STUDIES ON COMBINING SEQUENCE AND STRUCTURE FOR PROTEIN CLASSIFICATION

APPROVED BY SUPERVISORY COMMITTEE

Nick Grishin, PhD. ; Adviser

Steven Altschuler, Ph.D.

Zbyszek Otwinowski, Ph.D.

Rama Ranganathan, M.D., Ph.D.; Committee Chair

To my Parents

Kwang-Youl and OK-Soon Kim

&

my wife

Bo-Mi

&

my daughter

Erin

STUDIES ON COMBINING SEQUENCE AND STRUCTURE FOR PROTEIN CLASSIFICATION

by

BONG-HYUN KIM

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

August, 2009

STUDIES ON COMBINING SEQUENCE AND STRUCTURE FOR PROTEIN CLASSIFICATION

Bong-Hyun Kim, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, 2009

Nick Grishin, Ph.D.

The ultimate goal of our research is to develop a better understanding of how proteins evolve different structures and functions. A large scale protein clustering can provide a useful platform to identify such principles of protein evolution. Manual classification schemes accurately group homologous proteins, but they are slow and subjective. Automatic protein clustering methods are largely based on sequence information. Therefore, they often do not accurately reflect remote homologies that can be recognized by structural information. We hypothesized that combining evolutionary signals from protein sequence and 3D structure will improve automated protein classification. To test this hypothesis, we clustered proteins into evolutionary groups

using both sequence and structure by a fully automated method. We developed a stringent algorithm, self-consistency grouping (SCG) method, which clusters proteins if all the proteins in the group are more similar to each other than to proteins outside the group. Comparison of SCG and other commonly used clustering methods to a widely accepted manual classification scheme, Structural Classification of Protein (SCOP), showed SCG groups to better reflect the reference classification. In depth analysis of SCG clusters highlights new non-trivial evolutionary links between proteins. SCG clustering can be further developed as a reference for evolutionary classification of proteins.

Acknowledgements

I sincerely thank my mentor, Dr. Nick Grishin, for his continuous guidance and encouragements. He taught me the beautiful world of protein structure and evolution.

I am very grateful to the members in my dissertation committee, Drs. Rama Ranganathan, Zbyszek Otwinowski, and Stephen Altschuler, for their helpful advices and discussions.

I thank everybody in Grishin lab. Their help and friendship have made my graduate work and life a lot easier. I thank especially Jimin Pei and Shuoyong Shi for their friendship and help. I also thank for Ming Tang for his support in running computers.

I thank Hua Cheng, Lisa Kinch, and Chalam Chituri for their contributions to the work. Hua showed me how to infer the evolutionary history of proteins. Lisa always gave me good ideas and critical comments for my work and presentations. Chalam helped me express my algorithmic idea in formal mathematical presentation.

I am grateful for the supports of my friends in Korea and in UT Southwestern. Their support greatly helped me to finish the work. I especially thank my friends, Sang-Kwon, Se-Young, Sung-Chun, Joon-Ho, Hae-Joon, Sung-Gon, Ji-Hye, Ji-Sun, Kyung-Ah, Seung-Yun and Hyun-Joo.

Last, but not least, I am grateful to my grandparents, parents, my family. I especially thank my father Kwang-Youl, my mother Ok-Soon, my wife Bomi, and my daughter Erin. Were it not for their love and support, none of my achievements would have been possible.

Table of Contents

ACKNOWLEDGEMENTS ..... **VII**

TABLE OF CONTENTS ..... **VIII**

PRIOR PUBLICATIONS ..... **X**

LIST OF FIGURES ..... **XII**

LIST OF TABLES ..... **XIV**

CHAPTER 1 INTRODUCTION **15**

1.1 HOMOLOGY INFERENCE 15
1.2 PROTEIN CLASSIFICATION 17

CHAPTER 2 METHODS **19**

2.1 DATASET 19
2.1.1 Selection of protein domains *19*
2.1.2 Preparation of protein domain structures and sequences *19*
2.2 GENERATION OF ALIGNMENTS AND SCORES 20
2.2.1 Generation of sequence alignments and scores *20*
2.2.2 Generation of structural alignments and scores *22*
2.2.3 Scaling and normalization *22*
2.3 SELF CONSISTENCY GROUPING 24
2.3.1 Self consistency grouping *24*
2.3.2 Iterative self consistency grouping *29*
2.4 COMPARING A CLUSTERING TO A REFERENCE CLASSIFICATION 30
2.4.1 Reference clustering *30*
2.4.2 Sensitivity *30*
2.4.3 Specificity *31*
2.4.4 F-measure *32*
2.5 INFERENCE OF PHYLOGENETIC TREE 33

CHAPTER 3 DEVELOPMENT OF PROTEIN CLASSIFICATION **34**

3.1 PRELIMINARIES 34
3.1.1 Dataset 34
3.1.2 Scoring schemes 35
3.2 SCG CLUSTERING BASED ON SINGLE MEASURES 36
3.2.1 Cluster size analysis shows that SCG clusters are very small on average. 38
3.2.2 The cluster membership projected on SCOP database shows that cluster contains proteins related by family or subfamily groups. 40
3.3 UNBIASED COMBINATION OF SCORES USING SCG ALGORITHM 43
3.3.1 Average cluster size can be a quality measure for a scoring scheme in SCG clustering. 43
3.3.2 Combining sequence and structural information makes cluster size bigger. 44
3.3.3 Combining sequence and structural information do not make cluster quality worse. 46
3.4 IMPROVEMENT OF SCG ALGORITHM: ITERATIVE SCG (ISCG) 49
3.4.1 iSCG makes cluster size bigger. 50
3.4.2 iSCG makes the coverage bigger but severely reduces specificity. 51
3.5 ESTABLISHMENT OF THE CUTOFF VALUE FOR ISCG 52
3.5.1 Cluster size distribution 53
3.5.2 Comparing to SCOP database 55
3.5.3 Internal score distribution 57

CHAPTER 4 ANALYSIS OF PROTEIN CLASSIFICATION 62

4.1 GLOBAL ANALYSIS OF PROTEIN CLASSIFICATION 62
4.1.1 Network like representation of classification 62
4.1.2 Superclusters 66
4.2 ANALYSIS OF DIFFERENCE BETWEEN SCOP AND ISCG PROTEIN CLASSIFICATION 75
4.2.1 Putative Homologs 77
4.2.2 Similar in fold but unclear in homology 101
4.2.3 Partial similarity, unlikely in homology 112
4.2.4 No Similarity, wrong cluster 116

CHAPTER 5 CONCLUDING REMARKS 118

BIBLIOGRAPHY 122

Prior Publications

Kim BH, Cheng H, Kinch L, Grishin NV. An automatic classification of proteins using sequence and structure by consistency. (In preparation)

Cheng H*, **Kim BH***, Grishin NV. Detecting remote protein homologs using structures and sequences. (In preparation)

Raman S, Vernon R, Thompson J, Tyka M, Sadreyev R, Pei J, Kim D, Kellogg E, Dimaio F, Lange O, Kinch L, Sheffler W, **Kim BH**, Das R, Grishin NV, Baker D. Structure prediction for CASP8 with all-atom refinement using Rosetta. Proteins. 2009 Jul 20. [Epub ahead of print]

Sadreyev RI, **Kim BH**, Grishin NV. Discrete-continuous duality of protein structure space. Curr Opin Struct Biol. 2009 Jun;19(3):321-8. Epub 2009 May 29.

Sadreyev RI, Tang M, **Kim BH**, Grishin NV. COMPASS server for homology detection: improved statistical accuracy, speed and functionality. Nucleic Acids Res. 2009 Jul 1;37(Web Server issue):W90-4. Epub 2009 May 12.

Kim BH, Cheng H, Grishin NV. HorA web server to infer homology between proteins using sequence and structural similarity. Nucleic Acids Res. 2009 Jul 1;37(Web Server issue):W532-8. Epub 2009 May 5.

Xie CQ, Jeong Y, Fu M, Bookout AL, Garcia-Barrio MT, Sun T, **Kim BH**, Xie Y, Root S, Zhang J, Xu RH, Chen YE, Mangelsdorf DJ. Expression profiling of nuclear receptors in human and mouse embryonic stem cells. Mol Endocrinol. 2009 May;23(5):724-33. Epub 2009 Feb 5.

Cheng H, **Kim BH**, Grishin NV. Discrimination between distant homologs and structural analogs: lessons from manually constructed, reliable data sets. J Mol Biol. 2008 Apr 4;377(4):1265-78. Epub 2008 Jan 5.

Pei J, **Kim BH**, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 2008 Apr;36(7):2295-300. Epub 2008 Feb 20.

Cheng H, **Kim BH**, Grishin NV. MALIDUP: a database of manually constructed structure alignments for duplicated domain pairs. Proteins. 2008 Mar;70(4):1162-6.

Cheng H, **Kim BH**, Grishin NV. MALISAM: a database of structurally analogous motifs in proteins. Nucleic Acids Res. 2008 Jan;36 (Database issue):D211-7. Epub 2007 Sep 12.

Qi Y, Sadreyev RI, Wang Y, **Kim BH**, Grishin NV. A comprehensive system for evaluation of remote sequence similarity detection. BMC Bioinformatics. 2007 Aug 28;8:314.

Sadreyev RI, Tang M, **Kim BH**, Grishin NV. COMPASS server for remote homology inference. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W653-8. Epub 2007 May 21.

Pei J, **Kim BH**, Tang M, Grishin NV. PROMALS web server for accurate multiple protein sequence alignments. Nucleic Acids Res. 2007 Jul;35(Web Server issue):W649-52. Epub 2007 Apr 22.

Kim BH, Sadreyev R, Grishin NV. COG4849 is a novel family of nucleotidyltransferases. J Mol Recognit. 2005 Sep-Oct;18(5):422-5.

List of Figures

FIGURE 1 OVERVIEW OF SELF CONSISTENCY GROUPING. 25

FIGURE 2 AVERAGE CLUSTER SIZE OF SCG CLUSTERING BASED ON SINGLE SCORES 39

FIGURE 3 COMPARISON OF SCG CLUSTERING TO SCOP FAMILY 41

FIGURE 4 COMPARISON OF SCG CLUSTERING TO SCOP SUPERFAMILY 42

FIGURE 5 F-MEASURES BETWEEN EACH CLUSTERING AND SCOP FAMILY 44

FIGURE 6 AVERAGE CLUSTER SIZES OF SCG CLUSTERS FROM DIFFERENT SIMILARITY MEASURES 45

FIGURE 7 COMPARISON OF SCG CLUSTERS FROM DIFFERENT SIMILARITY MEASURES TO SCOP FAMILY 47

FIGURE 8 COMPARISON OF SCG CLUSTERS FROM DIFFERENT SIMILARITY MEASURES TO SCOP SUPERFAMILY 48

FIGURE 9 AVERAGE CLUSTER SIZES ARE SHOWN FOR EACH ITERATIONS FROM ISCG CLUSTERING BASED ON THE OPTIMALLY COMBINED 5 SEQUENCE AND STRUCTURE SCORES 50

FIGURE 10 ISCG CLUSTERS COMPARED TO SCOP SUPERFAMILY 51

FIGURE 11 CLUSTER SIZE DISTRIBUTION MEASURED BY NUMBER OF CLUSTERS ABOVE 5, 10, 15, 20, AND 40 53

FIGURE 12. COMPARISON OF CLUSTERS FROM SINGLE LINKAGE CLUSTERING METHOD TO SCOP SUPERFAMILY CLUSTERING 56

FIGURE 13. MEDIAN OF C-VALUES FOR CLUSTERING AT EACH CUTOFF VALUE 58

FIGURE 14. MEDIAN OF C-VALUES FOR CLUSTERING AT EACH CUTOFF VALUE FOR CLUSTERS BIGGER THAN 5 59

FIGURE 15. MEDIAN OF C-VALUES FOR CLUSTERING AT EACH CUTOFF VALUE FOR CLUSTERS BIGGER THAN 20 60

FIGURE 16 CLASSIFICATION RESULT REPRESENTED AS NETWORK 63

FIGURE 17 NUMBER OF CLUSTERS FOR EACH ISCG CLUSTER SIZE 64

FIGURE 18 NUMBER OF CLUSTERS FOR EACH ISCG CLUSTER SIZE. 65

FIGURE 19 SUPERCLUSTERS: TOP 10 BIGGEST CLUSTERS 66

FIGURE 20 STRUCTURES OF SUPERFOLD PROTEINS SELECTED BY ORENGO AND THE COWORKERS 68

FIGURE 21 REPRESENTATIVE STRUCTURES OF SUPERCLUSTERS 69

FIGURE 22 NUMBER OF OB FOLD STRUCTURES DEPOSITED INTO PDB AT EACH YEAR. 69

FIGURE 23 ANALOGOUS FERREDOXIN FOLD PROTEINS 71

FIGURE 24 DIFFERENT BETA-SHEET TOPOLOGY IN NUCLEOTIDE KINASE AND G PROTEIN 73

FIGURE 25 SIMILARITY BETWEEN NUCLEOTIDE KINASE AND NITROGENASE. 74

FIGURE 26 SIMILARITY BETWEEN NITROGENASE AND G PROTEIN 74

FIGURE 27 MANUAL CHECKING RESULT OF 83 CLUSTERS THAT CONTAIN DIFFERENT SCOP SUPERFAMILIES 75

FIGURE 28 REPRESENTATIVES OF FLAVODOXIN-LIKE FOLD PROTEIN STRUCTURES 89

FIGURE 29 HIERARCHICAL TREE FOR CLUSTER 43 94

FIGURE 30 BETA-PROPELLER FOLD PROTEIN STRUCTURES 97

FIGURE 31 DIFFERENT TYPES OF BLADES IN BETA-PROPELLER 98

FIGURE 32 REPRESENTATIVE STRUCTURES IN CLUSTER 4431 100

FIGURE 33 NETWORK VIEW OF CLUSTER 10 109

FIGURE 34 REPRESENTATIVE STRUCTURES IN CLUSTER 2859 112

FIGURE 35 REPRESENTATIVE STRUCTURES FROM CLUSTER 964 115

FIGURE 36 REPRESENTATIVE STRUCTURES FROM CLUSTER 1729 116

FIGURE 37 THE REPRESENTATIVE STRUCTURES IN CLUSTER 363 117

List of Tables

TABLE 1. LIST OF PROTEINS IN 10 BIGGEST CLUSTERS. 67

TABLE 2 POTENTIALLY HOMOLOGOUS CLUSTERS CONTAINING DIFFERENT SCOP SUPERFAMILIES. 83

TABLE 3 POTENTIAL HOMOLOG CLUSTERS CONTAINING DIFFERENT FOLDS. THE LEGENDS ARE SAME AS TABLE 2. 87

TABLE 4 POTENTIAL HOMOLOG CLUSTERS CONTAINING DIFFERENT CLASS. THE LEGENDS ARE SAME AS TABLE 2..... 88

TABLE 5. SUMMARY OF SUPERFAMILIES IN CLUSTER 15. 95

TABLE 6 CLUSTERS CONTAINING DIFFERENT SUPERFAMILIES WITHIN SAME FOLD IN "FOLD SIMILAR; HOMOLOGY UNCLEAR" CATEGORY..... 107

TABLE 7 SUMMARY FOR CLUSTERS IN "PARTIAL SIMILARITY, UNLIKELY HOMOLOGY" CATEGORY. 113

TABLE 8 SUMMARY OF CLUSTER IN "NO SIMILARITY, WRONG CLUSTER" CATEGORY. 117

CHAPTER 1

Introduction

"Nothing in Biology Makes Sense Except in the Light of Evolution."

(Dobzhansky, 1964)

1.1 Homology Inference

As Dobzhansky wrote in his article, evolutionary theory made a profound impact on biological sciences. Besides all the great philosophical and theoretical innovations that were brought by evolutionary theory, almost all biologists make predictions and hypotheses based on homology arguments in their everyday practices. The inference of homology is, however, sometimes difficult because of the distant relationships (Kinch & Grishin, 2002). Or sometimes the homology inference is proved to be erroneous because of the complications in their evolutionary history such as multi-domain problem (Gilks et al, 2002).

To infer homologous relationships, many homology search methods have been developed. The first milestone was the development of famous sequence similarity search method called basic local alignment search technique, BLAST (Altschul et al, 1990). BLAST finds homologous proteins (hit's) in the protein sequence database for a given protein (query) using pairwise sequence comparison. After BLAST, position specific

iteration BLAST (PSI-BLAST) was developed (Altschul et al, 1997). PSI-BLAST finds homologous sequences (hit's) in the database for a protein (query) using not only the sequence information of hit and query proteins but also homologous proteins of query proteins. This position specific matrix used in PSI-BLAST was generally called profile. After PSI-BLAST, many researches were done how to effectively compare profiles, profile-profile alignment (Rychlewski et al, 2000; Sadreyev & Grishin, 2003; Sunyaev et al, 1999). COMPASS (Sadreyev & Grishin, 2003) and HHsearch (Soding et al, 2005) are the state-of-the-art methods among those profile-profile comparison methods. After development of those profile-profile alignment programs, protein structural information as well as sequences or profiles were used to find remote homology such as HorA (Cheng et al, 2008). This HorA method extended the limit of homology detection since the structures are generally conserved longer than sequences.

One of key features of HorA method is that it uses previously known homologous protein sets (called training set) to derive the principles. This kind of research, extracting information from training set, is referred as supervised learning. The name supervised learning came from the fact that the training set is consist of examples to supervise the "learning" process of extracting the rule.

The classification of proteins developed in this study can be viewed as a methodology to detect remotely homologous proteins without pre-defined training set. Compared to the supervised learning procedure like HorA, the classification of protein

has intrinsic strength to find previously unknown information because the supervised learning procedure essentially mimics pre-defined examples.

1.2 Protein Classification

Classification is one of most fundamental logical activities of human being. Also, modern biology is essentially built on the famous classification of Linnaeus classification of species. Since classification is helpful in reducing the information and organize the relationships between objects, proteins naturally became subject of classification just like species were the subject of classification in 1700 to Linnaeus. A large body of work is already done in classification of proteins. Here, two orthogonal perspectives were chosen to briefly overview previous work in protein classification. One perspective is the major source of information for classification, i.e. sequence information or structural information. The other perspective is the main body of classifier, i.e. human experts (manual) or computers (automatic).

There are quite many protein classification schemes utilizing sequence information or group of homologous sequences, such as Pfam (Sonnhammer et al, 1997), CDD (Marchler-Bauer et al, 2007), and COG (Tatusov et al, 1997). There are three major protein structure classification database; SCOP (Murzin et al, 1995), CATH (Orengo et al, 1997) and Dali Domain Dictionary (Dietmann et al, 2001).

Protein classifications can also be grouped into manual or automatic schemes. SCOP is one of most popular manual method and Dali Domain Dictionary is one of famous resources in automatic schemes. Although manual classification schemes like

SCOP or semi-manual classification, CATH uses both sequence and structural information, both classifications use sequence and structure in different hierarchical levels not in combination of the two at the same time. Here, the protein classification developed in this study uses combined information of sequence and structure for the first time.

CHAPTER 2

Methods

2.1 Dataset

2.1.1 Selection of protein domains

Since proteins are composed of semi-independent folding, functional, or evolutionary units in general, we used domains defined in SCOP version 1.71 (Murzin et al, 1995) as our unit of proteins. The non-redundant protein set was prepared. A representative was selected from a group of proteins (more specifically domains) with more than 40% sequence identity. This selection procedure essentially gives a representative domain set with the less than 40% sequence identity among the members of this representative set. This selection procedure was developed by ASTRAL compendium (Brenner et al, 2000). From the downloaded list of non-redundant protein domains, we selected proteins from all-alpha, all-beta, alpha+beta, and alpha/beta classes defined in SCOP database. The other classes in SCOP database are artificially categorized, so we excluded them from our dataset, i.e. small, membrane, and multi-domain classes. The total number of domains in our final dataset is 7058.

2.1.2 Preparation of protein domain structures and sequences

Domain structures were prepared by ASTRAL compendium (Brenner et al, 2000). Domain sequences were prepared based on residues appeared in ATOM records in PDB

files. This procedure was developed to ensure simple 1:1 relationship between the residues in structure and sequence.

2.2 Generation of alignments and scores

5 different programs were used to align protein sequences or structures and the scores were reported. Each program uses its own scoring scheme to find the optimum alignment. In general, sequence alignment programs find real optimum since the search space is relatively small. In contrast, structural alignment programs do not guarantee to find the optimal alignment.

2.2.1 Generation of sequence alignments and scores

More exact and specific term for sequence alignment in this research is profile-profile alignment. Sequence alignment in general means the alignment of two sequences based on the amino acid from the two sequences subject to the alignment. Profile-profile alignment is an alignment of two sequences using not only the two subject sequences but also their homologs found by database search. Since the conservation pattern in many homologous sequences helps greatly to find more remotely related proteins, profile-profile alignment methods were used. Here, the title is given as “sequence alignments and scores” to emphasize in meaning that the profile-profile alignments use sequence information only and do not use structural information like the next section (2.2.2).

2.2.1.1 Compass

The raw score from COMPASS (Sadreyev et al, 2007) was used as a sequence score. The profiles for COMPASS alignments were generated by the script (buildali.pl, kindly provided by Dr. Soding in his HHsearch (Soding et al, 2005) package version 1.5). The parameters for the alignment building script were default and database searching was done on NRE90 (NRE90: NR+ENV with sequence identity 90% representatives, NR: Non redundant database in National Center for Biological Information, ENV: sequences gathered from environmental sampling, i.e. deep sea water) and NRE70 (NR+ENV with sequence identity 70% representatives) as suggested by the author. Notably, buildali.pl script generated multiple sequence alignments (MSA) equivalent to the PSI-BLAST 8th iteration profile. Then the columns in MSA having gaps in the query sequence were removed. The MSA was converted to numerical profile by COMPASS program.

2.2.1.2 HHsearch

HHsearch (Soding et al, 2005) probability was parsed from HHM alignments. HHM files that are equivalent to the profile in COMPASS were generated by the script provided by Dr. Soding in the HHsearch package. MSA files were generated the exactly same way as in MSA for COMPASS profile. Then the MSA files were converted in HHM using hhmake program in HHsearch package. The parameters were all default.

2.2.2 Generation of structural alignments and scores

2.2.2.1 DaliLite

DaliLite (Holm & Sander, 1993) raw score was used as a structural similarity score. PDB files were downloaded from ASTRAL compendium (Brenner et al, 2000). Seleno-methionines in PDB file were changed into methionine. Other kinds of non-standard amino acids were represented as X or Unknown. For multiple models in PDB files based on NMR, the first model was selected as a representative structure. In case of multiple alternative structural alignments, highest Z score alignments were selected as suggested by the author.

2.2.2.2 FAST & TM-align

PDB files were prepared similarly as in DaliLite program. One difference is that FAST (Zhu & Weng, 2005) cannot use non-standard amino acids in alignment process. All the non-standard amino acids where replace with Alanine for the sake of alignment generation, since the residue type does not affect the alignment process at all. FAST raw scores were parsed from the result.

For TM-align, PDB files were prepared the exactly same way as in FAST program. TMalign (Zhang & Skolnick, 2005) raw scores were parsed from the alignment result.

2.2.3 Scaling and normalization

2.2.3.1 Scaling with simple random model

$$S = \frac{S_{12} - S_{random}}{S_{self} - S_{random}}$$

S_{12}: Score between two proteins (or domains) 1 and 2

S_{self}: Average scores of two scores to proteins to themselves

$$S_{self} = \frac{S_{11} - S_{22}}{2}$$

S_{random} : Random score between two proteins

S_{random} was defined in two different ways.

$S_{random\ 1}$: The given alignment of proteins 1 and 2 was changed by reversing protein 1 in the aligned region. Then the score were calculated based on the modified alignment. This random score was good in keeping the composition (in sequence) or the connectivity (in structure) of amino acids in the alignment.

