...KI.KPIFTIQEFDFKIRQYLMQTYKI...KGQLEIAI...ESFNL...SDIFKKYKDNKTINM...DIYL.
Cluster:1   d1et9a2    .P...V.DKSKQ...TV.KPKVTAQEVDIKVRKLLIKKYDI...KGTVTLDL...IVFDL...NSMLKIYSNNERIDS...DVSI.
Cluster:1   d1lo5d2    .P...L.LDGKQ...KT.KKNVTVQELDLQARRYLQEKYNL...RGLIVFHT...VNYDL...NTLLRIYRDNKTINS...AIYL.
Cluster:1   d1sebd2    .T...V.EDGKN...QT.KKKVTAQELDYLTRHYLVKNKKL...TGYIKFIE...FWYDM...SKYLMMYNDNKMVDS...EVYL.
Cluster:1   d1hqrd2    .L...L.ISGES...IL.KDIVTFQEIDFKIRKYLMDNYKI...SGRIEIGT...EQIDL...SDIFAKYKDNRIINM...DIYL.
Cluster     Name       .####...#####...#############...#################...###########.
Cluster:2   d1m4va2    .VIKK...YIYKE...KELDFKLRQYLIQ...KIKVIMKDGGYYTFELN...DGRNIEKMEAN.
Cluster:2   d2tssa2    .KVKV...KFDKK...STLDFEIRHQLTQ...YWKITMNDGSTYQSDLS...NIDEIKTIEAE.
Cluster:2   d1aw7a2    .KVKV...KFDKK...STLDFEIRHALTQ...YWKITMNDGSTYQSDLS...NIDEIKTIEAE.
Cluster:2   d1v1pa2    .FVNK...LIQKE...KELDFKIRQQLVN...KIIINLKDENKVEIDLG...NSKDIRGISVT.
Cluster:2   1yn3_A     .TITV...TFNKN...KDLEGKVKSVLES...KYTVNFKNGTKKVIDLK...NSSDIKSININ.
Cluster:2   1yn4_A     .TISV...VFPEN...QEIDSKVKNELAS...TYTLTLNDGNKKVVNLK...DPSTIKQIQIV.
Cluster:2   1yn5_A     .TIAV...NLPKD...LDLGNKVKALLYD...VYTITWKDGSKKEVDLK...DSNSIKQIDIN.
```

The biological functions of proteins 1yn3-5 are still unknown, but the results from calculated structure classification suggest close homology to four other structures from PDB. These results will guide future experimental work to help understand the functional role of these proteins.

Performed experiments. In our tests we have performed STRALCP calculations for over 50 SCOP families. The “quality” of the data deposited in PDB is comprised of protein structures having resolutions ranging from 0.54Å (X-ray crystallography) to greater than 15Å (electron microscopy). Despite this noise, the robust nature of our clustering method detects relationships on the level of the SCOP family with very good agreement with manually maintained SCOP classification. Below is an example that shows the difference in two approaches: when a single measure is used to evaluate the overall level of structure similarity versus the multiple criteria based clustering (as implemented in STRALCP). R version 2.1.1, a computer language and environment for statistical computing and graph programming (see http://www.r-project.org/) was used for hierarchical clustering of analyzed structures shown as a dendrogram in the provided example.



```
Cluster:1   d1deeh_   a.8.1.1
Cluster:1   d1deeg_   a.8.1.1
Cluster:1   d1fc2c_   a.8.1.1
Cluster:1   d1edj__   a.8.1.1
Cluster:1   d1edl__   a.8.1.1
Cluster:1   d1edi__   a.8.1.1
Cluster:1   d1edk__   a.8.1.1
Cluster:1   d1lp1a_   a.8.1.1
Cluster:1   d1lp1b_   a.8.1.1
Cluster:1   d1q2na_   a.8.1.1
Cluster:1   d2spza_   a.8.1.1
Cluster:1   d1ss1a_   a.8.1.1
Cluster:1   d1h0ta_   a.8.1.1
Cluster:1   d1h0tb_   a.8.1.1

Cluster:2   d1ps3a1   a.8.3.1
Cluster:2   d1qwna1   a.8.3.1
Cluster:2   d1qx1a1   a.8.3.1
Cluster:2   d1htya1   a.8.3.1
Cluster:2   d1hxka1   a.8.3.1
Cluster:2   d1hwwa1   a.8.3.1
Cluster:2   d1qwua1   a.8.3.1
Cluster:2   d1o7d.1   a.8.3.1

Cluster:3   d1oksa_   a.8.5.1
Cluster:3   d1t6oa_   a.8.5.1
Cluster:3   d1r4ga_   a.8.5.1

Cluster:4   d1k1xa1   a.8.3.2
Cluster:4   d1k1xb1   a.8.3.2
Cluster:4   d1k1ya1   a.8.3.2
Cluster:4   d1k1yb1   a.8.3.2
Cluster:4   d1k1wa1   a.8.3.2

Cluster:5   d1dkya1   a.8.4.1
Cluster:5   d1dkza1   a.8.4.1
Cluster:5   d1dkxa1   a.8.4.1
Cluster:5   d1dkyb1   a.8.4.1

Cluster:6   d1nu9f1   a.8.6.1
Cluster:6   d1nu9c1   a.8.6.1
Cluster:6   d1nu7d1   a.8.6.1
Cluster:6   d1nu7h1   a.8.6.1

Cluster:7   d1nu7h2   a.8.6.1
Cluster:7   d1nu9c2   a.8.6.1
Cluster:7   d1nu9f2   a.8.6.1
Cluster:7   d1nu7d2   a.8.6.1