$S_{random\ 2}$: The given alignment of proteins 1 and 2 was changed by shifting protein 1 fixed number of residues in the aligned region. The number of shift is decided by number of aligned residues in protein 1 divided by 11. The shift of aligned sequence of 1 will be repeated 10 times with scores calculated for each shift. The random score was calculated as the median of 10 scores from each shift. Since the problem in $S_{random\ 1}$ is that it uses one random score, if the random score estimation can be easily biased. This $S_{random\ 2}$ uses median value of many different shift random trials, it has less bias.

2.2.3.2 Normalization with empirical score distribution

Another way of normalization is using Z-score. In this study, we used slightly modified version of Z-score compared to other studies.

$$Z = \frac{S_{12} - Mean_{12}}{Sigma_{12}}$$

S_{12}: Score between proteins 1 and 2

$Mean_{12}$: mean of scores between protein1 against all proteins in the dataset and between protein2 against all proteins. “12” means that it is from both proteins 1 and 2.

$Sigma_{12}$: Standard deviation of scores between protein1 against all proteins in the dataset and between protein2 against all proteins.

This modification is based on the observation that the score distribution does not only depend on the query protein (protein1) but also depends on hit protein (protein2).

2.3 Self consistency grouping

2.3.1 Self consistency grouping

Self consistency grouping (SCG) generates clusters when all the members of the cluster are best n (number of members in the cluster) hits. This procedure ensures that all the members of a cluster are closer to each other than the members outside of the group. SCG procedure ends when the next cluster is trivial cluster containing everything in representative set.

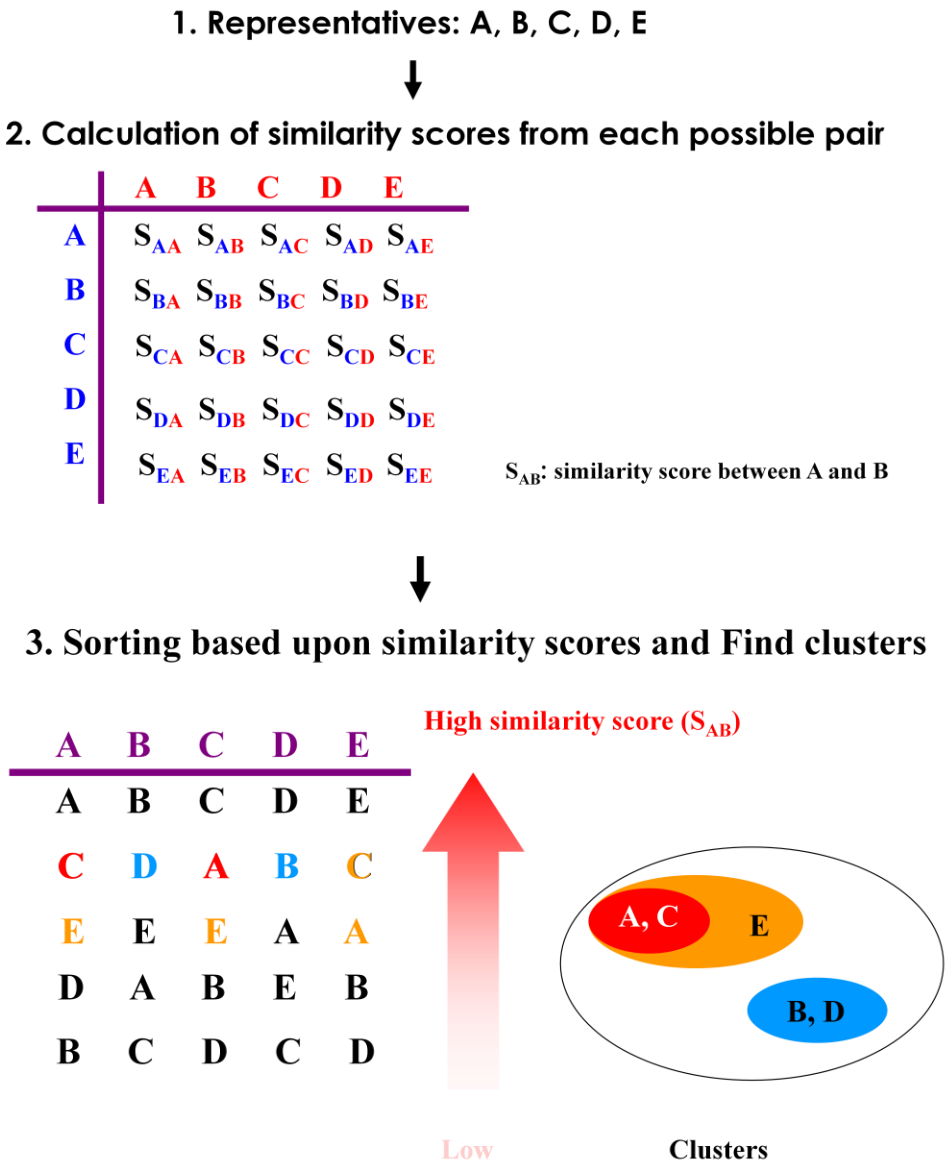


Figure 1 Overview of Self Consistency Grouping. Starting From selected representatives (step1) then similarities between all proteins are calculated (step2, e.g. for 5 selected proteins the similarity matrix will be 5x5). Next, the similarities are converted into ranking (step3) and the ranks are used to build clusters. In step3, all other proteins below the purple line are ordered according to the similarity to the proteins above purple line in each column. Then the ranking list is used to build clusters by grouping according to the consistency rule: all proteins within the group are closer than others outside the group are. Protein A and C are clustered first because they are consistently more closer to each other than others as shown in the ranking list; e.g. columns A and C the most closest proteins are itself and the second most similar proteins are C and A (red colored proteins in the ranking list). Similarity, B and D (blue colored proteins) are grouped. Notably, E formed cluster after A and C formed cluster because A and C have E as their 3rd closest proteins in their ranking list (columns A and C).

2.3.1.1 Pseudo codes for SCG algorithm

Problem:

Find groups of proteins. For each group, the group should be the biggest (encompass as many as proteins as possible). All proteins in one group are more similar to the member of the same group than other proteins outside the group. Also, the biggest group cannot be a trivial solution of one big group containing every protein in the dataset.

Inputs:

1. *Positive integer **N** for total number of proteins in the dataset*
2. *Rank matrix N by N **rank_mat** for protein 1…N*

   An element of **rank_mat[n,m]** is the rank of similarity score between protein n and m for query protein n.
3. *Sorted hit matrix N by N **sorted_mat** for proteins 1…N*

   If an element of **sorted_mat[n,m] = k** then the m^{th} similar protein to a query protein n is protein k. Here, all indices n, m, and k are between 1 and N.

Outputs:

An array of arrays represents clustering, clusters. **clusters[i]** is an array of indices of proteins in the cluster i. Cluster index, i is given by the order of cluster formation, so this cluster index i does not have any meaning except the uniqueness of cluster.

```
int clusters[][] SCG ( int N,
        const int rank_mat[][],
        const int sorted_mat[][] )
{
        terminal_nodes = initialize ( terminal_node for each proteins [1…N] )
        #check for grouping possibility for every terminal node (individual proteins)
        for cn = 2 … N :
                for t in terminal_nodes :
                        check_for_consistency( t, cn, rank_mat, sorted_mat )

        #Convert terminal_note_list into cluster_arrays;
        root_set = find all non-redundant root nodes from terminal_nodes
        n=0
        clusters = []
        for root in root_set :
                clusters[n] = Root.get_terminals()
                n++;
}
```

This function, check_for_consistency, checks if all the proteins of the top cn in the hit list of protein t.

```
Void check_for_consistency( Node t,
        int cn,
        rank_mat,
        sorted_mat )
{
        current_cluster_size = t.get_size()
        new_terminals = sorted_mat[ t ][ current_cluster_size : cn ] #This selects all terminal nodes between current_cluster_size and cn.
        root_set = roots from new_terminals and current root
        for root1 in root_set :
                for root2 in root_set and not root1:
                        if is_consistent( root1, root2, rank_mat ) then :
                                continue to next loop
                        else :
                                return
        combine_roots( root_set )
}
```

This function checks for all terminal nodes under the two roots are ranks within given integer number cn.

```
Boolean is_consistent( Node root1,
```

```
        Node root2,
        int cn,
        int rank_mat[][] )
{
        for terminal1 in root1.get_terminal_set() :
                for terminal2 in root2.get_terminal_set() :
                        if Rank_mat[Terminal1][terminal2] <= cn :
                                continue
                        else :
                                Return false
}
```

The function combine_roots makes a new root node and all current root nodes make as child_node of a new root node.

2.3.2 Iterative self consistency grouping

iSCG is an extension of SCG method by building cluster of clusters. The similarities between clusters are defined by maximum score between members of two clusters determined in the previous iteration. iSCG procedure ends when the current iteration does not merge clusters from the previous iterations.

2.4 Comparing a clustering to a reference classification

2.4.1 Reference clustering

Reference clusters are used for monitoring quality of test clusters. Since SCOP database (Murzin et al, 1995) is one of the best resources and has most comprehensive definition for evolutionarily related proteins (SCOP hierarchy of superfamily), SCOP is suitable for serving as reference clustering. Sometimes SCOP families or folds can serve as reference. SCOP family relationship is much tighter than superfamily and SCOP fold is much looser than SCOP superfamily.

2.4.2 Sensitivity

Sensitivity (Sam et al, 2006) is a measure of the proportion of a test clustering covering a reference clustering. Sensitivity is also called as recall in some literatures or other fields. Since sensitivity is originally defined for binary classification problem, the test and reference clusters are reduced to binary relationships.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Here,

TP = number of pairs of objects in the same cluster for both test and reference clusters

FN = number of pairs of objects in different clusters for test clustering but in the same reference cluster.

Again, the terms, TP and FN, came from the binary classification tradition. TP means True Positive and FN means False Negative. Here True and False mean whether the test result (here is a pair of objects) is right or wrong (or same/different) compared to reference. Positive or Negative mean that the pairs of objects are in the same cluster or not.

2.4.3 Specificity

Specificity (Sam et al, 2006) is a measure of the bad or wrong proportion in a test clustering compared to the reference clustering.

$$\text{Specificity} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

Here,

TP = number of pairs of objects in the same cluster for both test and reference clusters

FP = number of pairs of objects in the same test cluster but not in the same reference cluster.

Specificity has several variant formulas in different disciplines. The form used in this research is common among biomedical researchers and it is also known as positive predictive value in statistics.

In statistics, specificity is defined as follows.

$$\text{Specificity} = \frac{\text{TP}}{\text{TP}+\text{TN}}$$

TN: the number of links between proteins exists in different clusters in both test clustering and reference clustering. (FP is defined same way as before)

Even though the definition of specificity in statistics is very attractive measure to compare clustering in general, it is not a good measure to compare clustering of proteins. This is simply because the protein clustering is generally very small or sparse. For sparse clustering, TN becomes too high, and specificity becomes indistinguishable between good and bad clustering. Compared to specificity in statistics, specificity in biomedical sciences (or PPV) similarly measures number of wrong links, but do not have the problem involving big TN, since PPV does not use TN value in the equation. Therefore, in this study PPV is used instead of specificity. This also fits the tradition in biomedical sciences.

2.4.4 F-measure

F-measure (Van Rijsbergen, 1979) is a single value representation that combines sensitivity and specificity. It is defined as Harmonic mean of sensitivity and specificity.

$$\mathrm{F} = \frac{2 \times \mathrm{Sensitivity} \times \mathrm{Specificity}}{\mathrm{Sensitivity} + \mathrm{Specificity}}$$

This F-measure means how similar is the test clustering to the reference clustering. If the test and reference is same F-measure will be 1 and 0 means they are mostly dissimilar. Note that original form of F-measure use the term precision (same as specificity) and recall (same as sensitivity) commonly used in machine learning field (Vens et al, 2008).

2.5 Inference of phylogenetic tree

Sequence based phylogenetic trees for proteins in same clusters were inferred by the following steps.

1. MSA generation: Promals3D (Pei et al, 2008) generated MSA by combining pairwise structural alignments from DaliLite (Holm & Sander, 1993).
2. Distance inference: Protdist in Phylip package (Felsenstein, 1989) inferred distances between proteins
3. Tree building: Weighbor (Bruno et al, 2000) built trees based on the distance matrix from step2.

Structure based phylogenetic trees were inferred slightly modified steps.

1. Distance inference: DaliLite Z-score were converted by simple $\frac{Z_{self}}{Z} - 1$ transformation.
2. Tree building: Weighbor built trees based on the distance matrix from step1.

Trees based on the combined score were inferred by the same procedure as structure based trees but based on the combined score instead of DaliLite Z-score.

CHAPTER 3

Development of Protein Classification

3.1 Preliminaries

3.1.1 Dataset

Selecting proteins to be classified is the first step toward protein classification. Since it is our goal to provide comprehensive classification, to cover the whole protein universe, preferably all proteins found so far should be used. Practically this brute force approach is hard to achieve because of current computational limit of CPU time and memory. Moreover, if there are many redundant proteins in the dataset, they might hinder more accurate representation of protein fold space revealed by the classification process (Park et al, 2000). So, more reasonable choice is selecting non-redundant representatives to the dataset. 40% representatives, all proteins in the dataset share less than 40% sequence identity, were chosen because proteins generally shares structure and function above 40% sequence identity (Heger & Holm, 2003).

Another consideration for the dataset is whether to use whole length proteins or domains. Domains were chosen as dataset. By definition, evolutionary classification needs the objects in the dataset to have one ancestor (Reeves et al, 2006). Domain is, in evolutionary context, defined as an evolutionary unit share its evolutionary history in a protein. This definition of domain suits well with our purpose of evolutionary

classification. However the domain definition, i.e. how to divide a whole length protein defined in one open reading frame (ORF) in genomes, is very difficult to define (Marsden et al, 2002). In fact the correct domain definition depends on the correct delineation of evolutionary history of the protein, which is the subject of this study. To avoid this circular problem, this study used SCOP domains (Andreeva et al, 2008). SCOP domain is currently one of most widely accepted evolutionary domain definition and it was determined by experts.

Among SCOP40 representatives (SCOP domains shares less than 40% sequence identities in between), proteins do not belong regular globular protein classes, (all alpha, all beta, alpha+beta, and alpha/beta classes in SCOP) were excluded from the dataset. Those excluded proteins are mostly, membrane proteins, very small proteins, artificially generated proteins. The total number of proteins in SCOP40 ver. 1.71 without non regular globular proteins is 7085 proteins.

3.1.2 Scoring schemes

It is well established that sequence and structural changes are correlated (Anna R. Panchenko, 2005), but there is also difference between sequence and structures. Especially for higher sequence identity (more than 40%), structures are largely same. Also the sequence similarity drops quite rapidly compared to the structural scores. I.e. the structural signal lasts much longer than sequence signal. So the sequence similarity

and structural similarity can be complementary to each other to have better similarity measure.

Three structural similarity scores and two sequence similarity scores were used to evaluate similarities between domains. Three structural similarity scores came from DaliLite (Holm & Sander, 1993), FAST (Zhu & Weng, 2005), and TMalign (Zhang & Skolnick, 2005). Two sequence similarity scores were calculated by COMPASS (Sadreyev et al, 2003) and HHsearch (Soding et al, 2005). Those programs are state-of-the-art similarity comparison programs in structure and sequence. We made similarity score matrix between all proteins in dataset for 5 different similarity scores.

3.2 SCG clustering based on single measures

SCG algorithm is a clustering algorithm which is based on the classical concept of clustering that all members in a cluster need to be more similar (or closer in distance) than any others outside the cluster (Everitt, 1974). The name, SCG (self consistency grouping), came from this strict consistency requirement that all members in a cluster should be consistently more similar than non-members. SCG algorithm is built on two assumptions; (i) there are natural groups in the dataset and (ii) the similarity measure gives relatively high similarity scores between members in the same natural groups compared to similarity score between members in different natural groups. If there is an ideal similarity measure without error then SCG algorithm will correctly cluster most of natural groups, because this ideal similarity measure will give accurately high similarity scores between the members of the same natural groups. And the ideal similarity

measure will give low similarity scores for members in different natural groups. In contrast, if the similarity measure is not ideal and there is an erroneous value in the similarity matrix (i.e. high similarity score between non-natural groups) then the strict consistency requirement will remove the member with the erroneous similarity score from the cluster determined by SCG. Also, the consistency requirement will prevent to form a bad cluster based on erroneous scores, because it is very hard to have consistently high similarity scores between non-related object by random error. In summary, SCG is a clustering algorithm that is highly specific and will not form many bad clusters because of its strict consistency requirement. However, it will also be sensitive to errors. Errors in similarity measure likely make the SCG clusters smaller.

In the context of other clustering algorithms, SCG is a kind of agglomerative hierarchical clustering algorithm with certain differences compared to popular algorithms in this category such as single linkage clustering, average linkage clustering and complete linkage clustering. Among the popular methods, SCG is most close to complete linkage clustering. SCG can be seen as an algorithm that clusters similar cluster to complete linkage algorithm except that the cutoff is dynamically determined for each cluster. SCG method determines the stopping point of clustering where there is inconsistency that is not solved by increasing cluster size, hence SCG finds biggest consistent group of proteins in the ranks. In general, agglomerative hierarchical clustering methods require cutoff value to stop clustering. Without the cutoff value, those hierarchical methods cluster everything in the dataset.

SCG algorithm is also can be seen as a generalized algorithm of Clusters of Orthologous Groups (COG) database (Tatusov et al, 1997). In COG database, clusters of orthologs are defined by more than three proteins mutually most similar proteins between the genomes. COG and SCG are similar because both algorithms use information of mutually most similar proteins. SCG is, however, more general than COG algorithm because COG algorithm only use a mutually most similar pairs as a unit to define the cluster and SCG use most similar N (not pre-defined, bigger than 2 and smaller than the size of whole dataset) proteins.

3.2.1 Cluster size analysis shows that SCG clusters are very small on average.

Clusters were built by SCG based on similarity scores from 3 structural alignment methods and 2 profile comparison methods. (See details in section 2.2.) The raw similarity scores from each program were transformed into modified Z-scores. This modified Z-score are same as common Z-scores except that the distribution is combination of lump sum of two scoring distribution of two proteins in comparison. (See details in section 2.2.3.2.) Then, SCG algorithm built clusters.

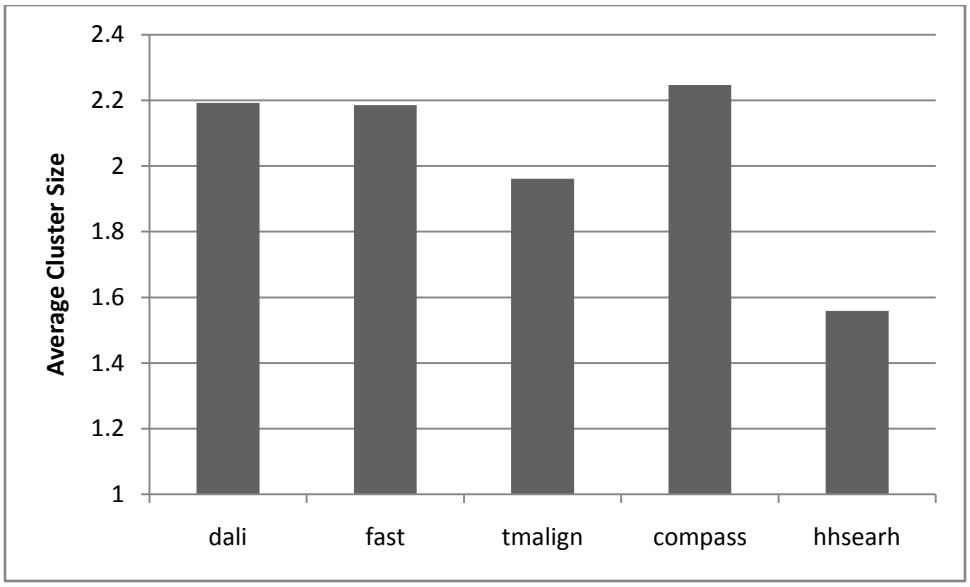


Figure 2 Average cluster size of SCG clustering based on single scores. dali, fast, and tmalign are clusterings based on structural similarity scores from DaliLite, FAST, and TMalign programs respectively. compass and hhsearch are clusterings based on sequence similarity measured by profile comparison programs COMPASS and HHsearch, respectively. dali and compass shows relatively bigger cluster size and hhsearch shows smallest.

Figure 2 shows that the clusters based on different similarity measures are all small. The average cluster sizes are about 2 and this is very small size. Average cluster size of SCOP (Murzin et al, 1995) families is about 3. So by the measuring cluster size only, generally SCG clusters are smaller than SCOP families.

Figure 2 also shows that relatively smaller clusters for clusters based on TMalign (Zhang & Skolnick, 2005) and HHsearch (Soding et al, 2005) scores. First, TMalign is generally more erroneous in measuring structural similarity than DaliLite (Holm & Sander, 1993) or FAST (Zhu & Weng, 2005). There are possibly two reasons for that. One reason is the aligning algorithm of TMalign. The algorithm of TMalign is started from the three initial alignments and iteratively finds the optimal alignment. However, if the

initial guesses are too far from the optimal alignment, TMalign probably miss the global optimum and TMalign is probably trapped into a local optimum alignment. The other is because of the TMalign scoring function. The scoring function of TMalign is only positive. So if the alignment is longer than the score will be higher. This makes TMalign to have longer alignments than other structural alignment programs. Second, HHsearch has smaller cluster size than other methods possibly because of the normalization procedure. Since the most reliable score reported by HHsearch is its probability value (the probability of being homologs from the comparison of Hidden Markov Models), the probability was converted into Z-score. However, the average cluster size is reduced after the Z-score normalization compared to the raw HHsearch probability. This problem remains to be fixed.

3.2.2 The cluster membership projected on SCOP database shows that cluster contains proteins related by family or subfamily groups.

The contents of clusters, what kind of proteins are in each cluster, can be more accurately established by comparing pre-existing classification (Fowlkes & Mallows, 1983). SCOP database provides very good framework for the comparison. Sensitivity and specificity is measured by comparing SCG clusters and SCOP family and superfamily clustering.

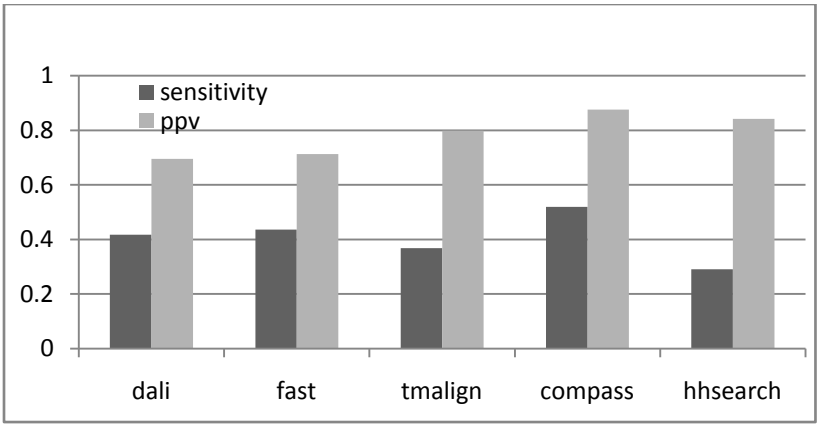


Figure 3 Comparison of SCG clustering to SCOP family. All clustering labels are same as Figure 2 Sensitivity and specificity (or ppv) values are calculated as described in methods. Both sensitivity and specificity have ranges from 0 to 1. Higher sensitivity indicates a bigger proportion of domains in the same SCOP families are also clustered in the same SCG clusters. Higher specificity indicates a smaller proportion in SCG clusters are from different SCOP families. When both sensitivity and specificity are 1, SCG clustering is exactly same content as SCOP family classification; whereas both are 0, SCG clustering is totally different from SCOP family.

Figure 3 shows the comparison result of SCG clusters based on each alignment methods to SCOP family classification. Notably, all the methods show less than 0.6 sensitivities. This supports the same conclusion derived by average cluster size analysis done in previous section. In general SCG clusters are smaller than SCOP family clustering and the clusters generally contain sub-family groups. COMPASS based SCG clustering shows higher sensitivity and specificity compared to other methods. It is not surprising that COMPASS show better performance than other structural alignment program because sequence method is better to distinguish the differences between proteins in family or subfamily level classification (Russell et al, 1997).

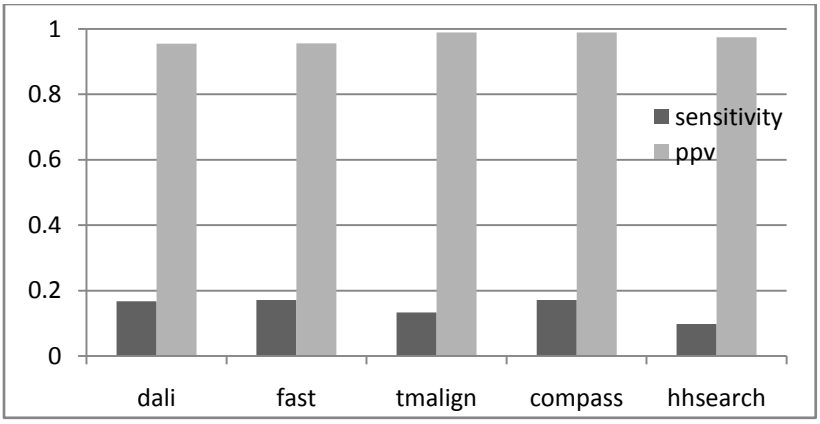


Figure 4 Comparison of SCG clustering to SCOP superfamily. All labels are same as Figure 3. This figure shows notably smaller sensitivity compared to Figure 3, due to much bigger definition of SCOP superfamilies.

Figure 4 shows SCG clusters based on different information compared to SCOP superfamily clustering. Since SCOP superfamilies contains more remote relationships than SCOP family (Figure 3), sensitivities are low and specificities are much high. Notably the relative performance of SCG clusters remain the same. Clustering based on COMPASS was most similar to SCOP superfamily, even though the difference is much smaller than that to SCOP family.

Even though it is very necessary to check our clustering or classification to previously established knowledge, we should not biased toward only recapturing reference database or SCOP database in our case. It is very important that the classification should be justified by itself, not justified by comparison with SCOP database.

3.3 Unbiased combination of scores using SCG algorithm

The sensitivity of SCG method to errors (mentioned in section 3.2), however, allowed us to combine different scores without referring (or biasing towards) any previous knowledge. SCG method can be a means to find a better or more accurate combined scoring scheme than single scores by finding optimum weights to maximize the average cluster size. The optimally combined score is probably more accurate than other single measures in reflecting groups exist in the dataset. It is also known that homologous proteins exist as groups (families or superfamilies). Therefore, the combined score is probably more accurate in reflecting homologous relationships between proteins.

3.3.1 Average cluster size can be a quality measure for a scoring scheme in SCG clustering.

As we discussed already, average cluster size correlated to the quality of similarity measure. This correlation is shown clearly in the following Figure 5.

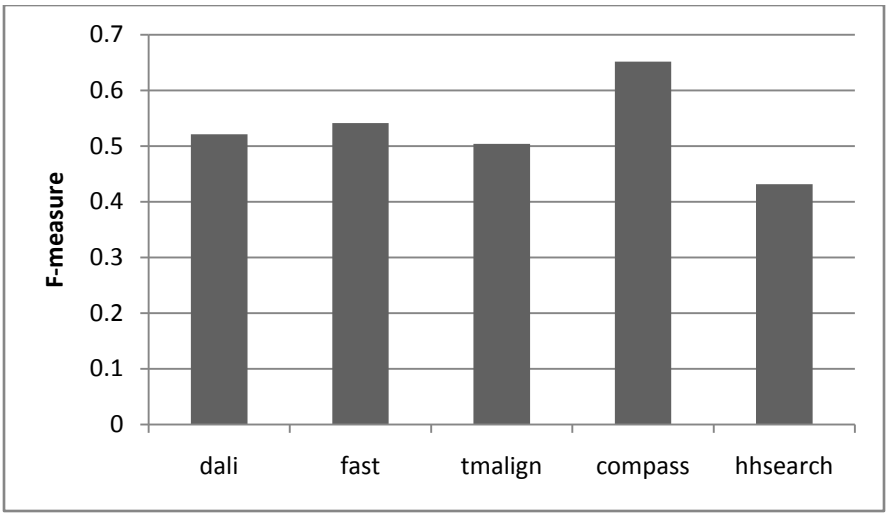


Figure 5 F-measures between each clustering and SCOP family. Same clustering labels are used as previous figures. F-measure is a combined measure of sensitivity and specificity used in previous figures. This value ranges from 0, indicating big difference between SCG clusters and SCOP family, to 1, indicating exact similarity between SCG and SCOP family. According to F-measure, SCG clustering based on profile comparison method COMPASS (labeled as compass) is most similar to SCOP family.

Since F-measure is a harmonic mean of sensitivity and specificity (see section 2.4.4), it summarizes the sensitivity and specificity into one easy interpretable value. Figure 5 shows that COMPASS is most similar to SCOP family and HHsearch is least similar. And structural measures are in between. This pattern of similarity (or quality) measured by SCOP family is well correlated to average cluster sizes shown in Figure 2.

3.3.2 Combining sequence and structural information makes cluster size bigger.

The combined score is a linear combination of previously mentioned 5 different scores from sequence and structural similarities. Scores from each program were converted by modified Z-score transformation. (See section 2.2.3.2.) The weights of the

linear combination of scores were determined by maximizing the average cluster size in clustering built by self consistency grouping (SCG) algorithm. To overcome technical difficulties in finding maximum average cluster size, we used genetic algorithm (Kikuchi et al, 2003). Genetic algorithm (GA) is a common stochastic method to find optimum value in multidimensional search space. GA is inspired by genetic inheritance in biological systems hence the name.

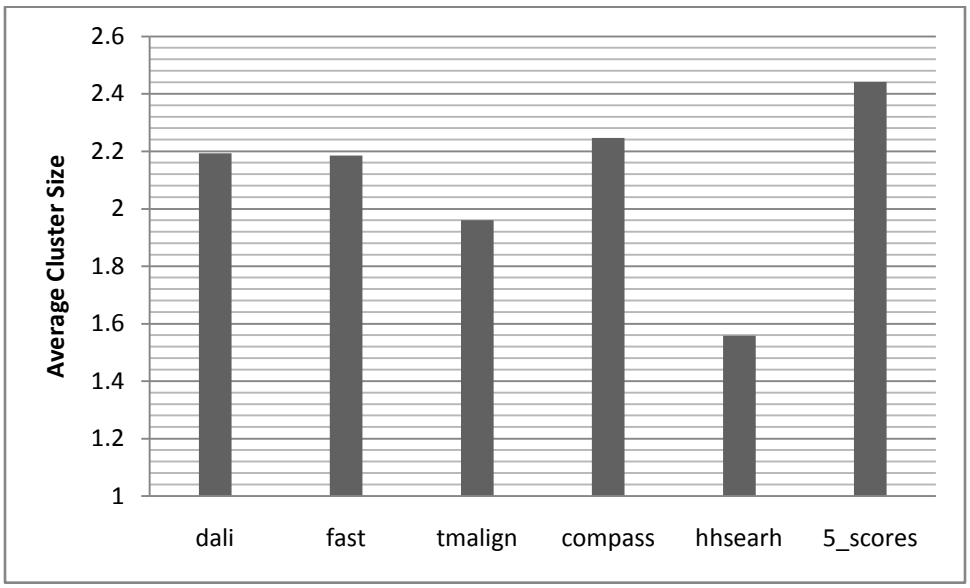


Figure 6 Average cluster sizes of SCG clusters from different similarity measures. Labels dali, fast, tmalign, compass, and hhsearch are same as previous figures. 5_scores denotes average cluster size of SCG clustering based on the optimally combined similarity measure using 5 different structural and sequence similarity scores (DaliLite, FAST, TMalign, COMPASS and HHsearch). All average cluster sizes are the same as used in Figure 2 except for the newly added 5_scores. 5_scores shows biggest average cluster size.

Figure 6 shows the improvement of optimally combined similarity measure compared to single measures. [Is it possible to show statistical measure that it is improved?] This relative increment is about 10% from the biggest average cluster size

from single similarity measures. Notably, the average cluster size is still much smaller than that of SCOP superfamily (~5) which is reasonable estimation of average cluster size of homologous proteins.

There are possible reasons for small increment. First, although the increment in terms of average cluster size is not big, the actual number of new evolutionary links defined by the combined score is not small. Second, the combination is linear. It is probable that the non-linear combination of scores helps to find better scoring function for evolutionary distance. Since it is not trivial to have natural form of non-linear combination, the non-linear combination of scores is one of the future directions of our research.

3.3.3 Combining sequence and structural information do not make cluster quality worse.

The quality of SCG clusters based on the combined score was measured by comparing to SCOP family and superfamily, since the average cluster size alone does not guarantee that the combined score is better similarity measure than the single measures. The comparison is done by measuring sensitivity and specificity as previous comparisons.

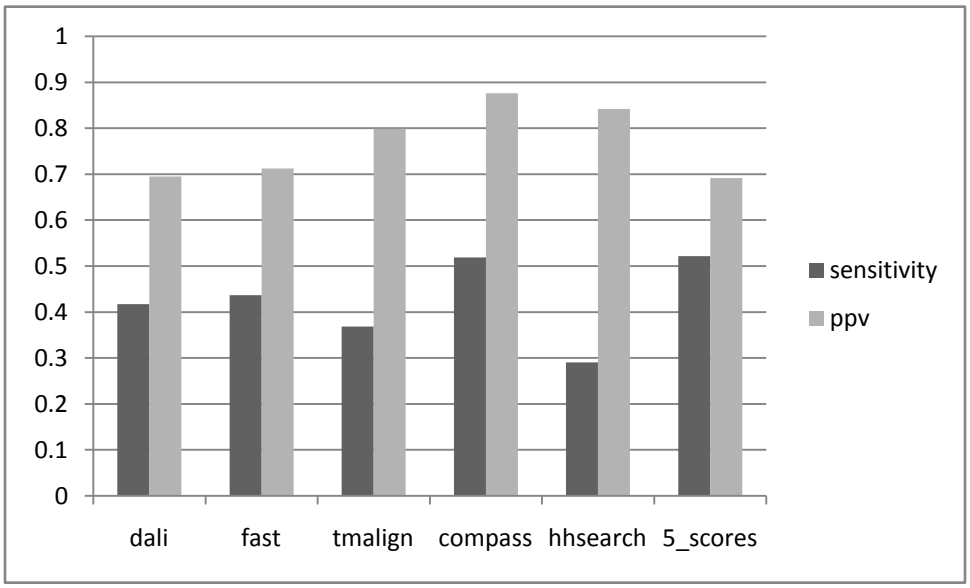


Figure 7 Comparison of SCG clusters from different similarity measures to SCOP family. All labels are same as Figure 6. All sensitivity and specificity values are from Figure 3 except the newly added 5_scores. Comparing to SCOP family, SCG clustering based on compass shows highest specificity. Notably sensitivity of 5_scores, the combined sequence and structural score, is relatively high.

Figure 7 shows that the comparison of SCG clusters to SCOP family classification. This figure shows that SCG clustering based on the combined score is similar to COMPASS in sensitivity yet the specificity is lower than COMPASS. Since SCG clustering based on similarity measure from COMPASS is better than the SCG clustering based on the combined score, this apparently unexpected result needs to be explained.

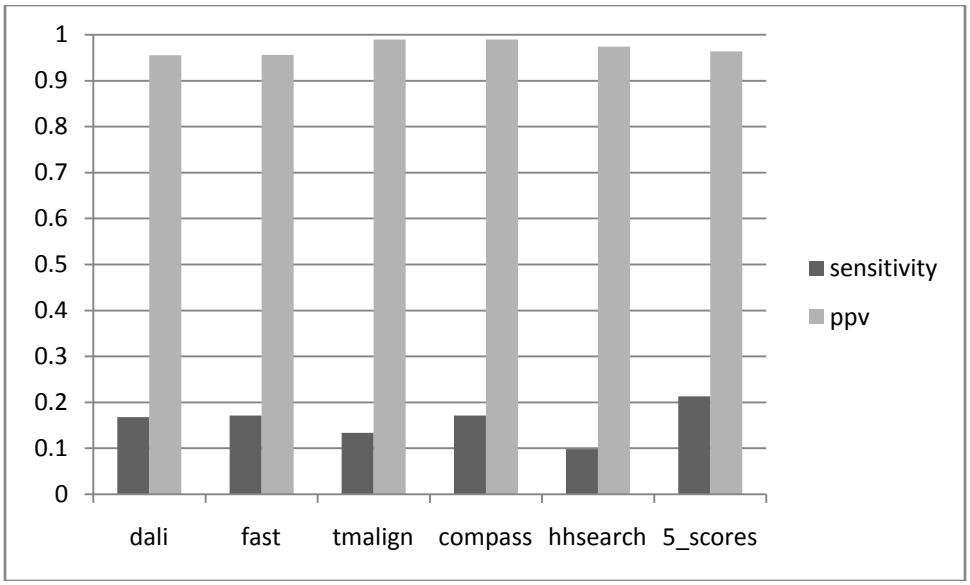


Figure 8 Comparison of SCG clusters from different similarity measures to SCOP superfamily. All labels are same as Figure 7. Notably, 5_scores, SCG clustering based on the optimally combined sequence and structural scores, shows highest sensitivity and maintain relatively good specificity. All sensitivity and specificity values are from Figure 4 except for 5_scores.

Figure 8 shows that the quality of the SCG clustering from combined score (label 5_score in the figure) is better than that from HHsearch, unlike Figure 7. This is because that SCOP family relationship is heterogeneous and not monophyletic, i.e. many close homologous proteins are divided into SCOP families and the division is not well defined.

Figure 7 and

Figure 8 show the importance of reference clustering and the caveat of measuring clustering quality by comparing to reference. Since the quality measure totally depends on reference clustering, it is very important to have very well defined reference clustering. The good reference cluster is, however, often not available as in the case of this study. And the idea of comparing to reference or “gold standard” is somewhat crossing borderline of trustworthy research, since clustering procedure needs

to be “unsupervised” or independent from previous knowledge by definition. It is not rare to see such a mistake that the clustering procedure is geared toward mimicking reference clustering from published literatures. This conceptual contamination was always big concern to us and we tried to prevent this problem all the time.

3.4 Improvement of SCG algorithm: Iterative SCG (iSCG)

Since the average cluster size of SCG clusters is still small (~2.5) after the combination of scores (SCOP superfamily average cluster size is 5), the very strict requirement of SCG was needed to be loosened. The iteration method was chosen to loosen SCG. This iteration idea is simple. After the SCG clustering is built based on a given similarity matrix, the next round is building a clustering of clusters. Each cluster in the previous iteration is treated as an object. And the similarity matrix is updated by maximum score between two clusters or a protein to a cluster. Then SCG algorithm is applied to the new similarity matrix for clusters. This iteration procedure can be done until there is no change in the clustering or convergence point. This iteration procedure effectively reduces the strict consistency rule of SCG algorithm. It is because the previously inconsistent objects can be consistent after updating the similarity matrix. The previous round of SCG and updating new similarity matrix between clusters in effect remove data points and in turn this functions as reduction in strictness of the method.

The iteration with updating by getting maximal scores between clusters is similar to single linkage clustering. (See detail in section 2.3.2.) This updating procedure can also be viewed as applying transitivity rule for the members in each cluster. If the

clusters contain homologs or evolutionarily related proteins then the transitivity rule is very natural rule to derive the relationships between clusters, because the evolutionary process is, in principle, transitive.

The iterative SCG (iSCG) has attractive features in theory, but it has one major drawback. The procedure will introduce more and bigger errors than SCG. The loosening of strict requirement of consistency rule will inevitably bring many bad clusters.

3.4.1 iSCG makes cluster size bigger.

iSCG clusters were built based on the combined scores. As expected from the thought experiment, the average cluster size is increasing as the SCG clustering goes iterations.

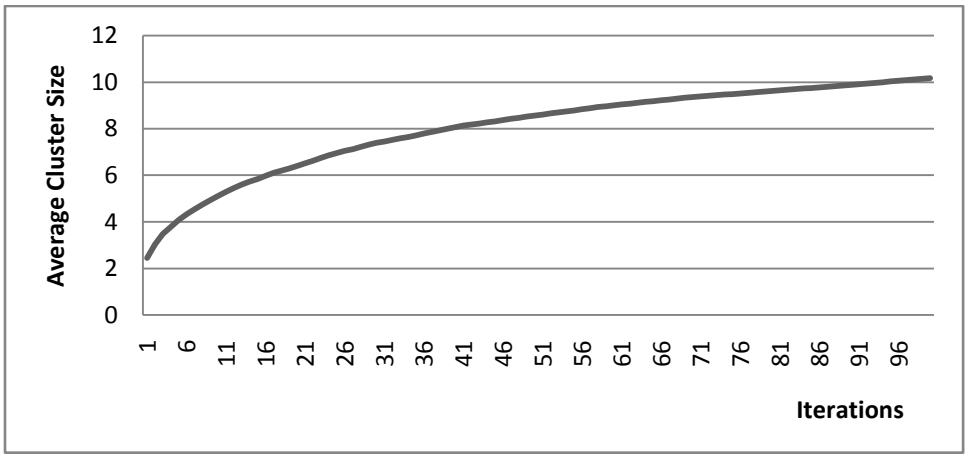


Figure 9 Average cluster sizes are shown for each iterations from iSCG clustering based on the optimally combined 5 sequence and structure scores. iSCG (or iterative SCG) builds clusters of clusters in each iteration until there is no new clusters can be formed. In each subsequent iterations, the ranking list is re-evaluated based on between clusters using maximum score between members against the other cluster. Thus, the first iteration result is just SCG clustering and subsequently the clusters are monotonically larger.

Since it is rather hard to determine natural point to stop the iteration, our first attempt was repeat iteration up to the point where there is not changes in clusters, i.e. repeat the iteration up to the convergence point. Figure 9 shows the trend of average cluster size. The iteration procedure did not converge up to 100 iterations shown in the figure.

3.4.2 iSCG makes the coverage bigger but severely reduces specificity.

The quality of iSCG clusters can be monitored using SCOP superfamily as before. The sensitivity and specificity was also used to show the similarity of iSCG clusters to SCOP superfamily. Notably, iSCG gives a cluster at all iterations. So the sensitivity and specificity can be shown and continuous lines showing the trends of the two values according to the iteration.

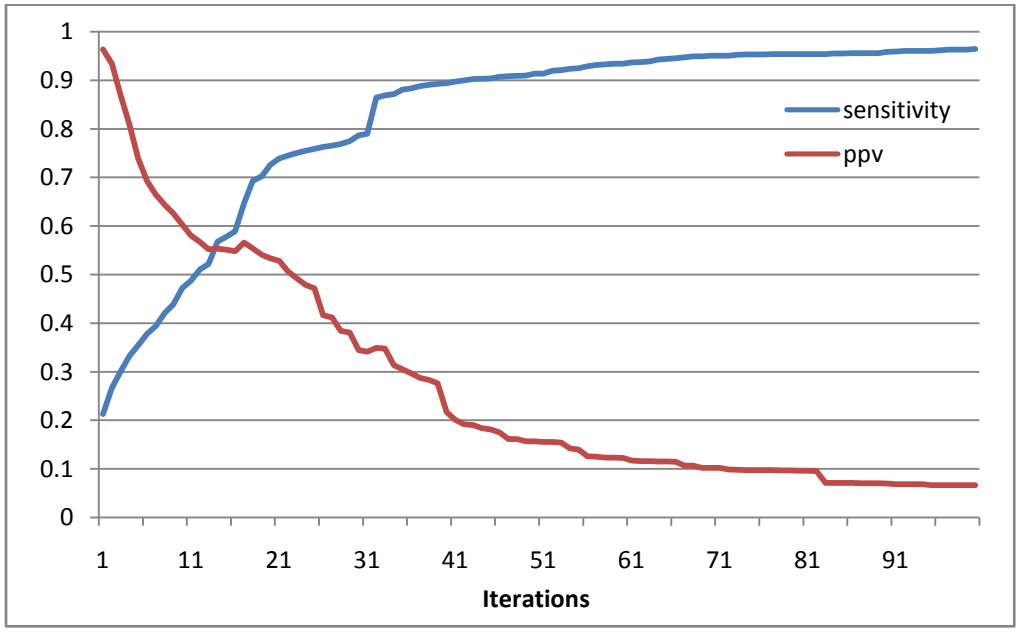


Figure 10 iSCG clusters compared to SCOP superfamily. Sensitivity and specificity (or ppv) was measured. Since iSCG gives a different clustering in each iterations, there are a

trajectories of sensitivity and specificities. Sensitivity increases with iteration because the larger clustering naturally includes more domains (or proteins) from the same SCOP superfamily. Specificity decreases because larger clustering naturally decreases the proportion of domains from the same SCOP superfamily. The sensitivity and specificity trajectories cross around 13th iteration.

Figure 10 shows that the quality of iSCG clusters compared to SCOP superfamily. The similarity was most similar around at 13th iteration. As expected from our thought experiments, the quality of clustering is decreasing after the 13th iteration. With this observation and other direct observation of bad clusters in iSCG clustering, iSCG was linking clusters based on trivial or random similarity scores (scores < 2) above 14 iterations. So the iSCG algorithm need to use certain cutoff to avoid linking clusters based on random similarities. Using transitivity principle, iSCG method clusters more homologous proteins. However iteration also made iSCG more vulnerable to errors and iSCG have to use cutoff value to keep the quality of clusters good.

3.5 Establishment of the cutoff value for iSCG

A cutoff is a value to determine boundaries of clusters. A good cutoff value is important for clustering, since too loose cutoff will include non-homologous proteins into the same cluster and too strict cutoff will divide homologous proteins into different clusters.

To find a good cutoff value, clusters were built at various cutoffs. Single linkage clustering method (de Hoon et al, 2004) was used because single linkage clustering is

relatively simple and straight forward method to see the changes of clustering according to the cutoff. Clusters were built based on the combined scores.

3.5.1 Cluster size distribution

Clusters were built by single linkage clustering method based on the combined score at various cutoffs ranging from 1 to 8. Then, the number of clusters bigger than 5, 10, 15, 20 and 40 (Figure 11) is counted at each cutoff.

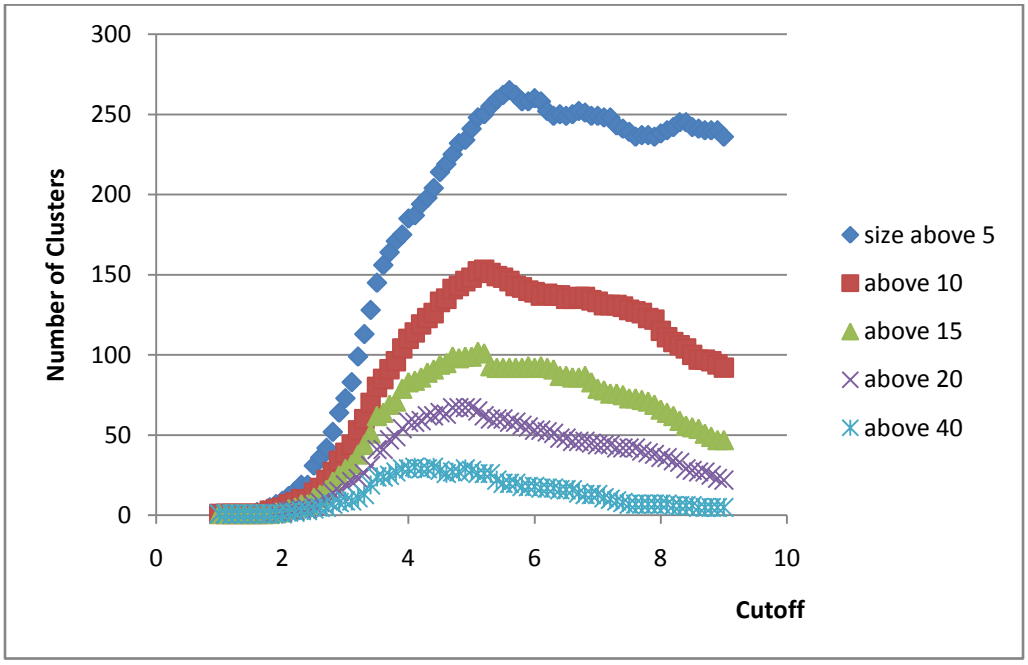


Figure 11 Cluster size distribution measured by number of clusters above 5, 10, 15, 20, and 40. X axis is cutoff score and Y axis is number of clusters meets the condition of size. The points in difference series (different size conditions) corresponding to the same cutoff value (value in X axis) are from the same single linkage clustering determined by the cutoff score.