Cluster:8   d1ud0a_   a.8.4.1
Cluster:8   d1ud0b_   a.8.4.1
Cluster:8   d1ud0c_   a.8.4.1
Cluster:8   d1ud0d_   a.8.4.1

Cluster:9   d1tf0b_   a.8.1.2
Cluster:9   d1gab__   a.8.1.2
Cluster:9   d1prb__   a.8.1.2
Cluster:9   d1gjta_   a.8.1.2
Cluster:9   d1gjsa_   a.8.1.2

Cluster:10  d1gvnc_   a.8.2.1
Cluster:10  d1gvna_   a.8.2.1

Cluster:11  d1bdd__   a.8.1.1
Cluster:11  d1bdc__   a.8.1.1

Cluster:12  d1r8ia_   a.8.7.1

Cluster:13  d1ufaa1   a.8.3.3
```

The dendrogram (to the left) shows the results of the LGA_S based clustering of SCOP entries from the fold a.8. Each code (entry_family) represents one protein from SCOP classification: entry and family number. Similar dendrograms result from applying one criterion based clustering such as RMSD. On the right side the results of clustering created using STRALCP approach are shown. Using STRALPC (multiple criteria based clustering) we can clearly separate proteins into appropriate clusters that correspond with a very high accuracy to the SCOP families (see the last column to the right which gives SCOP family codes).

Our AS2TS protein structure analysis and modeling system is being used in collaborative work with many scientific groups in their biology research (see [1] - [8]).

Publication/Presentations

Papers:

1) B. V. Geisbrecht, B. Y. Hamaoka, B. Perman, A. Zemla, D. J. Leahy: "Crystal Structures of Eap Domains from Staphylococcus Aureus Reveal an Unexpected Homology to Bacterial Superantigens", J.Biol.Chem, 2005, 280(17), pp. 17243-50. UCRL-JRNL-216376

2) J. B. Pesavento, M. Cosman, A. Zemla, P. T. Beernink, S. L. McCutchen-Maloney, J. P. Fitch, R. Balhorn, D. Barsky: "Identification of a thermo-regulated glutamine-binding protein from Yersinia pestis", Protein Science, (submitted), UCRL-JRNL-213767

3) R. Stanfield, A. Zemla, I.A. Wilson, and B. Rupp: "Antibody elbow angles are influenced by their light chain class", J.Mol.Biol., (in press), UCRL-JRNL-218128

4) P. J. Beuning, S. M. Simon, A. Zemla, D. Barsky, G. C. Walker: “A Non-Cleavable UmuD Variant that Acts as a UmuD’ Mimic”, J.Biol.Chem., (in press), UCRL-JRNL-216587

5) C. Ecale Zhou, A. Zemla, D. Roe, M. Young, M. Lam, J. Schoeniger, R. Balhorn: "Computational approaches for identification of conserved/unique binding pockets in the A chain of ricin", Bioinformatics 2005 21: pp. 3089-3096. UCRL-JRNL-209388

6) A. Zemla, C. Ecale Zhou, T. Slezak, T. Kuczmarski, D. Rama, C. Torres, D. Sawicka, D. Barsky: "AS2TS system for protein structure modeling and analysis", Nucleic Acids Research, 2005, 33, pp. W111-W115. UCRL-JRNL-209684

7) S. D. Goens, S. Botero, A. Zemla, C. Ecale Zhou, M. Perdue: "Bovine enterovirus type 2. Complete genomic sequence and molecular modeling of the reference strain and a wild type isolate from endemically infected US cattle", Journal of General Virology, 85, 2004, pp. 3195-3203. UCRL-JRNL-202639

8) K. A. Kanterdjieff, Ch. Y. Kim, C. Naranjo, G. S. Waldo, T. P. Lekin, B. W. Segelke, A. Zemla, M. S. Park, T. C. Terwilliger, B. Rupp: "Mycobacterium tuberculosis RmlC epimerase (Rv3465): a promising drug-target structure in the rhamnose pathway", Acta Cryst., 2004, D60, pp. 895-902. UCRL-JRNL-202415

Presentations:

9) A. Zemla, C. E. Zhou, M. Lam, J. Smith, B. Kirkpatrick. "A novel structure-driven approach for protein classification", a poster presented at the LLNL CAR Showcase Event, Nov. 3, 2005. UCRL-POST-216262

10) C. Zhou, M. Lam, J. Smith, A. Zemla, and T. Slezak. "Computational approaches for identification of targets for protein-based diagnostics" a poster presented at a meeting sponsored by the Dept. of Homeland
Security in Boston, April 26-28, 2005. UCRL-POST-211564

11) C. E. Zhou, A. Zemla, M. Lam, D. Roe. "Computational approaches for identification of signatures for protein-based diagnostics", a poster presented at a Gordon Conference in Buellton, CA, Jan 31 - Feb 4, 2005. UCRL-POST-209138

12) C. Zhou, M. Lam, A. Zemla, M. Yeh, T. Kuczmarski, C. Torres, J. Smith, T. Slezak, "Computational approaches to assay development for real-time detection of biothreat pathogens", a poster presented at a Keystone Symposium, Keystone, CO, Jan. 6-11, 2004. UCRL-POST-202649.

13) C. Zhou, A. Zemla, B. Vitalis T. Slezak. "Computational approaches for pathogen detection using protein-based signatures", a poster presented at the LLNL CBBB Media Event, Sept. 23, 2004. UCRL-POST-206450

14) C. Zhou, A. Zemla, T. Kuczmarski, M. Lam. "High-throughput selection of protein-based signature targets for detection of bio-threat agents", a poster presented at an American Society for Microbiology conference on Functional Genomics and Bioinformatics Approaches to Infectious Disease Research, Portland OR, October 6-9, 2004. UCRL-POST-207543