Figure 11 shows that the number of clusters reaches maximum value between cutoff 4 and 6. At the low cutoff 1, the cutoff is too permissive. There are only few clusters and those clusters contain almost all dataset. So the number of clusters is very

small, i.e. the number of clusters above size 40 is 1. This cluster contains almost all proteins except few. As the cutoff value gets higher the number of clusters becomes higher because clusters are separated from few gigantic clusters. At the high cutoff 8, the cutoff is too strict. The clusters reduced very small in size, i.e. the number of clusters above cluster size above 40 is again reduced to 2. But those two clusters are not gigantic clusters but quite tight homologous groups of proteins. This trend is similar in other sizes measurements. In clusters of size bigger than 20, 15, and 10 shows similar trend as in size bigger than 40; the number of clusters increase starting from cutoff 1 and then reach maximum around cutoff score 4 ~ 6.

According to the data shown in Figure 11, the reasonable cutoff is likely located between score 4 and 6. This relatively simple approach to find a reasonable cutoff is good because this approach does not require any pre-defined knowledge about the dataset or proteins. However, the cluster size might not represent meaningful relationships between cluster members, i.e. homologous relationships between proteins. The quality of clusters can be directly checked by comparing to known “gold standard” or reference clustering defined by homologous relationships.

There is another trend to note from Figure 11. The numbers of clusters around cutoff score 8 increase as the threshold for cluster counting decreases from 40, 20, 15, 10 and 5. This trend is due to the number of tight clusters of very similar proteins. The number of those very tight clusters increases because the cluster size threshold decreases. The number of small tight clusters so many the cluster size distribution

bigger than cluster size 5 does not show clear maximum point as other distributions show.

3.5.2 Comparing to SCOP database

We can measure the quality of clustering by comparing the clustering of interest to reference or "gold standard" clustering as already discussed in previous sections. This quality measures can also provide valuable information to determine right cutoff score for reasonable clustering since the sensitivity and specificity are also related to cutoff values.

Clusters were built by single linkage clustering method based on the combined score at various cutoffs ranging from 1 to 8. Then, the sensitivity and specificity were measured at each cutoff (Figure 12).

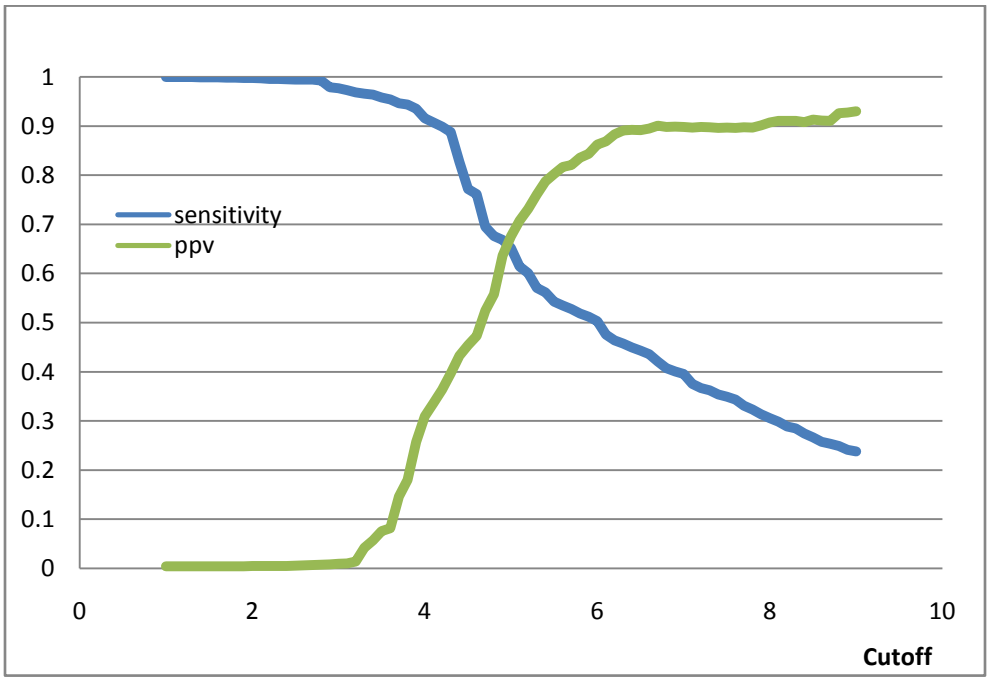


Figure 12. Comparison of clusters from single linkage clustering method to SCOP superfamily clustering. Single linkage clustering algorithm is used to build clusters based on the optimally combined 5 sequence and structure scores. Sensitivity and specificity (or ppv) is shown for different cutoff values. The sensitivity and specificity changes are in general same as the increasing iteration number in Figure 10. The sensitivity and specificity values cross each other around cutoff 5.

Figure 12 shows that the sensitivity is decreasing and specificity is increasing according to the cutoff value. As the size of clusters becomes smaller according to cutoff value, the general tendencies of two values are opposite. If a cluster is big (low cutoff value) and contains many proteins then it is easy to contain more correct links. But at the same time, it is also likely to have some wrong links within same cluster. In contrast if a cluster small (high cutoff value), sensitivity is small and specificity is high. So the sensitivity is generally decreasing and the specificity is generally increasing as the cutoff becomes higher. So the reasonable cutoff value can be the point where the two values, sensitivity and specificity are similar or balanced. And the balanced point, cutoff 5 where the sensitivity and specificity intersect, can be a reasonable cutoff value.

SCOP superfamily relationship is chosen to be a reference clustering. This is because superfamily relationship is where the proteins in the same group are mostly homologous and yet the similarities are not trivially high. In contrast, family relationship is too similar and many of the homologies are intertwined between families, i.e. the families are not generally monophyletic. SCOP fold relationship does not imply homology, so it is not suitable to be a reference for homologous protein classification.

3.5.3 Internal score distribution

Homogeneity within clusters and separation between clusters can be yet another way of monitoring cluster quality. However, traditionally developed methods were defined for distance matrices not similarity score matrices. So we used very simple statistics to approximately measure homogeneity and separation.

The minimum score between same cluster members represents the homogeneity of the cluster. The higher the minimum value the more cohesive the cluster. The maximum value between different cluster members represents the separation between clusters. The higher the value is the smaller the gap between clusters. However, the minimum and maximum values do not represent internal change of the clustering very well by themselves. So a balanced value of homogeneity and separation was devised by taking ratio of the two values.

$$C = \frac{\min_{x,y} S_{xy}}{\max_{i,j} S_{ij}}$$

Here, x and y are proteins within the same clusters. i and j are proteins belong to different clusters but i should belong to the same cluster which contains x and y. This ratio, C-value can be defined for a cluster bigger than size 1. There will be distribution of C-value for a clustering, since C-value is defined for a cluster. C-value above 1 means the minimum value is larger than the maximum value to other clusters and the cluster is relatively cohesive compared to the gap to the closest other cluster.

The median of C-value distribution in clustering at each cutoff was used to track the change of clustering property.

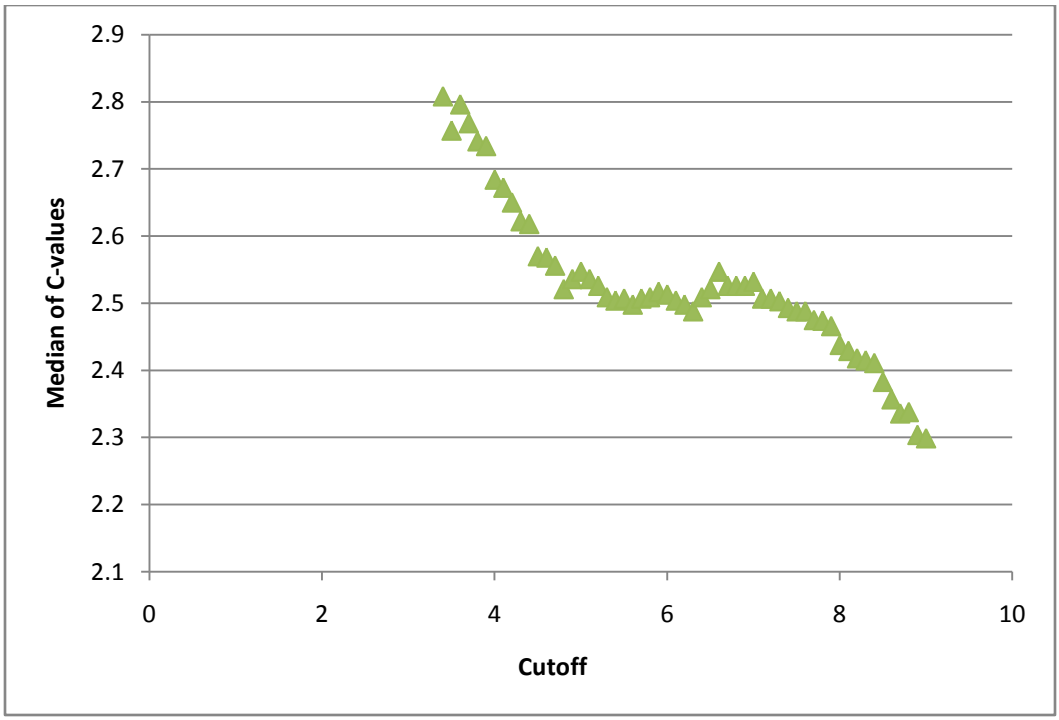


Figure 13. Median of C-values for clustering at each cutoff value. Clusters were built as in Figure 11 and Figure 12. The median of C-values are calculated from non-singletons (clusters size >2) in each clustering at different cutoff value. Notably, median of C-values form a plate around cutoff score 5-7.

The median C-value per clustering was above 1 and decreased from smallest cutoff to largest with 5-7 stable point. However, Figure 13 contradicts to the expected result. Since C-value is relative cohesiveness of clusters, the median value of cluster expected to increase as cutoff value increases. The higher cutoff is the stricter cutoff and the stricter cutoff makes should make smaller tighter clusters. Since the result from Figure 13 is based on every cluster in a cluster without singletons, different size threshold was tried (as in section 3.5.1) to dissect the behavior of median C-value per clustering.

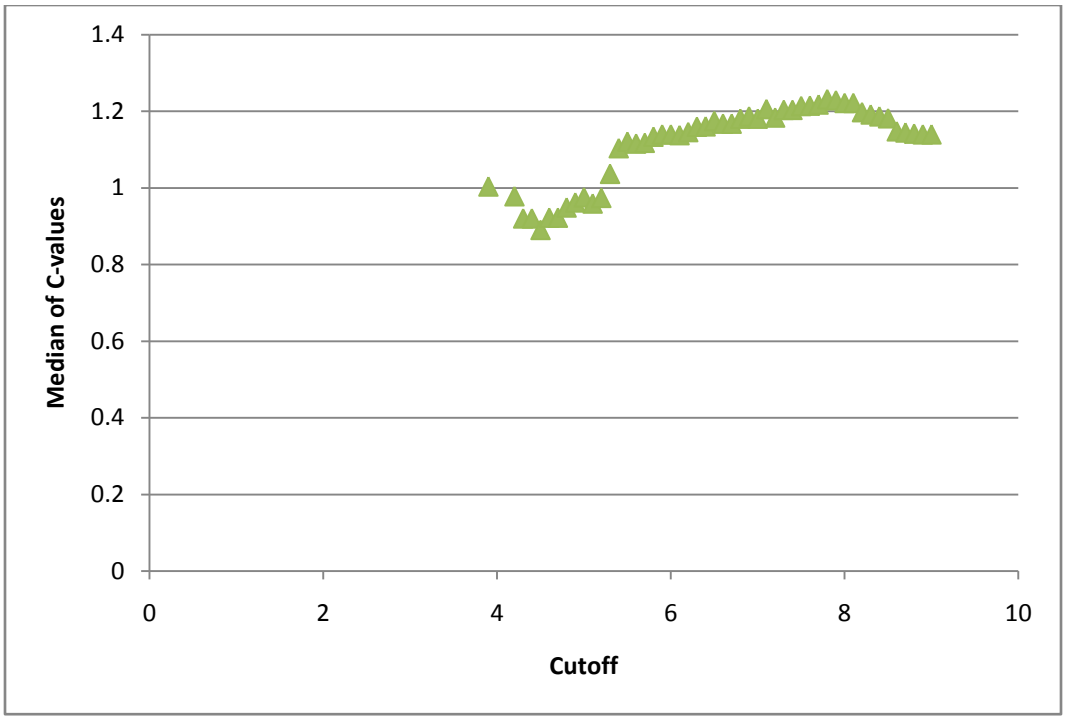


Figure 14. Median of C-values for clustering at each cutoff value for clusters bigger than 5. The median of C-values are from clusters bigger than size 5 in each clustering at different cutoff value.

Cluster size above 5 shows that the medians of C- values are relatively stable above cutoff larger than 6 and there were changes at cutoff value 5. One notable difference is that the change according to various cutoffs is inverse direction of previous figure, all clusters without singletons. Notably, this Figure 14 shows the expected trend. The difference between Figure 13 and Figure 14 are the small clusters between sizes 2-5.

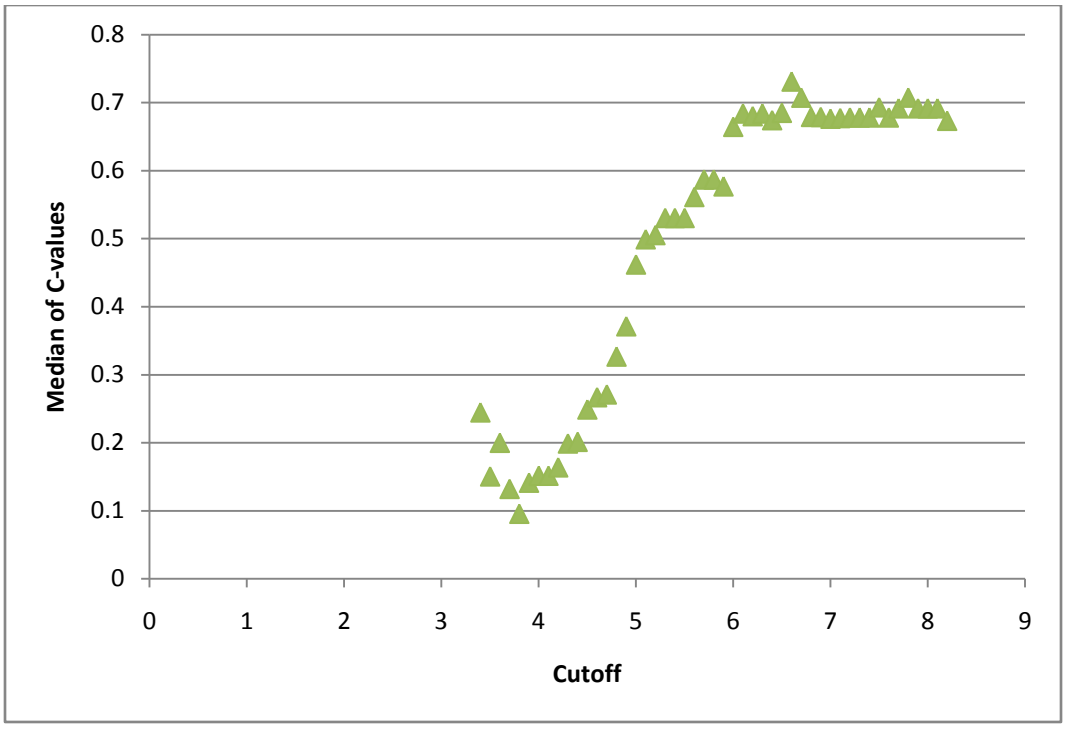


Figure 15. Median of C-values for clustering at each cutoff value for clusters bigger than 20. The median of C-values are from clusters bigger than size 20 in each clustering at different cutoff value.

The expected trend is more obvious in cluster size of bigger than 20. From median of C-values change at different cluster sizes, it is likely that the clusters from cutoff above 6 and below 4 might be quite different in their quality.

The difference between Figure 13 and other two figures, Figure 14 and Figure 15, is very intriguing problem. Especially the trend in Figure 13 is hard to intuitively understand. The reason behind the unexpected trend in Figure 13 might be those clusters that are resistant to be clustered with gigantic clusters are super-tight and very far (larger distance to closest cluster, smaller in terms of score) from other clusters. Those "super-tight" clusters outnumber normally behaving clusters for lower score

cutoff, thus the median C-value is high at low cutoff. The median C-value for high cutoff is lower because the relative effect of “super-tight” clusters is smaller. There are more “normal” clusters separated because of higher cutoff (Figure 13). This might still affect for the clusters size above 5 in Figure 14. Since the two different groups (“super-tight” versus “normal”) might be more balance for clusters size above 5, Figure 14 probably shows small change in absolute value of median C-values according to the change in cutoff.

Finally, based on the conclusions from the sections 3.5.1, 3.5.2, and 3.5.3 it is very likely that the clusters are different around the cutoff value of combined score 5. This cutoff value was also observed in our manual analysis of clusters during the course of experiments. It was observed that many wrong clusters were prevented to form at cutoff 5.

CHAPTER 4

Analysis of Protein Classification

After established all methods for protein classification, iSCG at cutoff 5 clustered ~7000 proteins into ~1500 proteins. In this chapter, the classification result will be discussed.

4.1 Global Analysis of Protein Classification

4.1.1 Network like representation of classification

The clustering result can be represented as network of above significant score cutoff 5. The representation was prepared using Cytoscape program (Shannon et al, 2003). As shown in Figure 16, there are few big clusters and many singleton clusters (clusters having only one domain as a member) up to 10% of total dataset. This circular layout of network visualization also helps to identify well connected clusters and sparsely connected clusters. This layout forms nice round circle for a cluster when all members in the cluster are well connected with similarity scores above the threshold.

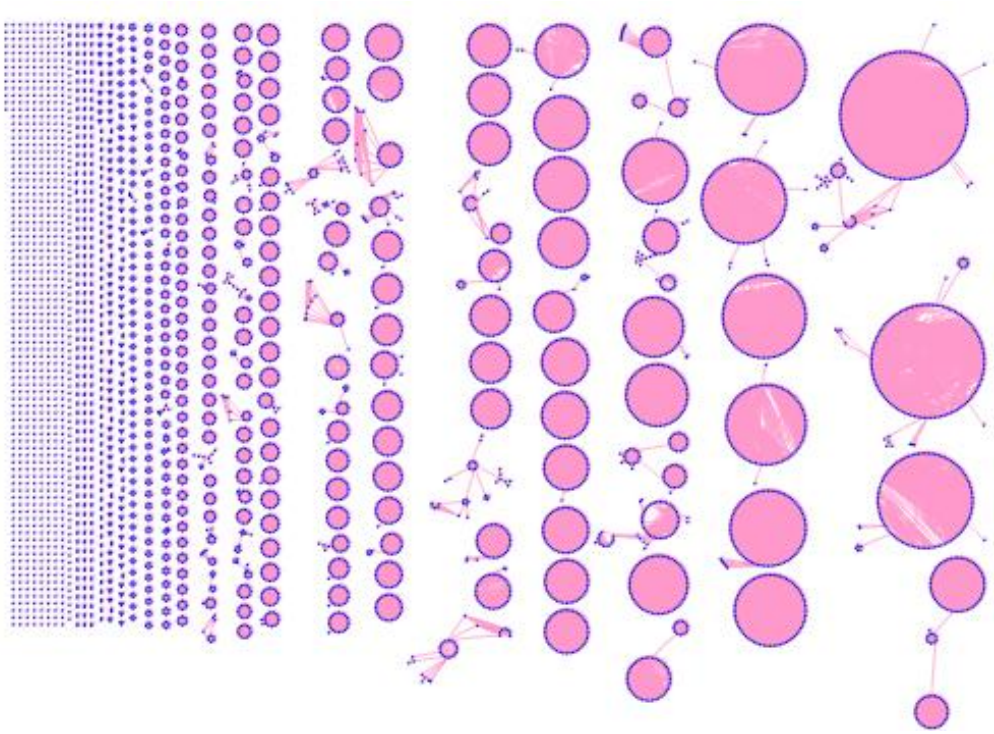


Figure 16 Classification result represented as network. The blue dot (node) represents a domain and magenta lines (edges) represent significant similarity above threshold, the combined score bigger than 5. Groups of inter connected domains (clusters) are ordered by their sizes, the number of nodes.

Outliers (sparsely connected members in a cluster) are represented like bristles (or small antennas) out of the circle. A cluster on upper right corner of Figure 16 shows the case clearly. If the members of clusters are form distinct groups and the relationship

between groups are rare, then the circular layout makes distinct circles connected by few lines. A cluster on the lower right corner of Figure 16 shows three distinct groups of domains connected by a line representing that there are three distinct groups of proteins in the cluster.

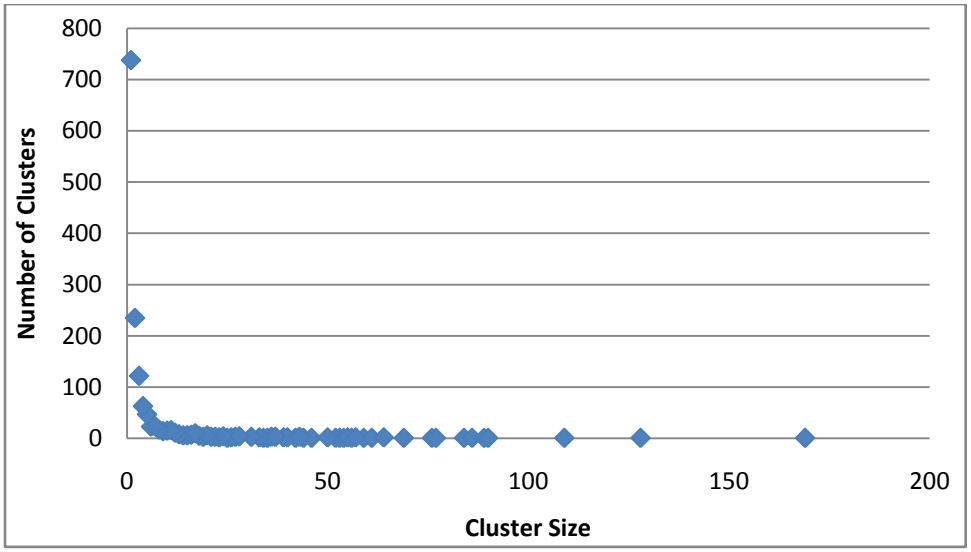


Figure 17 Number of clusters for each iSCG cluster size. This distribution of cluster size shows that there are very many clusters of small size in the classification and small numbers of big clusters.

Briefly mentioned sizes of clusters can be shown more clearly with Figure 17 and Figure 18. Both figure shows that the number of clusters decrease very sharply for increasing cluster size. The relationship between cluster size and number of clusters can be modeled by power-law distribution (Dokholyan, 2005). This power-law distribution is also called as scale-free network or “small world effect”. This power-law distribution of internet web site connectivity or actor network is very interesting property of social

networks, like hub node. The hub node is a node (corresponding a blue dot or domain in previous network figure) that has many connections. It has been proposed that the preferential attachment to the nodes that have many connections is the probable cause of power-law (Qian et al, 2001). In contrast to this idea of preferential attachment of node, the cluster size distribution can be power-law from random connection (Dokholyan et al, 2002). This implies that the proteins evolve randomly and shaped current protein space.

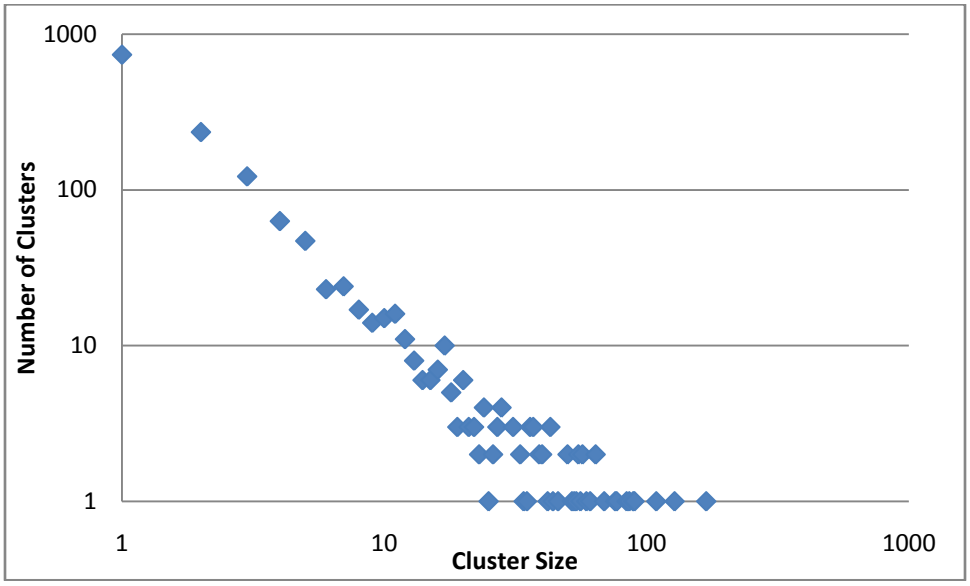


Figure 18 Number of clusters for each iSCG cluster size. This figure is representing same data as in Figure 17. X and Y axes are changed into log scales to show the power-law in cluster size distribution.

SCOP superfamily size also shows the same power-law distribution like our classification. However, all those analysis of general properties cannot tell much about the proteins. So we checked the top10 biggest clusters.

4.1.2 Superclusters

We checked the top 10 biggest clusters and named them as superclusters. In general they are from well known superfolds defined by Orengo and coworkers at 1994 (Orengo et al, 1994). More specifically, Orengo and coworker defined 9 superfolds, Globin, Beta-trefoil, Helical bundle, Immunoglobulin folds, Ferredoxin-like fold, Jelly roll, Rossmann-like fold, Beta-grasp, and TIM barrel (Figure 20).

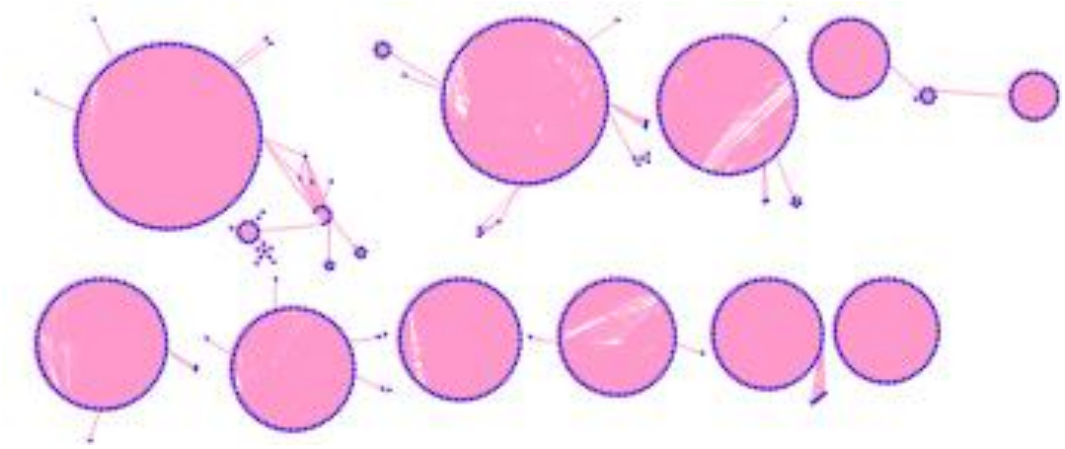
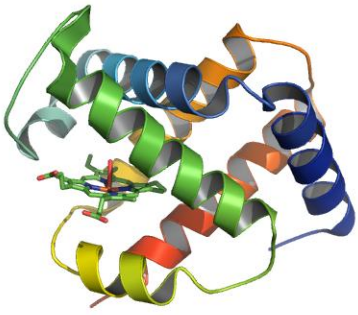


Figure 19 Superclusters: Top 10 biggest clusters. This is zoomed up view of 10 clusters biggest in size shown in previous figure. From left to right and top row to bottom the numbers, 1-10 are used to designate each cluster.

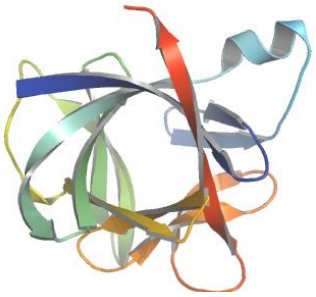
Supercluster number	Supercluster proteins	Supercluster size	Corresponding superfold
1	**Immunoglobulins**	169	Immunoglobulin
2	**TIM barrel**	128	TIM barrel
3	OB fold	109	-
4	**P-loop hydrolases**	90	Rossmann-fold
5	**beta-Grasp**	89	Beta-Grasp
6	**Galatose binding domain-like**	86	Jelly roll
7	HTH	84	-
8	Thioredoxin fold	77	-
9	beta-propellers	76	-

10	**alpha/beta hydrolases**	69	Rossmann-fold

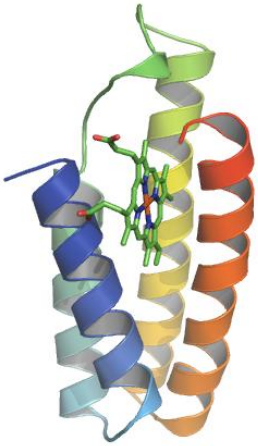
Table 1. List of proteins in 10 biggest clusters. Bold faced proteins are overlapping proteins with Orengo's superfold list. The superclusters are number by 1-10 by size. Supercluster proteins are the majority of proteins in each supercluster. Supercluster size is the number of proteins in each super cluster. Corresponding superfold column shows the matching superfold defined by Orengo et al. [ref] If there is no matching superfold for supercluster, then "–" is in the column.



Globin (1thb)

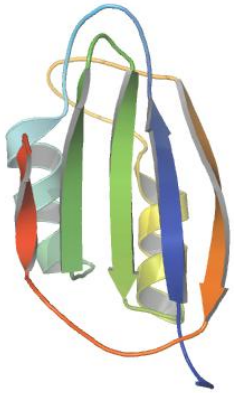


Beta-Trefoil (1i1b)

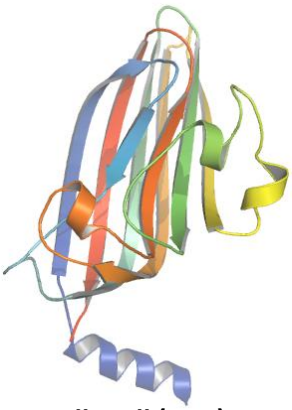


Helical bundle (256b)

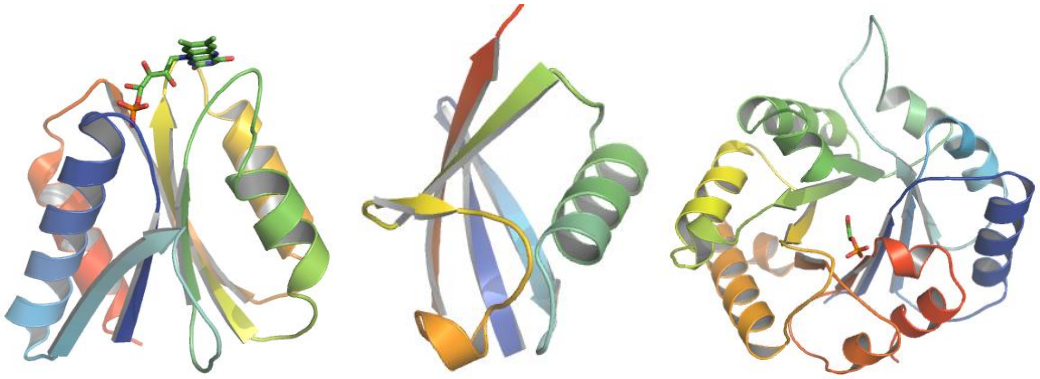
Immunogobulin (2hre)



Ferredoxin fold (1aps)

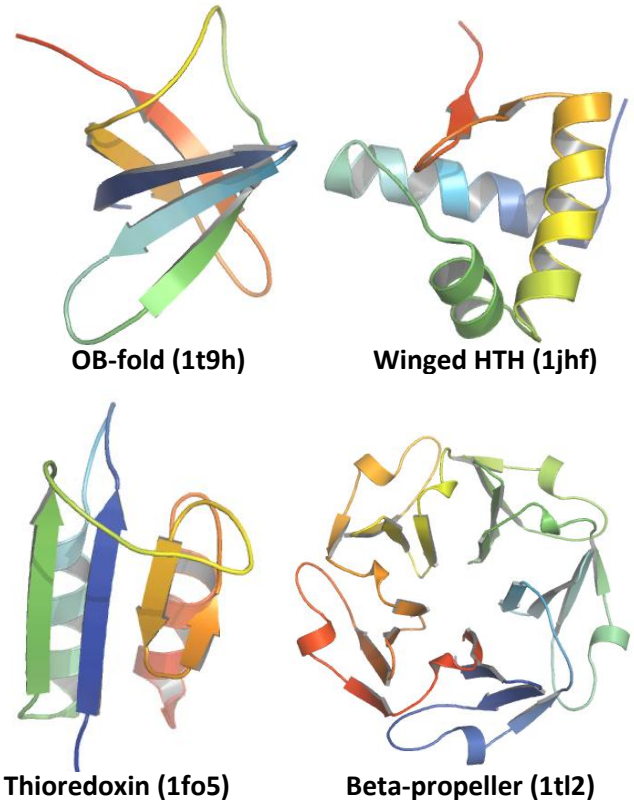


Jelly roll (2stv)



Rossmann-like (4fxn) **Ubiquitin-like (1ubq)** **TIM barrel (7tim)**

Figure 20 Structures of superfold proteins selected by Orengo and the coworkers (Orengo et al, 1994). All the structures are colored from blue to red. Blue end represents N-terminus end of proteins and red colored end represents C-terminal end of proteins. Each structure represents the superfold shown as labels. The structures are from the PDB codes shown in parentheses. The superfold names also appeared in supercluster are in bold face.



OB-fold (1t9h) **Winged HTH (1jhf)**

Thioredoxin (1fo5) **Beta-propeller (1tl2)**

Figure 21 Representative structures of superclusters. Only structures did not appear in superfold list (Figure 20). Structures are shown in blue to red colors. N-terminal end of proteins are blue and C-terminal ends are red.

When we compared superclusters and superfolds, 4 superclusters are not in superfold list; Supercluster 3, 7, 8 and 9 (Table 1, Figure 21). Since it was 1994 when Orengo and coworkers made the superfold list (Figure 20), it is plausible that the 4 supercluster proteins were not recognized at that time. To test this idea, the deposition time of OB-fold proteins were plotted in Figure 22.

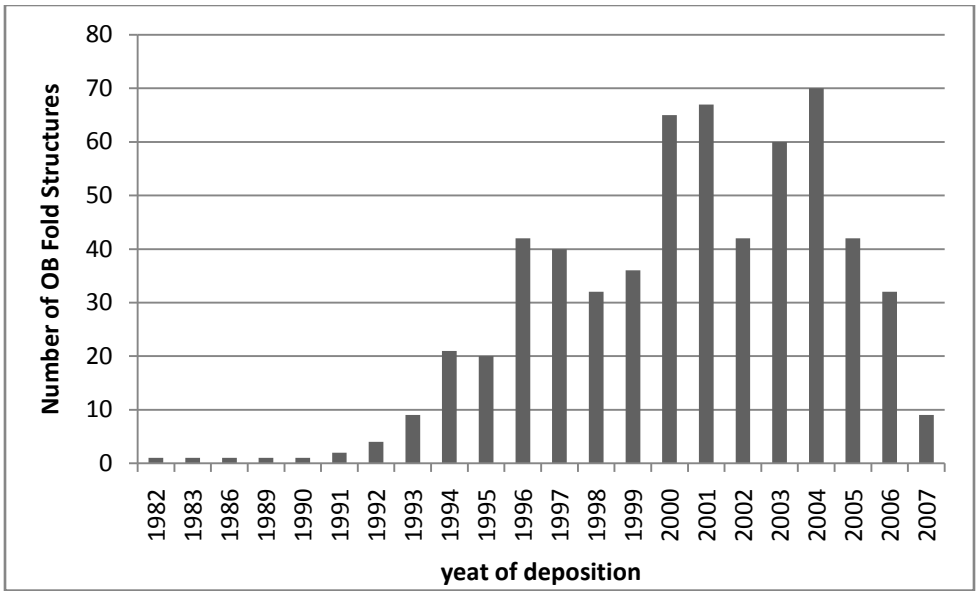


Figure 22 Number of OB fold structures deposited into PDB at each year.

As anticipated, many OB-fold proteins were solved after mid 1990s. It is out of range of this work, but it would be intriguing to follow the representations of folds in researchers' minds by monitoring the trajectory of structure deposition and citation of

those structures in the literatures for each structural fold. This result would give us the bias of PDB and structural biology field. Also if we distinguish the scientific literatures from other biological training than structural biology, that would be interesting result of how structures are represented in general biological scientists' mind other than structural biologists.

Another interesting difference between superclusters and superfolds is that there are folds that were in the superfold list but not in the supercluster list, i.e. Globin, Beta-trefoil, Helical bundle, Ferredoxin-like fold. This highlights the conceptual difference between superfolds and superclusters. Superfolds were defined as common folds among non-homologous proteins. Superclusters are, by definition of our classification, supposedly to be clusters of homologous proteins. For example ferredoxin-like fold is very common fold in current PDB. Ferredoxin-like fold proteins are divided into ~30 clusters in iSCG classification. Ferredoxin-like fold proteins are thought to be analogous not homologous. As shown in Figure 23, many ferredoxins share the core structure. They share two layers of alpha and beta structure and beta-alpha-beta-beta-alpha-beta topology with 2-3-1-4 beta strand order (Figure 23). Exopepidase dimerization domains and RNA binding domains are, however, very different in the secondary structure length. This means that if the two proteins diverged from the same ancestor, they have insertions or deletions (indels) on every secondary structural elements. Since proteins with simple topology can more easily evolved *de novo,*

ferredoxin proteins are more likely evolved independently than diverged from the ancestral protein.

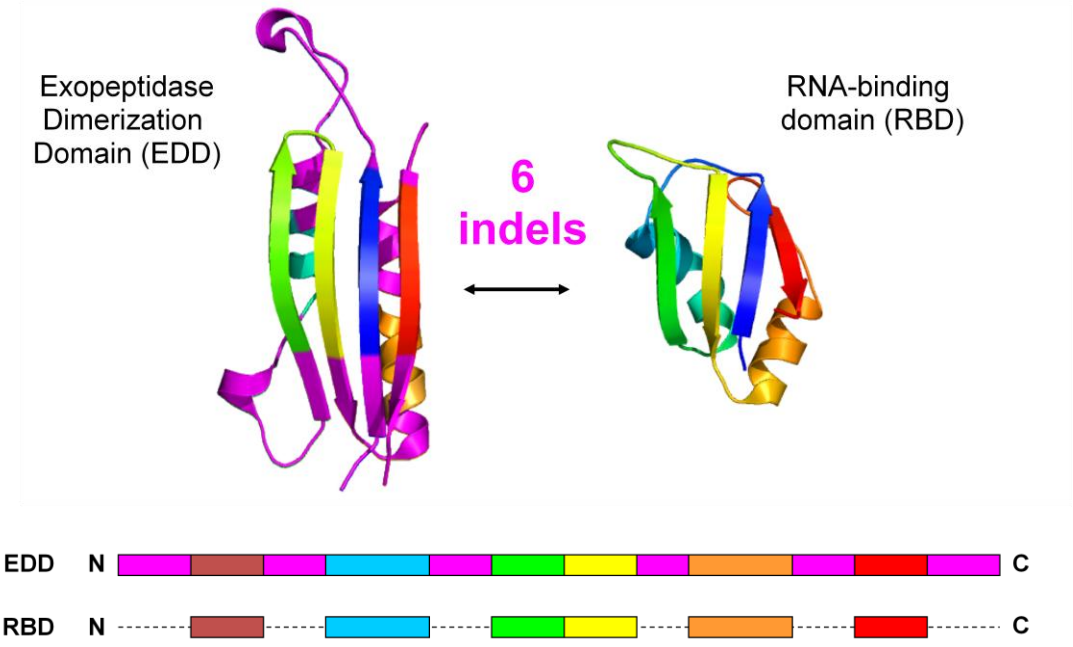


Figure 23 Analogous ferredoxin fold proteins. Exopeptidase dimerization domain and RNA-binding domains are shown in left and right respectively. In upper panel, two proteins are shown in structure figure. In lower panel, the two proteins are shown in linear line to show the corresponding regions. The two structures are shown from blue to red as in rainbow except purple colored region. Equivalent region of two proteins are shown in the same color. Regions in purple color represent insertions or deletions (indel).

Another interesting example among superclusters is supercluster #4 P-loop hydrolases (Figure 19). This cluster has three distinct groups. Two bigger groups are linked through the small connecting group. It is generally very hard to argue homology when the similarity is very remote (like in the case of ferredoxin-like fold example). If two remote groups of proteins are connected though intermediate groups of proteins,

then it is much more likely that the two groups of marginally similar proteins evolved from the same ancestor. This phenomenon is called transitivity principle and this transitivity is essentially the only practical method to infer remote homology reliably. This transitivity is analogous to the famous example of archaeopteryx (Kundrat, 2004) that was thought to link dinosaurs and current birds. Without the fossil records of archaeopteryx and other intermediate form between dinosaurs and birds, it would be much harder to establish evolutionary relationship between birds and dinosaurs. All the members in supercluster 4 are P-loop hydrolases. More specifically the two separate groups (not directly connected) are nucleotide kinases and G proteins. They are connected through small intermediate group of nitrogenase-iron like proteins. Now most researchers in the field regard them as homologous proteins, but they were not considered as homologous in mid 80's. Especially, Thomas Steitz suggested that the nucleotide kinases and G proteins are analogous proteins because the beta sheet topology of nucleotide kinases and G proteins are quite different (Figure 24). Both proteins contain family specific beta-hairpins at different places in beta-sheet and this difference is hard to be explained simple fold migration (Grishin, 2001). The intermediate group, nitrogenase iron like proteins, however can help to establish the evolutionary link very solidly. This middle group retain N-terminal half of protein very similar to nucleotide kinases. C-terminal halves of nitrogenases and G proteins are also very similar. This transitivity rule is in fact the workhorse of our iSCG methodology.

Without this transitivity rule, the homologous groups defined by our method would be much smaller than we have.

Figure 24 Different beta-sheet topology in Nucleotide kinase and G protein. The parallel beta-strands are colored in yellow. Red and blue are anti-parallel beta-strands. Each beta-strand is assigned numbers 1-6 from N-termini to C-termini in both structures.

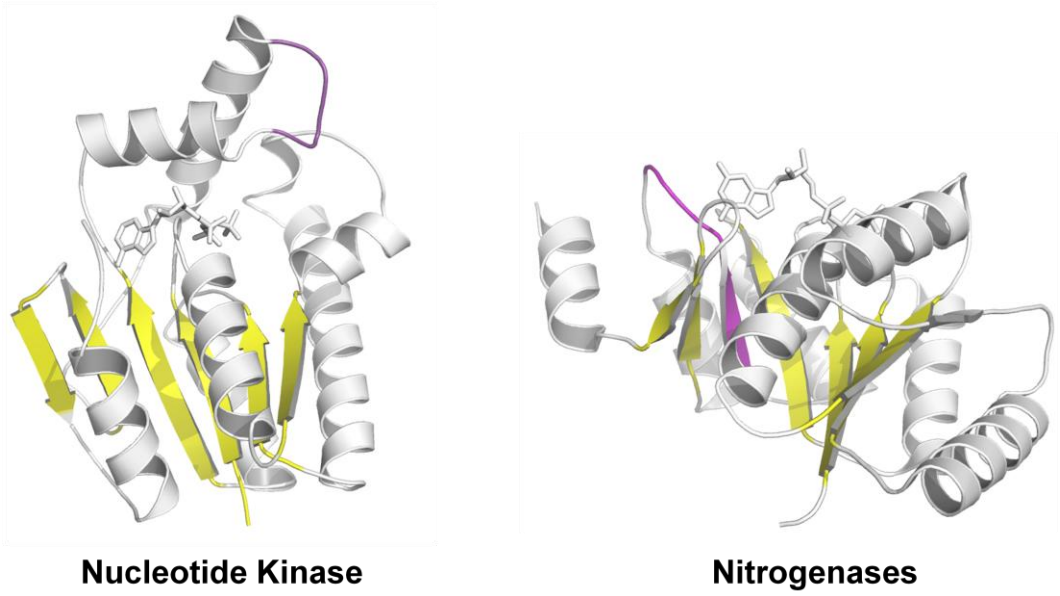


Figure 25 Similarity between nucleotide kinase and nitrogenase. Yellow color represents similar region in the two proteins. Differences between the two proteins are shown in purple color.

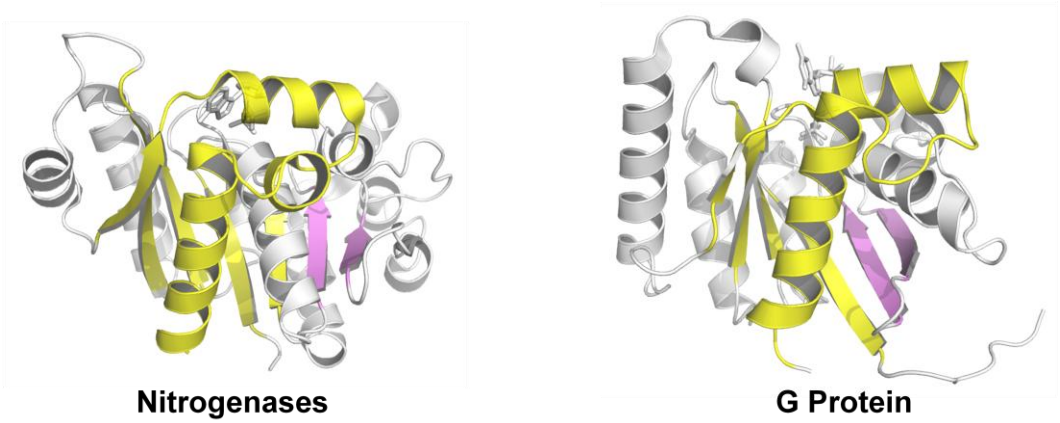


Figure 26 Similarity between nitrogenase and G protein. Yellow color represents similar region in the two proteins. Differences between the two proteins are shown in purple color.

4.2 Analysis of Difference between SCOP and iSCG Protein Classification

We compared iSCG protein classification and SCOP superfamilies. Among ~1500 clusters of iSCG clustering, we found 84 clusters contain different SCOP superfamilies.

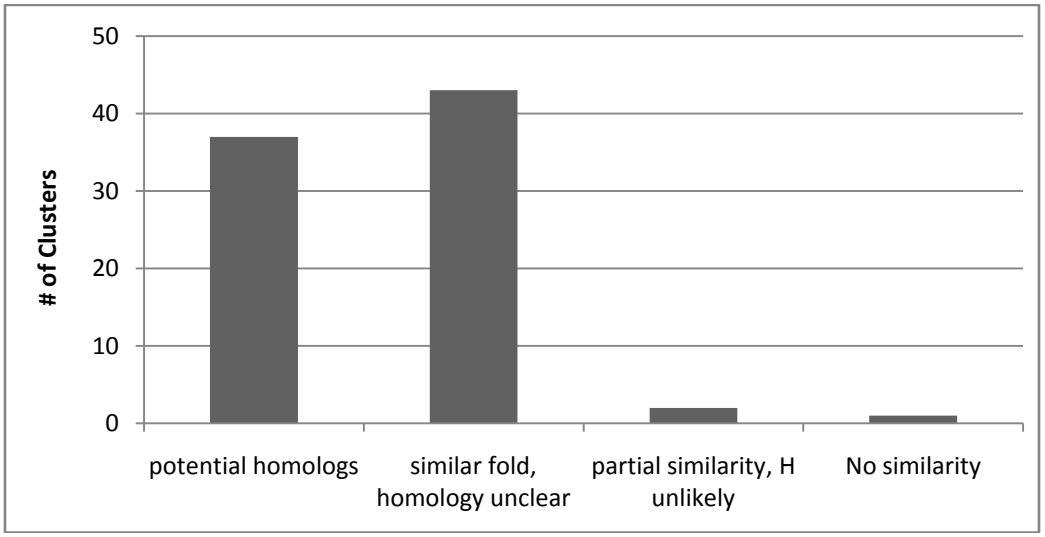


Figure 27 Manual checking result of 83 clusters that contain different SCOP superfamilies. Potential homologs are clusters that are very likely to be homologous proteins. Similar fold, homology unclear represents clusters that is possibly to be homologous but the authors cannot be convinced entirely the evolutionary relationships within the group. Partial similarity H unlikely is clusters that have low chance to be clusters of homologous proteins. No similarity is blatantly wrong clusters.

We divided these 84 clusters different from SCOP superfamily into four categories according to the quality of clusters. Potential homolog is a category for clusters that is quite likely to have homologs in them. The criterion of homology here is structural/sequence/functional features that are observed in homologous proteins, or

previous published established results even though SCOP superfamily did not reflected them. Similar fold, homology unclear is a category for clusters that contain proteins similar proteins but not obvious homologs like potential homolog category. This similar fold homology unclear category can be viewed as a category for clusters that contains dinosaurs and birds without archaeopteryx. As more sequences and structures are revealed that fill the gaps between two remote groups, the clusters belong to similar fold unclear homology will gradually move into potential homology category. Partial similarity H unlikely is a category that contains clusters of likely not homologous. Clusters in this category and no similarity category can be considered as wrong to us because the clusters are supposedly to contain homologous proteins. We put 80 clusters into two correct categories, potential homolog and similar fold homology unclear and classified three classes into two wrong categories. Clusters in each category will be discussed detailed fashion in next section.

It is important to compare our result to previously known information. If the analysis just cataloging how many were same as previously defined evolutionary groups and how many were different, then we would learn very little about nature by this research. So we compared our clusters to SCOP superfamilies since SCOP superfamily relationship is the most comprehensive and very accurate deposition of currently known homologous group of proteins and then deeply analyzed those clusters that are different from SCOP superfamilies.

Before discussing the detailed analyses about those selected 83 clusters, it is necessary to differentiate the over-clustering and under-clustering. When we compare to different clustering schemes, the difference can be classified as over-clustering and under-clustering. Those 83 clustered discussed in the previous paragraph, are essentially over-clustering of iSCG classification compared to SCOP superfamilies. In our research, we paid more attention to those over-clustering because even if we made many more potential homologous protein groups but the new clusters are useless if they are not reliable. And if we made under-clustering in iSCG classification, that is not that bad because we can consult both iSCG result and SCOP superfamilies and select bigger clusters. This under-clustering problem can be roughly assessed using the coverage or sensitivity already measured (Figure 10). The sensitivity of iSCG clustering compared to SCOP superfamily is about 60%. Therefore, in some clusters iSCG classification under-clustered compared to SCOP and in some clusters iSCG over-clustered than SCOP. But when iSCG clustered proteins into same cluster, those proteins are very likely to be homologous.

4.2.1 Putative Homologs

As stated in previous section, detailed analysis on clusters containing different SCOP superfamilies are described in this section. The main focus of this analysis is again to confirm if iSCG classification is indeed reasonable or not. Especially this subsection contains iSCG clusters that are more likely to be right.

Flavodoxin-like fold proteins (Cluster 6)				
family merge; superfamily break; fold split;				
Note: Significant sequence/structural similarity. Share active site region				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a/b	Flavodoxin-like	CheY-like	24(25)	d1r8ja2 d1a04a2 d1qo0d_ d1dcfa_
		Succinyl-CoA synthetase domains	2	d1eucb1
		Cobalamin (vitamin B12)-binding domain	5	d1bmta2

Leucine-rich repeats (Cluster 42)				
family merge; superfamily break; fold split;				
Note: LRRs show highly similar hydrophobic sequence pattern and high structural similarity. This cluster has a problem in breaking L domain family. L domain family proteins form another cluster (Cluster 2366) by themselves. All leucine-rich repeats are probably homologous and needs to be clustered together.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a/b	Leucine-rich repeat, LRR (right-handed beta-alpha superhelix)	RNI-like	5	d1io0a_ d1yrga_ d1fqva2
		L domain-like	12(18)	d1koha1 d1a9na_ d1dcea3 d1w8aa_ d1h6ta2 d1jl5a_ d1ogqa_
		Outer arm dynein light chain 1	1	d1m9la_

TIM barrels (Cluster 43)				
family merge; superfamily break; fold split;				
Note: All TIM barrel proteins share alpha/beta barrel topology and general active site region.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a/b	TIM beta/alpha-barrel	Triosephosphate isomerase (TIM)	3	d1aw1a_
		Aldolase	31	d1gzga_ d1f74a_ d1of8a_ d1vlia2 d1nvma2 d1dosa_
		Phosphoenolpyruvate/pyruvate domain	11(12)	d1dxea_ d1e0ta2 d1kbla1 d1m3ua_ d1dqua_
		Malate synthase G	1	d1d8ca_
		RuBisCo, C-terminal domain	4	d1geha1
		Bacterial luciferase-like	7	d1nfp__ d1ezwa_ d1nqka_ d1luca_
		PLC-like phosphodiesterases	5	d1o1za_ d2plc__ d1djxa3
		Cobalamin (vitamin B12)-dependent enzymes	5	d1ccwb_ d7reqa1 d1xrsa_ d1eexa_
		Ribulose-phoshate binding barrel	19	d1nsj__ d1qo2a_ d1dbta_ d1tqja_ d1y0ea_
		tRNA-guanine transglycosylase	2	d1iq8a1
		Dihydropteroate synthetase-like	4	d1eyea_ d1f6ya_
		UROD/MetE-like	4	d1j93a_ d1u1ja2
		FAD-linked oxidoreductase	2(3)	d1b5ta_
		Pyridoxine 5'-phosphate synthase	1	d1m5wa_
		Monomethylamine methyltransferase MtmB	1	d1ntha_

		Homocysteine S-methyltransferase	2	d1lt8a_
		(2r)-phospho-3-sulfolactate synthase ComA	1	d1qwga_
		Radical SAM enzymes	3	d1olta_ d1tv8a_ d1r30a_
		Thiamin phosphate synthase	2	d1xi3a_
		CutC-like	1	d1twda_
		ThiG-like	1	d1xm3a_
		FMN-linked oxidoreductases	15	d1viza_
		Inosine monophosphate dehydrogenase (IMPDH)	3	d1jr1a1

HeH motifs (Cluster 46)

family break; superfamily break; fold split;

Note: The two proteins aligned with loops/turns without gaps. Roh termination factor, N-terminal domain probably diverged from SAP domain superfamily. SAP domain superfamily should be clustered together but they are very diverse in sequence structure so SAP domain superfamily is split into three small clusters in iSCG clustering.

Class	Fold	Superfamily	Ndom	Representative SCOPid
a	LEM/SAP HeH motif	SAP domain	1(5)	d1y02a1
		Rho termination factor, N-terminal domain	1	d1a62_1

Beta-trefoils (Cluster 65)

family merge; superfamily merge; fold clean;

Note: They are probably homologous (Ponting & Russell, 2000).

Class	Fold	Superfamily	Ndom	Representative SCOPid
B	beta-Trefoil	Cytokine	9	d1l2ha_ d1rg8a_
		Ricin B-like lectins	14	d1qxma1 d1upsa2 d1dqga_
		Agglutinin	2	d1jlxa1
		STI-like	8	d1a8d_2 d1wba__
		Actin-crosslinking proteins	5	d1dfca2 d1hcd__
		MIR domain	2	d1t9fa_
		DNA-binding protein LAG-1 (CSL)	1	d1ttua3
		AbfB domain	1	d1wd3a2

Double psi beta-barrels (Cluster 89)

family merge; superfamily merge; fold clean;

Note: Structural similarity between the two superfamily is moderately high (DaliLite Z-score: 7) and sequence similarity is low. But from the point of view that this complex fold is hard to be independently evolved two times, they are more likely to be homologous.

Class	Fold	Superfamily	Ndom	Representative SCOPid
B	Double psi beta-barrel	Barwin-like endoglucanases	3	d1bw3__ d1n10a2 d2eng__
		ADC-like	15	d1aa6_1 d1cr5a1 d1ppya_

Calponin-homology domain-like proteins (Cluster 156)				
family merge; superfamily merge; fold split;				
Note: Very high sequence similarity. Peculiar fold.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
A	CH domain-like	Calponin-homology domain, CH-domain	10	d1aoa_1
		Hook domain	1	d1wixa_

Winged HTH proteins (Cluster 167)				
family break; superfamily break; fold split;				
Note: HTH proteins are probably homologous (Aravind et al, 2005).				
Class	Fold	Superfamily	Ndom	Representative SCOPid
A	DNA/RNA-binding 3-helical bundle	Winged helix DNA-binding domain	80(113)	d1ri7a1 d1aoy__ d1repc2 d1ldda_ d1w1we_ d1u5ta2 d1lvaa2 d2foka3 d1oywa1 d1ufma_ d1d5va_ d1hsta_ d1xn7a_ d1hw1a1 d1ucra_ d1oyia_ d1t6sa2 d1fzpb_ d1dpua_ d1omia2 d1in4a1 d1xd7a_ d1bia_1 d1jhfa1 d1fx7a1 d1mkma1 d1b9ma1 d1ixca1 d1fp1d1 d1bjaa_ d1r7ja_ d1tbxa_ d1stza1 d1tqia1 d1ulya_ d1yg2a_ d1ku9a_ d1sfxa_ d1bm9a_ d1y0ua_ d1q1ha_ d1mzba_ d1p6ra_ d1o57a1
		C-terminal effector domain of the bipartite response regulators	4(9)	d1gxqa_

RuvA C-terminal domain-like (Cluster 194)				
family break; superfamily break; fold split;				
Note: SCOP mentions that UBA and HBS1 are possibly related. Between UBA and CRAL the closest link is 4481 and 4353 with DALI 6.7 and HH 0.5.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
A	RuvA C-terminal domain-like	UBA-like	14(18)	d1oqya1 d1otra_ d1v92a_ d1efub3
		CRAL/TRIO N-terminal domain	3	d1aua_1
		HBS1-like domain	1	d1ufza_

Immunoglobulin-like beta-sandwich (Cluster 217)				
family break; superfamily break; fold split;				
Note: The two groups show very high sequence/structural similarities. Some proteins in Fibronectine type III superfamily are topologically changed to make more stable dimmers, hence they were not clustered.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
B	Immunoglobulin-like beta-sandwich	Purple acid phosphatase, N-terminal domain	1	d4kbpa1
		Fibronectin type III	56(60)	d1f42a2

Beta-grasp proteins (Cluster 223)				
family break; superfamily break; fold split;				
Note: Most beta-Grasp fold members are probably homologous (Iyer et al, 2006). Interesting exception is members of streptokinases and superantigens. Those outliers are quite different from other beta-grasp proteins.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a+b	beta-Grasp (ubiquitin-like)	Ubiquitin-like	36	d1k8rb_ d1gnua_ d1l7ya_ d1gg3a3 d1v2ya_ d1h8ca_
		TGS-like	3	d1jala2 d1tkea1
		Doublecortin (DC)	1	d1mg4a_
		TmoB-like	1	d1t0qc_
		CAD & PB1 domains	11	d1c9fa_ d1pqsa_
		MoaD/ThiS	8	d1rwsa_ d1fm0d_ d1wgka_
		2Fe-2S ferredoxin-like	17	d1feha2 d1czpa_
		Staphylokinase/streptokinase	3(4)	d1l4db_
		Superantigen toxins, C-terminal domain	9	d1enfa2

Alpha/alpha toroids (Cluster 228)				
family merge; superfamily break; fold split;				
Class	Fold	Superfamily	Ndom	Representative SCOPid
A	alpha/alpha toroid	Six-hairpin glycosidases	16	d1ayx__ d1lf6a1 d1h54a1 d1fp3a_ d1nc5a_ d1vd5a_ d1g9ga_
		Seven-hairpin glycosidases	4	d1dl2a_
		Chondroitin AC/alginate lyase	6	d1cb8a1 d1qaza_
		Terpenoid cyclases/Protein prenyltransferases	9(11)	d2sqca2 d1c3d__ d1dceb_
		Family 10 polysaccharide lyase	2	d1gxma_

Cluster 256				
family merge; superfamily break; fold split;				
YhbC-like is interesting that it is essential protein but its function is not known (Yu et al, 2001).				
Class	Fold	Superfamily	Ndom	Representative SCOPid
B	Sm-like fold	Sm-like ribonucleoproteins	10(11)	d1b34a_ d1hk9a_
		YhbC-like, C-terminal domain	1	d1ib8a1

Barrel-sandwich hybrids (Cluster 332)
family merge; superfamily break; fold split;
Note: Single hybrid motif and Rudiment single hybrid motif share significant structural similarity and moderate to high sequence similarity. Evolutionary trace can also be shown by tightly aligned turns. Duplicated hybrid motif is duplicated and dimerized single hybrid motif as the name implies. Rudiment single hybrid motif is splitted into two clusters (this cluster and cluster 283). This cluster and cluster 283 might be homologous.

Class	Fold	Superfamily	Ndom	Representative SCOPid
B	Barrel-sandwich hybrid	Single hybrid motif	8	d1bdo__
		Rudiment single hybrid motif	1(7)	d1e2wa2
		Duplicated hybrid motif	2	d1gpr__ d1qwya_

Gelsolin-like proteins (Cluster 552)				
family merge; superfamily merge; fold clean;				
Note: Moderately high sequence similarity (HHsearch 0.7) and high structural similarity (DaliZ 11) d1m2oa4 is very similar to d1svy_. d1svy_ is a hybrid of sec23/24 and actin depolymerizing protein.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a+b	Gelsolin-like	Actin depolymerizing proteins	14	d1cfya_ d1p8xa1
		C-terminal, gelsolin-like domain of Sec23/24	2	d1m2oa4

Immunoglobulin-like proteins (Cluster 662)				
family break; superfamily break; fold split;				
Note: Very high sequence similarity (HHsearch 0.97) and structural score (DaliZ 11.7). It is beyond reasonable doubt that all proteins in this cluster are homologous. The problem is in breaking (super)family relationship in SCOP database. This needs further investigation.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
B	Immunoglobulin-like beta-sandwich	Invasin/intimin cell-adhesion fragments	4(6)	d1cwva1
		E set domains	3(55)	d1qfha1

Beta-clip (Cluster 1207)				
family merge; superfamily merge; fold split;				
Note: Similar in structure (DaliZ 6.2). Share similarity in trimerization. SCOP also notes the similarity, too.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
B	beta-clip	dUTPase-like	6	d1euwa_
		Tlp20, baculovirus telokin-like protein	1	d1tul__

HTH proteins (Cluster 1528)				
family break; superfamily break; fold split;				
Note: all HTH's are probably homologous (Aravind et al, 2005).				
Class	Fold	Superfamily	Ndom	Representative SCOPid
A	DNA/RNA-binding 3-helical bundle	Sigma3 and sigma4 domains of RNA polymerase sigma factors	4(7)	d1xsva_ d1ku3a_
		C-terminal effector domain of the bipartite response regulators	4(9)	d1fsea_

Acyl-CoA binding protein & FERM second domain (Cluster 1721)
family merge; superfamily merge; fold clean;
Note: High sequence/structural similarity; In 2nd domain of FERM, short conserved helix and loop occupies the cleft between helix2 and helix 3. This cleft is the active site in Acyl-CoA binding proteins.

Class	Fold	Superfamily	Ndom	Representative SCOPid
A	Acyl-CoA binding protein-like	Acyl-CoA binding protein	2	d1hbka_
		Second domain of FERM	3	d1gg3a1

Table 2 Potentially homologous clusters containing different SCOP superfamilies. Each sub-table contains cluster title featuring the majority of proteins in the cluster with a light blue shadowed row. The title row is followed by a terse comparison to SCOP. This comparison is done at SCOP family, superfamily and fold level and given classification of "clean", "merge", "split", and "break". "clean" means the cluster contains exactly same as SCOP at the given level, i.e. "family clean" means that the cluster contains all the members in SCOP family. This label of "clean" usually means nice agreement on SCOP and iSCG clustering especially for the superfamily level. "merge" means that the cluster contains more than one SCOP groups without contradicting to SCOP, i.e. cluster 1721 in the table has the "superfamily merge" label since this cluster contains all members in the two superfamilies. Thus "merge" might mean over-clustering or interesting new finding with nice agreement with SCOP. "split" means the cluster contains subset of one SCOP group, i.e. cluster 1528 has "fold split" because this cluster subset (under-cluster) compared to the DNA/RNA binding 3-helical bundle fold in SCOP. "break" label is very interesting case where the cluster does not agree on SCOP boundary of the groups, i.e. cluster 1528 has "superfamily break" because this cluster contains subset of two superfamilies. Usually, this "break" label requires special attention to understand why the disagreement on the group boundary happened. The third row is for noting short information for the cluster, i.e. previous literatures about the proteins in the cluster or other argument for homology, and other interesting observations about the contents of the cluster. From the fourth row, the table contains summary for each cluster. The "class", "fold" and "superfamily" columns represent respective SCOP information. "Ndom" represents how many proteins in the cluster belongs the SCOP superfamily. If the cluster does not contain all the SCOP superfamily members , the total number of the superfamily in SCOP is shown in the parenthesis. "representative SCOP id" column shows the list of family representatives in the cluster.

Protein Kinase, SAICAR synthase, and ATP-grasp (Cluster 7)				
family merge; superfamily break; fold break;				
Note: Protein kinase and others in the cluster are probably homologous (Grishin, 1999). They share architecture of secondary structural elements remotely and have rather conserved active site structure.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a+b	ATP-grasp	Glutathione synthetase ATP-binding domain-like	17(19)	d1a9xa5 d1ehia2 d1uc8a2 d1i7na2 d2scub2 d1kbla3
	SAICAR synthase-like	SAICAR synthase-like	4	d1bo1a_ d1w2fa_ d1kuta_
	Protein kinase-like (PK-like)	Protein kinase-like (PK-like)	43	d1ia9a_ d1tqia2 d1j7la_ d1nw1a_ d1a06__ d1cjaa_ d1e7ua4

Beta-propellers (Cluster 15)				
family merge; superfamily merge; fold break;				
Note: Most beta-propellers are probably homologous (Chaudhuri et al, 2008).				
Class	Fold	Superfamily	Ndom	Representative

				SCOPid
b	4-bladed beta-propeller	Hemopexin-like domain	5	d1fbl_1
	5-bladed beta-propeller	Tachylectin-2	1	d1tl2a_
		Arabinanase/levansucrase/invertase	5	d1gyha_ d1uypa2 d1vkda_ d1oyga_
		Apyrase	1	d1s1da_
	6-bladed beta-propeller	Sialidases	11	d1f8ea_ d1v0ea1
		Kelch motif	1	d1u6dx_
		Soluble quinoprotein glucose dehydrogenase	1	d1crua_
		Thermostable phytase (3-phytase)	1	d1h6la_
		TolB, C-terminal domain	1	d1crza1
		YWTD domain	2	d1ijqa1
		Calcium-dependent phosphotriesterase	2	d1v04a_ d1pjxa_
		Tricorn protease N-terminal domain	1	d1k32a2
		Fucose-specific lectin	1	d1ofza_
		NHL repeat	2	d1q7fa_
	7-bladed beta-propeller	Galactose oxidase, central domain	1	d1k3ia3
		3-carboxy-cis,cis-mucoante lactonizing enzyme	1	d1jofa_
		Putative isomerase YbhE	1	d1ri6a_
		Sema domain	3	d1olza2
		Oligoxyloglucan reducing end-specific cellobiohydrolase	2	d1sqja1
		Nucleoporin domain	2	d1xipa_
		YVTN repeat-like/Quinoprotein amine dehydrogenase	5	d1l0qa2 d1jmxb_ d2bbkh_
		Nitrous oxide reductase, N-terminal domain	1	d1fwxa2
		WD40 repeat-like	12	d1sq9a_ d1yfqa_
		RCC1/BLIP-II	2	d1a12a_ d1jtdb_
		Clathrin heavy-chain terminal domain	1	d1utca2
		Peptidase/esterase 'gauge' domain	2	d1ve6a1 d1qfma1
		Integrin alpha N-terminal domain	1	d1txva_
		Tricorn protease domain 2	1	d1k32a3
	8-bladed beta-propeller	Quinoprotein alcohol dehydrogenase-like	3	d1flga_
		C-terminal (heme d1) domain of cytochrome cd1-nitrite reductase	1	d1qksa2
		DPP6 N-terminal domain-like	2	d1orva1

DNA polymerse III χ subunit & P-loop hydrolase (Cluster 17)				
family merge; superfamily break; fold break;				
Note: Probably homologous. Very high sequence/structural similarity. SCOP also mentioned this relationship.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a/b	DNA polymerase III chi subunit	DNA polymerase III chi subunit	1	d1em8a_
	P-loop containing nucleoside	P-loop containing nucleoside	30(190)	d1w36d2 d1a1va2

	triphosphate hydrolases	triphosphate hydrolases		d1gkub2 d1rifa_

Rossmann-fold (Cluster 39)				
family merge; superfamily break; fold break;				
Note: The two superfamilies share much conserved structure and sequence. (Dali Z-score:10.3, HHsearch: 0.93) And they share active site region. However the question remains that the this cluster might be over-split and this cluster should include many more Rossmann-fold proteins.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a/b	NAD(P)-binding Rossmann-fold domains	NAD(P)-binding Rossmann-fold domains	55(161)	d1b8pa1 d1lj8a4 d1a4ia1 d1vlla_ d1gdha1
	ATC-like	Aspartate/ornithine carbamoyltransferase	9	d1js1x2

Metallo-dependent hydrolase & 7 stranded β/α barrel (Cluster 41)				
family merge; superfamily merge; fold break;				
Note: The two structure share high structural similarity and metal binding sites.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a/b	TIM beta/alpha-barrel	Metallo-dependent hydrolases	18	d1itua_ d4ubpc2 d1o12a2 d1bf6a_ d1j6oa_ d1onwa2 d1gkpa2 d1j79a_ d1m7ja3 d1a4ma_ d1p1ma2 d1ra0a2 d1j5sa_
	7-stranded beta/alpha barrel	PHP domain-like	2	d1m65a_ d1v77a_

FAD/NAD(P)-binding domain & Nucleotide-binding domain (Cluster 260)				
family break; superfamily break; fold break;				
(Cheng et al, 2008)				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a/b	FAD/NAD(P)-binding domain	FAD/NAD(P)-binding domain	37(62)	d1ng4a1 d1chua2 d1d5ta1 d1fcda1
	Nucleotide-binding domain	Nucleotide-binding domain	9	d1c0pa1 d1i8ta1 d1cjca2

ATPase domain & Sporulation response regularoty protein (Cluster 275)				
family merge; superfamily merge; fold merge;				
Note: SCOP mentioned that spo0B has histidine kinase fold lacking the ATP-binding site (DALI 8.8 and HH 0.99).				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a+b	ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase	ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase	14	d1th8a_ d1jm6a2 d1b63a2

				d1uyla_
	Sporulation response regulatory protein Spo0B	Sporulation response regulatory protein Spo0B	1	d1ixma_

DnaG C-terminal domain & DnaB N-terminal domain (Cluster 290)				
family merge; superfamily merge; fold merge;				
Note: extremely high sequence/structural similarity. It should be noted as a SCOP error.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
A	DNA primase DnaG, C-terminal domain	DNA primase DnaG, C-terminal domain	1	d1t3wa_
	N-terminal domain of DnaB helicase	N-terminal domain of DnaB helicase	1	d1b79a_

Double psi beta-barrels (Cluster 692)				
family merge; superfamily break; fold break;				
(Coles et al, 2006)				
Class	Fold	Superfamily	Ndom	Representative SCOPid
B	Reductase/isomerase/elongation factor common domain	Translation proteins	13(14)	d1sqra_ d1wb1a2
	Elongation factor/aminomethyltransferase common domain	EF-Tu/eEF-1alpha/eIF2-gamma C-terminal domain	5	d1r5ba2
	Domain of alpha and beta subunits of F1 ATP synthase-like	N-terminal domain of alpha and beta subunits of F1 ATP synthase	2	d1w0ja2

Cluster 735				
family break; superfamily break; fold break;				
Note: The two superfamilies have high structural and sequence similarity (DALI z-score: 6.6, HHsearch: 0.8). Structural alignment and sequence alignments agree with each other in most parts with quite several identical residue pairs. Their fold is kind of peculiar because there are three consecutive helices on one side of the beta-sheet. But 1wfz binds a metal while 1t3q does not.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a+b	SufE/NifU	SufE/NifU	4	d1mzga_ d1su0b_
	CO dehydrogenase flavoprotein C-domain-like	FAD/NAD-linked reductases, dimerisation (C-terminal) domain	12(13)	d1xhca3
		CO dehydrogenase flavoprotein C-terminal domain-like	5	d1jroa3

Calpain middle domain, Galactose binding domain, & Collagen binding domain (Cluster 784)				
family break; superfamily break; fold break;				
Note: Calpain and Galactose binding domain have very high sequence similarity and structural similarity. CUB-like is rather distant from them. CUB-like have moderate structural similarity to calpain. This cluster should be clustered with cluster 30. So this cluster has over-split problem.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
B	Calpain large subunit, middle domain (domain III)	Calpain large subunit, middle domain (domain III)	1	d1df0a2
	Galactose-binding domain-like	Galactose-binding domain-like	1(47)	d1wmda1
	CUB-like	Collagen-binding domain	1	d1nqja_

Cluster 850				
family merge; superfamily break; fold break;				
(Aravind et al, 2002)				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a/b	Adenine nucleotide alpha hydrolase-like	Adenine nucleotide alpha hydrolases-like	5(20)	d1jmva_
	Cryptochrome/photolyase, N-terminal domain	Cryptochrome/photolyase, N-terminal domain	5	d1dnpa2

Cluster 3730				
family merge; superfamily merge; fold merge;				
(Kinch et al, 2005)				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a/b	DAK1/DegV-like	DAK1/DegV-like	4	d1mgpa_ d1oi2a_
	IIA domain of mannose transporter, IIA-Man	IIA domain of mannose transporter, IIA-Man	1	d1pdo__

Cluster 4431				
family merge; superfamily merge; fold merge;				
Note: High structural similarity. Long insertion and extension are in gp5 protein.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a+b	Major capsid protein gp5	Major capsid protein gp5	1	d1ohga_
	Hypothetical protein PF0899	Hypothetical protein PF0899	1	d1shea_

Table 3 Potential homolog clusters containing different folds. The legends are same as Table 2.

Barrel-sandwich hybrid & Hammerheads (Cluster 283)				
family merge; superfamily break; fold break;				
Sequence and structural motif of hammerhead				
Class	Fold	Superfamily	Ndom	Representative SCOPid
B	Barrel-sandwich hybrid	Rudiment single hybrid motif	6(7)	d1b6ra1
a+b	alpha/beta-Hammerhead	CO dehydrogenase molybdoprotein N-domain-like	6	d1jrob1
		Nicotinate/Quinolinate PRTase N-terminal domain-like	5	d1vlpa1 d1o4ua2
		Pyrimidine nucleoside phosphorylase C-terminal domain	3	d1brwa3
		Ribosomal protein L16p/L10e	2	d1jj2h_ d1wkia_

HTH (Cluster 398)				
family merge; superfamily break; fold break;				
(Aravind et al, 2005)				
Class	Fold	Superfamily	Ndom	Representative SCOPid
A	Putative DNA-binding domain	Putative DNA-binding domain	2(14)	d1l8ra_
a+b	DNA-binding domain of Mlu1-box binding protein MBP1	DNA-binding domain of Mlu1-box binding protein MBP1	1	d1bm8__

HTH (Cluster 2756)				
family merge; superfamily break; fold break;				
(Aravind et al, 2005)				
Class	Fold	Superfamily	Ndom	Representative SCOPid
A	Putative DNA-binding domain	Putative DNA-binding domain	2(14)	d1jjcb1
a+b	SRP19	SRP19	3	d1jida_

Table 4 Potential homolog clusters containing different class. The legends are same as Table 2.

4.2.1.1 Flavodoxin-like fold proteins (Cluster 6)

Among Flavodoxin-like fold proteins defined in SCOP, three different superfamily proteins were clustered in this cluster; CheY-like superfamily, cobalamin

binding-domain superfamily and succinyl-CoA synthetase domain superfamily (Cluster 6 in Table 2). CheY-like superfamily proteins are response regulator in bacterial two-component regulatory systems. Cobalamin binding-domains (Bandarian et al, 2002) bind to the cofactor cobalamin and provide this cofactor to other enzymatic domains. Succinyl-CoA synthetase domains (Fraser et al, 1999) carry out phosphorylation in citric acid cycle. Even though the three superfamily do not share the key functional and conserved residues, they share structural similarity and general sequence conservation pattern much higher than unrelated proteins in the fold, DaliLite Z-score above 9 and COMPASS E-value 1.0e-9. Along with this general sequence/structural similarity, they share the active site region. So it is likely scenario that they share the ancestor and diverged to have their respective functions.

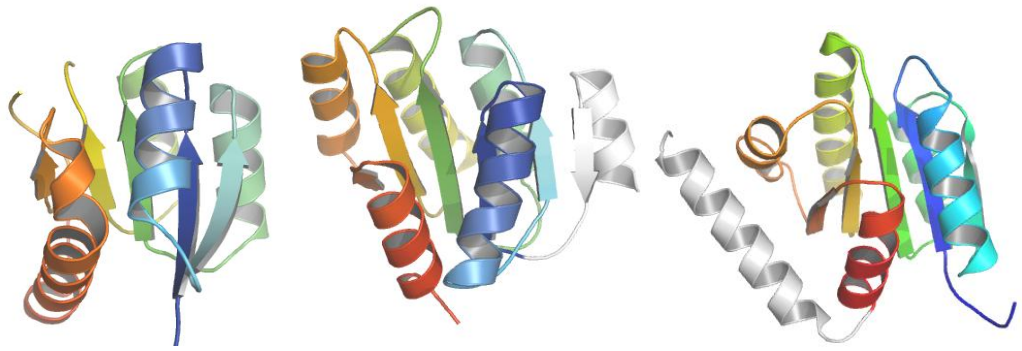


Figure 28 Representatives of Flavodoxin-like fold protein structures. The left structure (SCOPid: d1r8ja2) is the representative from CheY-like superfamily. The middle structure (SCOPid: d1eucb1) is the representative from succinyl-CoA synthetase superfamily. The right structure (SCOPid: d1bmta2) is the representative from cobalamin binding-domain superfamily. All structures are colored from blue to red starting from N-terminus to C-terminus of proteins. Relative insertions are colored in white. All structures share the common core of parallel beta-sheets (5-4-3-1-2) connected by alpha helices. The middle structure succinyl-CoA synthetase has extended in N-terminus with a beta-strand and an alpha-helix. The right structure

cobalamin binding domain has an extra alpha-helix at the C-terminal end. All extended structures are colored in white.

4.2.1.2 Cluster 43: TIM barrel fold proteins

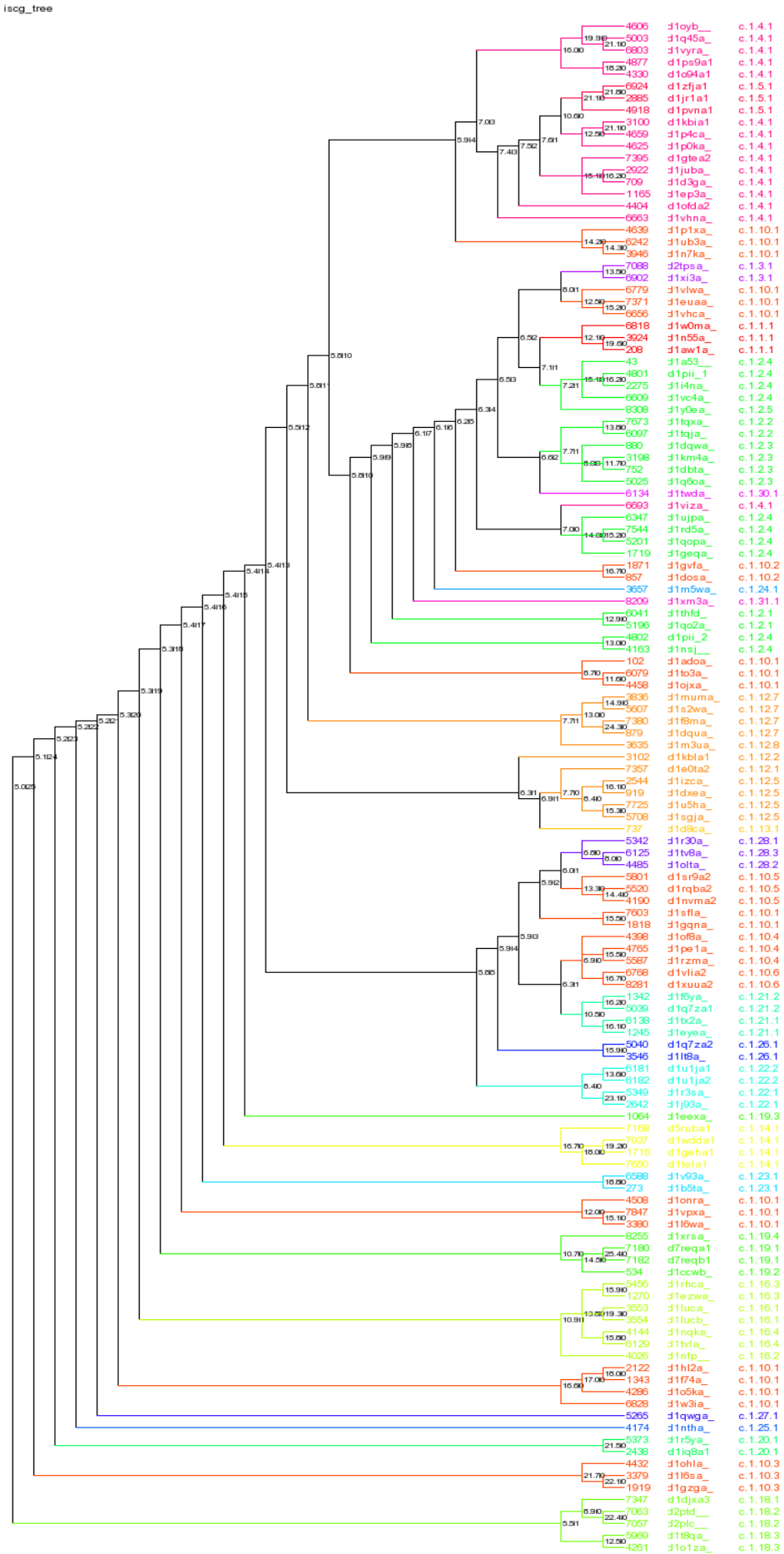
TIM beta/alpha barrel fold is one of the most versatile folds to catalyze many different kinds of chemical reactions. Most TIM barrel enzymes have active site at the C-terminal end of barrel which might be an indication of their common evolutionary origin (Nagano et al, 2002).

Class I aldolases are grouped into 6 different subgroups in iSCG and widely distributed among other phosphate binding enzymes (Figure 29). Aldolase catalyzes the fusion or cleavage of two carbonyl compounds. Class I aldolases have catalytic lysine forming Schiff base with substrate and phosphate binding sites. According to SCOP database all class I aldolases are classified into one family. Sub-classification of class I aldolases with other non-aldolase enzymes in iSCG (Figure 29) is contradicting to generally accepted hypothesis of their divergent evolution from common ancestor based on enzymatic function and active site residues.

Current version of CATH database (ver. 3.1.0) shows similar result in classifying phosphate-binding TIM barrel enzyme. CATH classified the aldolases and other phosphate-binding enzymes in the same homologous superfamily. Also, the aldolases are sub-divided into smaller groups with other enzymes at the family level. PSI-BLAST (query: KDPG aldolase, 1eua) found Ribulose-phosphate binding enzymes (located in the same subtree in Figure 29) with significant E-value (1e-32 at 5th iteration) before finding

other class I aldolase subgroups. Other PSI-BLAST searched based on queries in different subclass of aldolases shows similar result in finding enzymes in the same subtree non-aldolase TIM barrel enzymes before finding other different subgroup aldolases. Additionally, each subgroups of aldolase shows distinctive structural features outside the active site, i.e. extensions in N- or C-termini.

This seemingly contradiction between functional classification and similarity based classification might mean polyphyletic origin of class I aldolases. Further research will be conducted to understand the evolutionary origin of proteins in TIM barrel fold.


iscg_tree

d1oyb__	c.1.4.1	TIM beta/alpha-barrel	FMN-linked oxidoreductases	FMN-linked oxidoreductases	Old yellow enzyme (OYE)
d1q45a_	c.1.4.1	TIM beta/alpha-barrel	FMN-linked oxidoreductases	FMN-linked oxidoreductases	12-oxophytodienoate reductase (OPR, OYE~
d1vyra_	c.1.4.1	TIM beta/alpha-barrel	FMN-linked oxidoreductases	FMN-linked oxidoreductases	Pentaerythritol tetranirate reductase
d1ps9a1	c.1.4.1	TIM beta/alpha-barrel	FMN-linked oxidoreductases	FMN-linked oxidoreductases	2,4-dienoyl-CoA reductase, N-terminal d~
d1o94a1	c.1.4.1	TIM beta/alpha-barrel	FMN-linked oxidoreductases	FMN-linked oxidoreductases	Trimethylamine dehydrogenase, N-termina~
d1zfja1	c.1.5.1	TIM beta/alpha-barrel	Inosine monophosphate dehydro~	Inosine monophosphate dehydro~	Inosine monophosphate dehydrogenase (IM~
d1jr1a1	c.1.5.1	TIM beta/alpha-barrel	Inosine monophosphate dehydro~	Inosine monophosphate dehydro~	Inosine monophosphate dehydrogenase (IM~
d1pvna1	c.1.5.1	TIM beta/alpha-barrel	Inosine monophosphate dehydro~	Inosine monophosphate dehydro~	Inosine monophosphate dehydrogenase (IM~
d1kbia1	c.1.4.1	TIM beta/alpha-barrel	FMN-linked oxidoreductases	FMN-linked oxidoreductases	Flavocytochrome b2, C-terminal domain
d1p4ca_	c.1.4.1	TIM beta/alpha-barrel	FMN-linked oxidoreductases	FMN-linked oxidoreductases	Membrane-associated (S)-mandelate dehyd~
d1p0ka_	c.1.4.1	TIM beta/alpha-barrel	FMN-linked oxidoreductases	FMN-linked oxidoreductases	Isopentenyl-diphosphate delta-isomerase
d1gtea2	c.1.4.1	TIM beta/alpha-barrel	FMN-linked oxidoreductases	FMN-linked oxidoreductases	Dihydropyrimidine dehydrogenase, domain~
d1juba_	c.1.4.1	TIM beta/alpha-barrel	FMN-linked oxidoreductases	FMN-linked oxidoreductases	Dihydroorotate dehydrogenase
d1d3ga_	c.1.4.1	TIM beta/alpha-barrel	FMN-linked oxidoreductases	FMN-linked oxidoreductases	Dihydroorotate dehydrogenase
d1ep3a_	c.1.4.1	TIM beta/alpha-barrel	FMN-linked oxidoreductases	FMN-linked oxidoreductases	Dihydroorotate dehydrogenase
d1ofda2	c.1.4.1	TIM beta/alpha-barrel	FMN-linked oxidoreductases	FMN-linked oxidoreductases	Alpha subunit of glutamate synthase, ce~
d1vhna_	c.1.4.1	TIM beta/alpha-barrel	FMN-linked oxidoreductases	FMN-linked oxidoreductases	Putative flavin oxidoreducatase TM0096
d1p1xa_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	Deoxyribose-phosphate aldolase DeoC
d1ub3a_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	Deoxyribose-phosphate aldolase DeoC
d1n7ka_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	Deoxyribose-phosphate aldolase DeoC
d2tpsa_	c.1.3.1	TIM beta/alpha-barrel	Thiamin phosphate synthase	Thiamin phosphate synthase	Thiamin phosphate synthase
d1xi3a_	c.1.3.1	TIM beta/alpha-barrel	Thiamin phosphate synthase	Thiamin phosphate synthase	Thiamin phosphate synthase
d1vlwa_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	KDPG aldolase
d1euaa_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	KDPG aldolase
d1vhca_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	Hypothetical protein HI0047
d1w0ma_	c.1.1.1	TIM beta/alpha-barrel	Triosephosphate isomerase (TI~	Triosephosphate isomerase (TI~	Triosephosphate isomerase
d1n55a_	c.1.1.1	TIM beta/alpha-barrel	Triosephosphate isomerase (TI~	Triosephosphate isomerase (TI~	Triosephosphate isomerase
d1aw1a_	c.1.1.1	TIM beta/alpha-barrel	Triosephosphate isomerase (TI~	Triosephosphate isomerase (TI~	Triosephosphate isomerase
d1a53__	c.1.2.4	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	Tryptophan biosynthesis enzym~	Indole-3-glycerophosphate synthase, IPGS
d1pii_1	c.1.2.4	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	Tryptophan biosynthesis enzym~	Indole-3-glycerophosphate synthase, IPGS
d1i4na_	c.1.2.4	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	Tryptophan biosynthesis enzym~	Indole-3-glycerophosphate synthase, IPGS
d1vc4a_	c.1.2.4	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	Tryptophan biosynthesis enzym~	Indole-3-glycerophosphate synthase, IPGS
d1y0ea_	c.1.2.5	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	NanE-like	Putative N-acetylmannosamine-6-phosphat~
d1tqxa_	c.1.2.2	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	D-ribulose-5-phosphate 3-epim~	D-ribulose-5-phosphate 3-epimerase
d1tqja_	c.1.2.2	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	D-ribulose-5-phosphate 3-epim~	D-ribulose-5-phosphate 3-epimerase
d1dqwa_	c.1.2.3	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	Decarboxylase	Orotidine 5'-monophosphate decarboxylas~
d1km4a_	c.1.2.3	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	Decarboxylase	Orotidine 5'-monophosphate decarboxylas~
d1dbta_	c.1.2.3	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	Decarboxylase	Orotidine 5'-monophosphate decarboxylas~
d1q6oa_	c.1.2.3	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	Decarboxylase	3-keto-L-gulonate 6-phosphate decarboxy~
d1twda_	c.1.30.1	TIM beta/alpha-barrel	CutC-like	CutC-like	Copper homeostasis protein CutC
d1viza_	c.1.4.1	TIM beta/alpha-barrel	FMN-linked oxidoreductases	FMN-linked oxidoreductases	PcrB protein homolog YerE
d1ujpa_	c.1.2.4	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	Tryptophan biosynthesis enzym~	Trp synthase alpha-subunit
d1rd5a_	c.1.2.4	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	Tryptophan biosynthesis enzym~	Trp synthase alpha-subunit
d1qopa_	c.1.2.4	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	Tryptophan biosynthesis enzym~	Trp synthase alpha-subunit
d1geqa_	c.1.2.4	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	Tryptophan biosynthesis enzym~	Trp synthase alpha-subunit
d1gvfa_	c.1.10.2	TIM beta/alpha-barrel	Aldolase	Class II FBP aldolase	Tagatose-1,6-bisphosphate aldolase
d1dosa_	c.1.10.2	TIM beta/alpha-barrel	Aldolase	Class II FBP aldolase	Fructose-bisphosphate aldolase (FBP ald~
d1m5wa_	c.1.24.1	TIM beta/alpha-barrel	Pyridoxine 5'-phosphate synth~	Pyridoxine 5'-phosphate synth~	Pyridoxine 5'-phosphate synthase
d1xm3a_	c.1.31.1	TIM beta/alpha-barrel	ThiG-like	ThiG-like	Thiazole biosynthesis protein ThiG
d1thfd_	c.1.2.1	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	Histidine biosynthesis enzymes	Cyclase subunit (or domain) of imidazol~
d1qo2a_	c.1.2.1	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	Histidine biosynthesis enzymes	Phosphoribosylformimino-5-aminoimidazol~
d1pii_2	c.1.2.4	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	Tryptophan biosynthesis enzym~	N-(5'phosphoribosyl)anthranilate isomera~
d1nsj__	c.1.2.4	TIM beta/alpha-barrel	Ribulose-phoshate binding bar~	Tryptophan biosynthesis enzym~	N-(5'phosphoribosyl)anthranilate isomera~
d1adoa_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	Fructose-1,6-bisphosphate aldolase
d1to3a_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	Putative aldolase YihT
d1ojxa_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	Archaeal fructose 1,6-bisphosphate aldo~
d1muma_	c.1.12.7	TIM beta/alpha-barrel	Phosphoenolpyruvate/pyruvate ~	Phosphoenolpyruvate mutase/Is~	2-methylisocitrate lyase
d1s2wa_	c.1.12.7	TIM beta/alpha-barrel	Phosphoenolpyruvate/pyruvate ~	Phosphoenolpyruvate mutase/Is~	Phosphoenolpyruvate mutase
d1f8ma_	c.1.12.7	TIM beta/alpha-barrel	Phosphoenolpyruvate/pyruvate ~	Phosphoenolpyruvate mutase/Is~	Isocitrate lyase
d1dqua_	c.1.12.7	TIM beta/alpha-barrel	Phosphoenolpyruvate/pyruvate ~	Phosphoenolpyruvate mutase/Is~	Isocitrate lyase
d1m3ua_	c.1.12.8	TIM beta/alpha-barrel	Phosphoenolpyruvate/pyruvate ~	Ketopantoate hydroxymethyltra~	Ketopantoate hydroxymethyltransferase P~
d1kbla1	c.1.12.2	TIM beta/alpha-barrel	Phosphoenolpyruvate/pyruvate ~	Pyruvate phosphate dikinase, ~	Pyruvate phosphate dikinase, C-terminal~
d1e0ta2	c.1.12.1	TIM beta/alpha-barrel	Phosphoenolpyruvate/pyruvate ~	Pyruvate kinase	Pyruvate kinase, N-terminal domain
d1izca_	c.1.12.5	TIM beta/alpha-barrel	Phosphoenolpyruvate/pyruvate ~	HpcH/HpaI aldolase	Macrophomate synthase
d1dxea_	c.1.12.5	TIM beta/alpha-barrel	Phosphoenolpyruvate/pyruvate ~	HpcH/HpaI aldolase	2-dehydro-3-deoxy-galactarate aldolase
d1u5ha_	c.1.12.5	TIM beta/alpha-barrel	Phosphoenolpyruvate/pyruvate ~	HpcH/HpaI aldolase	Citrate lyase, beta subunit
d1sgja_	c.1.12.5	TIM beta/alpha-barrel	Phosphoenolpyruvate/pyruvate ~	HpcH/HpaI aldolase	Citrate lyase, beta subunit
d1d8ca_	c.1.13.1	TIM beta/alpha-barrel	Malate synthase G	Malate synthase G	Malate synthase G
d1r30a_	c.1.28.1	TIM beta/alpha-barrel	Radical SAM enzymes	Biotin synthase	Biotin synthase
d1tv8a_	c.1.28.3	TIM beta/alpha-barrel	Radical SAM enzymes	MoCo biosynthesis proteins	Molybdenum cofactor biosynthesis protei~
d1olta_	c.1.28.2	TIM beta/alpha-barrel	Radical SAM enzymes	Oxygen-independent coproporph~	Oxygen-independent coproporphyrinogen I~
d1sr9a2	c.1.10.5	TIM beta/alpha-barrel	Aldolase	HMGL-like	2-isopropylmalate synthase LeuA, cataly~
d1rqba2	c.1.10.5	TIM beta/alpha-barrel	Aldolase	HMGL-like	Transcarboxylase 5S subunit, N-terminal~
d1nvma2	c.1.10.5	TIM beta/alpha-barrel	Aldolase	HMGL-like	4-hydroxy-2-oxovalerate aldolase DmpG, ~
d1sfla_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	Type I 3-dehydroquinate dehydratase
d1gqna_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	Type I 3-dehydroquinate dehydratase
d1of8a_	c.1.10.4	TIM beta/alpha-barrel	Aldolase	Class I DAHP synthetase	3-deoxy-D-arabino-heptulosonate-7-phosp~
d1pe1a_	c.1.10.4	TIM beta/alpha-barrel	Aldolase	Class I DAHP synthetase	3-deoxy-D-manno-octulosonate 8-phosphat~
d1rzma_	c.1.10.4	TIM beta/alpha-barrel	Aldolase	Class I DAHP synthetase	3-deoxy-D-arabino-heptulosonate-7-phosp~
d1vlia2	c.1.10.6	TIM beta/alpha-barrel	Aldolase	NeuB-like	Spore coat polysaccharide biosynthesis ~
d1xuua2	c.1.10.6	TIM beta/alpha-barrel	Aldolase	NeuB-like	Capsule biosynthesis protein SiaC, N-te~
d1ff6ya_	c.1.21.2	TIM beta/alpha-barrel	Dihydropteroate synthetase-li~	Methyltetrahydrofolate-utiluz~	Methyltetrahydrofolate: corrinoid/iron-~
d1q7zaa1	c.1.21.2	TIM beta/alpha-barrel	Dihydropteroate synthetase-li~	Methyltetrahydrofolate-utiluz~	Cobalamin-dependent methionine synthase~
d1tx2a_	c.1.21.1	TIM beta/alpha-barrel	Dihydropteroate synthetase-li~	Dihydropteroate synthetase	Dihydropteroate synthetase
d1eyea_	c.1.21.1	TIM beta/alpha-barrel	Dihydropteroate synthetase-li~	Dihydropteroate synthetase	Dihydropteroate synthetase
d1q7za2	c.1.26.1	TIM beta/alpha-barrel	Homocysteine S-methyltransfer~	Homocysteine S-methyltransfer~	Cobalamin-dependent methionine synthase~
d1lf8a_	c.1.26.1	TIM beta/alpha-barrel	Homocysteine S-methyltransfer~	Homocysteine S-methyltransfer~	Betaine-homocysteine S-methyltransferase
d1u1ja1	c.1.22.2	TIM beta/alpha-barrel	UROD/MetE-like	Cobalamin-independent methion~	5-methyltetrahydropteroyltriglutamate--~
d1u1ja2	c.1.22.2	TIM beta/alpha-barrel	UROD/MetE-like	Cobalamin-independent methion~	5-methyltetrahydropteroyltriglutamate--~
d1r3sa_	c.1.22.1	TIM beta/alpha-barrel	UROD/MetE-like	Uroporphyrinogen decarboxylas~	Uroporphyrinogen decarboxylase, UROD
d1j93a_	c.1.22.1	TIM beta/alpha-barrel	UROD/MetE-like	Uroporphyrinogen decarboxylas~	Uroporphyrinogen decarboxylase, UROD
d1eexa_	c.1.19.3	TIM beta/alpha-barrel	Cobalamin (vitamin B12)-depen~	Diol dehydratase, alpha subun~	Diol dehydratase, alpha subunit
d5rubal	c.1.14.1	TIM beta/alpha-barrel	RuBisCo, C-terminal domain	RuBisCo, large subunit, C-ter~	Ribulose 1,5-bisphosphate carboxylase-o~
d1wdda1	c.1.14.1	TIM beta/alpha-barrel	RuBisCo, C-terminal domain	RuBisCo, large subunit, C-ter~	Ribulose 1,5-bisphosphate carboxylase-o~
d1geha1	c.1.14.1	TIM beta/alpha-barrel	RuBisCo, C-terminal domain	RuBisCo, large subunit, C-ter~	Ribulose 1,5-bisphosphate carboxylase-o~
d1fxla1	c.1.14.1	TIM beta/alpha-barrel	RuBisCo, C-terminal domain	RuBisCo, large subunit, C-ter~	Ribulose 1,5-bisphosphate carboxylase-o~
d1v93a_	c.1.23.1	TIM beta/alpha-barrel	FAD-linked oxidoreductase	Methylenetetrahydrofolate red~	Methylenetetrahydrofolate reductase
d1b5ta_	c.1.23.1	TIM beta/alpha-barrel	FAD-linked oxidoreductase	Methylenetetrahydrofolate red~	Methylenetetrahydrofolate reductase
d1onra_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	Transaldolase
d1vpxa_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	Decameric fructose-6-phosphate aldolase~
d1fi6wa_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	Decameric fructose-6-phosphate aldolase~
d1xrsa_	c.1.19.4	TIM beta/alpha-barrel	Cobalamin (vitamin B12)-depen~	D-lysine 5,6-aminomutase alph~	D-lysine 5,6-aminomutase alpha subunit,~
d7reqa1	c.1.19.1	TIM beta/alpha-barrel	Cobalamin (vitamin B12)-depen~	Methylmalonyl-CoA mutase, N-t~	Methylmalonyl-CoA mutase alpha subunit,~
d7reqb1	c.1.19.1	TIM beta/alpha-barrel	Cobalamin (vitamin B12)-depen~	Methylmalonyl-CoA mutase, N-t~	Methylmalonyl-CoA mutase beta subunit, ~
d1ccwb_	c.1.19.2	TIM beta/alpha-barrel	Cobalamin (vitamin B12)-depen~	Glutamate mutase, large subun~	Glutamate mutase, large subunit
d1rhca_	c.1.16.3	TIM beta/alpha-barrel	Bacterial luciferase-like	F420 dependent oxidoreductases	Coenzyme F420 dependent secondary alcoh~
d1ezwa_	c.1.16.3	TIM beta/alpha-barrel	Bacterial luciferase-like	F420 dependent oxidoreductases	Coenzyme F420 dependent tetrahydrometha~
d1luca_	c.1.16.1	TIM beta/alpha-barrel	Bacterial luciferase-like	Bacterial luciferase (alkanal~	Bacterial luciferase alpha chain, LuxA
d1lucb_	c.1.16.1	TIM beta/alpha-barrel	Bacterial luciferase-like	Bacterial luciferase (alkanal~	Bacterial luciferase beta chain, LuxB
d1nqka_	c.1.16.4	TIM beta/alpha-barrel	Bacterial luciferase-like	Ssud-like monoxygenases	Alkanesulfonate monooxygenase SsuD
d1tvla_	c.1.16.4	TIM beta/alpha-barrel	Bacterial luciferase-like	Ssud-like monoxygenases	Putative monooxygenase YtnJ
d1nfp__	c.1.16.2	TIM beta/alpha-barrel	Bacterial luciferase-like	Non-fluorescent flavoprotein ~	Non-fluorescent flavoprotein (luxF, FP3~
d1hl2a_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	N-acetylneuraminate lyase
d1f74a_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	N-acetylneuraminate lyase
d1o5ka_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	Dihydrodipicolinate synthase
d1w3ia_	c.1.10.1	TIM beta/alpha-barrel	Aldolase	Class I aldolase	2-keto-3-deoxy gluconate aldolase Eda
d1qwga_	c.1.27.1	TIM beta/alpha-barrel	(2r)-phospho-3-sulfolactate s~	(2r)-phospho-3-sulfolactate s~	(2r)-phospho-3-sulfolactate synthase Co~
d1ntha_	c.1.25.1	TIM beta/alpha-barrel	Monomethylamine methyltransfe~	Monomethylamine methyltransfe~	Monomethylamine methyltransferase MtmB
d1r5ya_	c.1.20.1	TIM beta/alpha-barrel	tRNA-guanine transglycosylase	tRNA-guanine transglycosylase	Queosine tRNA-guanine transglycosylase
d1iq8a1	c.1.20.1	TIM beta/alpha-barrel	tRNA-guanine transglycosylase	tRNA-guanine transglycosylase	Archaeosine tRNA-guanine transglycosyla~
d1ohla_	c.1.10.3	TIM beta/alpha-barrel	Aldolase	5-aminolaevulinate dehydratas~	5-aminolaevulinate dehydratase, ALAD (p~
d1l6sa_	c.1.10.3	TIM beta/alpha-barrel	Aldolase	5-aminolaevulinate dehydratas~	5-aminolaevulinate dehydratase, ALAD (p~
d1gzga_	c.1.10.3	TIM beta/alpha-barrel	Aldolase	5-aminolaevulinate dehydratas~	5-aminolaevulinate dehydratase, ALAD (p~
d1djxa3	c.1.18.1	TIM beta/alpha-barrel	PLC-like phosphodiesterases	Mammalian PLC	Phospholipase C isozyme D1 (PLC-D1)
d2ptd__	c.1.18.2	TIM beta/alpha-barrel	PLC-like phosphodiesterases	Bacterial PLC	Phosphatidylinositol-specific phospholi~
d2plc__	c.1.18.2	TIM beta/alpha-barrel	PLC-like phosphodiesterases	Bacterial PLC	Phosphatidylinositol-specific phospholi~
d1f8qa_	c.1.18.3	TIM beta/alpha-barrel	PLC-like phosphodiesterases	Glycerophosphoryl diester pho~	Glycerophosphodiester phosphodiesterase~
d1o1za_	c.1.18.3	TIM beta/alpha-barrel	PLC-like phosphodiesterases	Glycerophosphoryl diester pho~	Hypothetical protein TM1621

Figure 29 Hierarchical tree for cluster 43. iSCG result is shown as hierarchical tree. Each superfamily is represented by distinct colors. The labels in the internal nodes are minimum score between the children nodes with iteration number. First column in terminal node label is the unique identifier for each protein for this study. Second column is SCOP id. Third column is short SCOP family ID. The second part of the figure displays same order of SCOP domains as the first part in six columns: SCOPid, SCOP family ID, fold name, superfamily name. Finally, sixth column is family name.

4.2.1.3 Cluster 15: Beta propeller folds

Cluster 15 contains 31 beta propeller superfamilies defined in SCOP database. Beta propellers consist of 4 to 8 anti-parallel beta-sheets. The beta-sheets form toroid around a central axis of pseudo-symmetry. Each beta sheet is called blade from the analogy of the architecture to propellers in fans. Beta propellers are one of most versatile proteins that they bind from macro-molecules such as proteins or DNA/RNA to very small inorganic molecules. Many beta propeller proteins also carry out enzymatic reactions (Table 5).

Superfamily	NB	Active Site	Ligand	F	Note
Hemopexin-like domain	4	Between two domains	Heme	B	Homo-dimer forms active site.
Tachylectin-2	5	Between every blades	Sugar	B	
Arabinanase/levansucrase/invertase	5	Canonical	Sugar	E	
Apyrase	5	Canonical	Sugar	E	Cellular function in preventing blood clotting
Sialidases	6	Canonical	Sugar	E	
Soluble quinoprotein glucose dehydrogenase	6	Canonical	Sugar	E	
Thermostable phytase (3-phytase)	6	Canonical	Small	E	
TolB, C-terminal domain	6	?	?	?	Function in Gram negative Bacterial cell envelop integrity
YWTD domain	6	Canonical	Protein	B	
Calcium-dependent phosphotriesterase	6	Canonical	Small	E	

Tricorn protease N-terminal domain	6	Canonical	Peptide ?	B?	Same chain with other 7 blade Topologically unclosed!
Fucose-specific lectin	6	Between blades except 1&6	Sugar	B	5 sugars binding. Somewhat controversy to previous 1:1 stoicheometry
NHL repeat	6	Canonical	Small	B	Sensor connected to receptor
Kelch motif	6	Canonical	Protein	B	E3 ligase complex substrate binding,
Galactose oxidase, central domain	7	Canonical	Sugar	E	
YVTN repeat-like/Quinoprotein amine dehydrogenase	7	Canonical	Protein	B/E	
Nitrous oxide reductase, N-terminal domain	7	Canonical	Small	E	
WD40 repeat-like	7	Canonical	Protein	B	
RCC1/BLIP-II	7	Bottom	Protein	B	RCC1 binds DNA at canonical surface
Clathrin heavy-chain terminal domain	7	Canonical	Peptide	B	
		Between blade 1 and 2	Peptide	B	
Peptidase/esterase 'gauge' domain	7	Canonical	Peptide	B	Topologically unclosed for prolylpeptidase.
Integrin alpha N-terminal domain	7	?	?	B?	
Tricorn protease domain 2	7	Canonical	?	?	Topologically unclosed!
3-carboxy-cis,cis-mucoante lactonizing enzyme	7	Canonical	Small	E	
Putative isomerase YbhE	7	?	?	?	
Sema domain	7	Canonical	Protein	B	Do not show clear repeat of blades
Oligoxyloglucan reducing end-specific cellobiohydrolase	7	Canonical	Sugar	E	Bacterial cell wall rigidity, Duplicated domains bind with 90 degree form active site
Nucleoporin domain	7	?	Protein /RNA	B	
Quinoprotein alcohol dehydrogenase-like	8	Canonical	Small	E	
C-terminal (heme d1) domain of cytochrome cd1-nitrite reductase	8	Canonical	Heme	B?	
DPP6 N-terminal domain-like	8	Canonical	Peptide	B	Topologically unclosed, Glycosylation, Blade VII was

Table 5. Summary of superfamilies in Cluster 15. Superfamilies are as defined in SCOP database. NB denotes number of blades. Active site column describes the general location of active sites

of the domains in the superfamily. Canonical means the pseudo-symmetric axis of top side. Bottom means the opposite side of top. Ligand column describes the general types of ligand for the superfamily. Small means general small inorganic or organic molecules except heme, sugar, DNA, RNA, or proteins. F denotes crude functional category. B and E represent binding function and enzymatic functions respectively. Question marks are used for unknown or putative information.

Cluster 15 in iSCG classification contains all known beta-propeller families in SCOP except one superfamily from 6-bladed beta-propellers. Although many literatures considered the proteins from each number of blades are closer homologs than proteins from the other number of blades, i.e. 7 blades beta-propellers are considered closer homologs than 8 blades beta-propellers. This view is now changing with known examples from Kelch motifs that form 6 or 7 blades and other examples. Our classification of beta-propellers is in agreement with this recent literature (Chaudhuri et al, 2008). Although 5-8 blades beta-propellers clearly sharing their ancestors, 4 blades beta-propellers are quite distinct and might not share the same evolutionary origin

One interesting feature among the beta-propellers is the “molecular velcro”. DPP6, POP and Tricorn (6 blades and 7 blades) (Engel et al, 2003) do not form molecular velcros in contrast to the rest of beta-propellers. It was speculated that the absence of molecular Velcro might mean those proteins are more flexible than other beta-propellers.

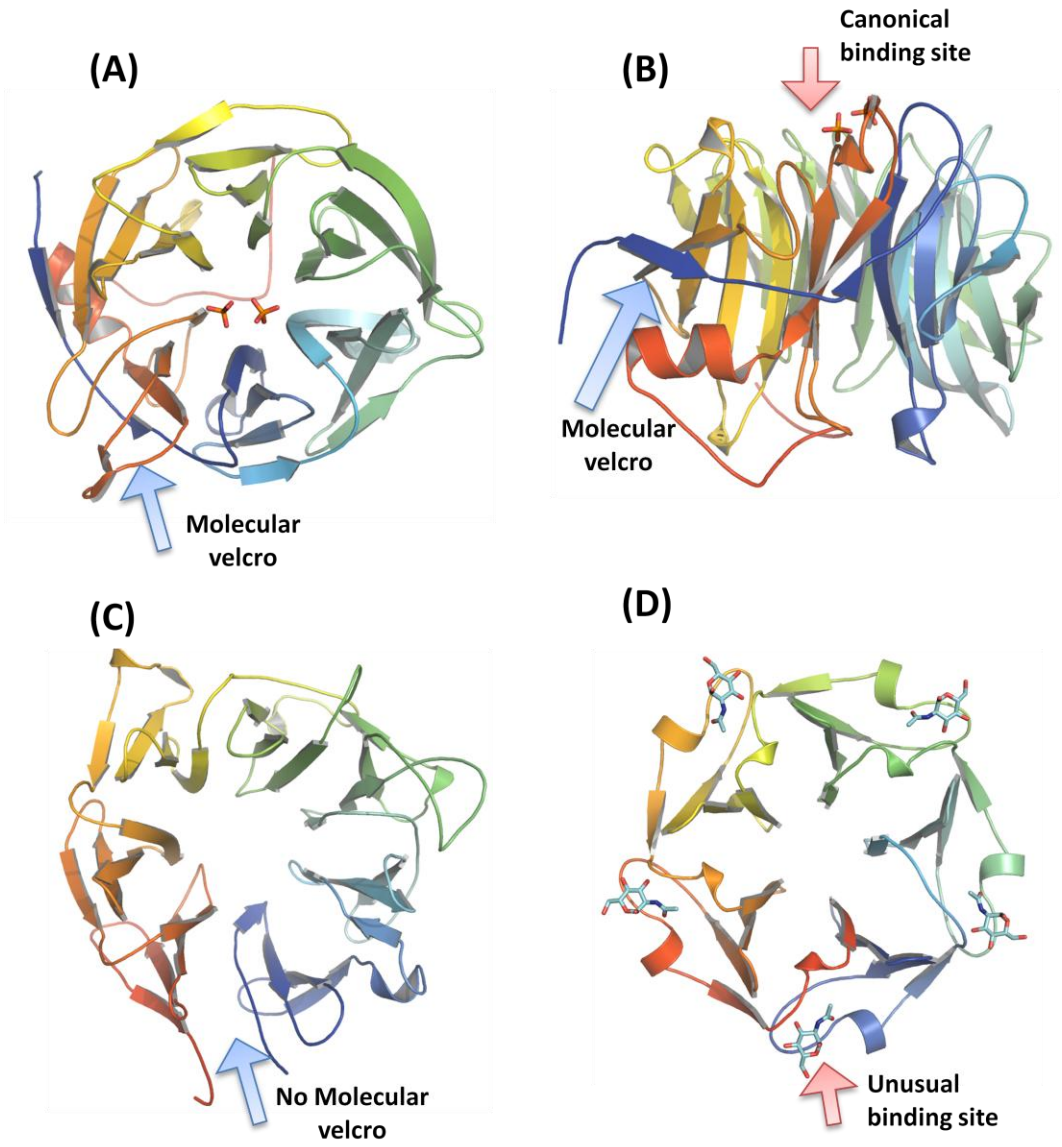


Figure 30 Beta-propeller fold protein structures. All structures shown colored from blue (N-terminus) to red (C-terminus). (A) Top view and (B) side view of 6 bladed beta-propeller Thermostable phytase (SCOPid: d1h61a_). Blue arrows pointing molecular velcros. Red arrows pointing ligand binding or active sites. (C) Top view of prolyl oligopeptidase (SCOPid: d1qfma1). No molecular velcros for this protein. (D) Top view of Tachylectin (SCOPid: d1tl2a_). Tachylectin has unusual sugar binding sites between blades.

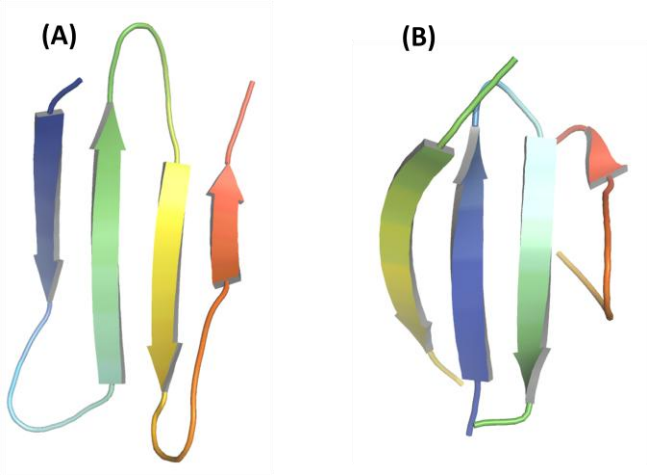


Figure 31 Different types of blades in beta-propeller. (A) Common type of blade and (B) uncommon type of 6 bladed beta propeller GyrA protein (Fifth blade from SCOPID: d1wp5a_)

Two proteins in beta-propeller fold formed different cluster (cluster 5835). Their beta-blade structures are different from other beta-propellers. This might indicates that the two gyrase domains are not homologous to the other common beta-propellers (Corbett et al, 2004).

4.2.1.4 Cluster 41: Metallo-dependent hydrolase and PHP domain

Cluster 41 contains metallo-dependent hydrolase superfamily and PHP domain-like superfamily proteins. Metallo-dependent hydrolases are TIM barrel fold proteins and these enzymes catalyze various reactions using one or more divalent metal ions (Lisa Holm 1997 JMB) coordinated by Histidines. PHP domain-like proteins are known as putative phosphate esterase and nucleases. PHP proteins belong to 7 stranded beta/alpha barrel fold. SCOP classified PHP domains and metallo-dependent hydrolases into different folds according to their number of repeating beta/alpha units.

PHP domains and metallo-dependent hydrolases have similar metal binding sites with conserved Histidines (HXH motif). They also have similar barrel cap with C-termini helices which is different from many other TIM barrels with barrel cap by N-termini helices or beta hairpins. Similar to Cluster 41, CATH database also merged PHP domains into metallo-dependent hydrolase Homologous superfamily.

The similarity between PHP domains and other superfamilies in 7 stranded beta/alpha barrel fold are not as high as PHP and MDH. Also the similarity appears to be largely come from the topological similarity of 7 beta/alpha units. This relatively closer similarity between PHP and MDH supports the hypothesis of the polyphyletic origin of 7 stranded beta/alpha barrels.

4.2.1.5 Cluster 4431: Pyrococcus furiosus hypothetical protein and major capsid protein gp5

Hypothetical protein PF0899 is a functionally unknown protein in *Pyrococcus furiosus*, a hyperthermophilic archeon. Major capsid protein gp5 is structural protein to assemble procapsid of bacteriophage HK97 (Wikoff et al, 2006). PF0899 and gp5 proteins are classified into different folds in SCOP.

Common region of PF0899 and gp5 are colored from blue to red. Major capsid protein gp5 has insertion and extension (white colored region in the Figure 32) compared to PF0899. The insertion and extension stabilize oligomeric complex of gp5 proteins. PF0899 might be transferred from viruses infecting hyperthermophilic archaea.

There are three reasons; 1) Many genes are transferred from phages to bacteria (Salzberg et al, 2001), 2) Other close pyrococci do not have homolog of PF0899 protein except the few species having close homolog, and 3) PF0899 is the closest protein to Major capsid protein gp5 among known proteins and the two proteins are quite different from other proteins. The hypothesis of horizontal gene transfer of PF0899 from viral capsid protein will be clearer, as we accumulate more genome sequences of archaeal viruses.

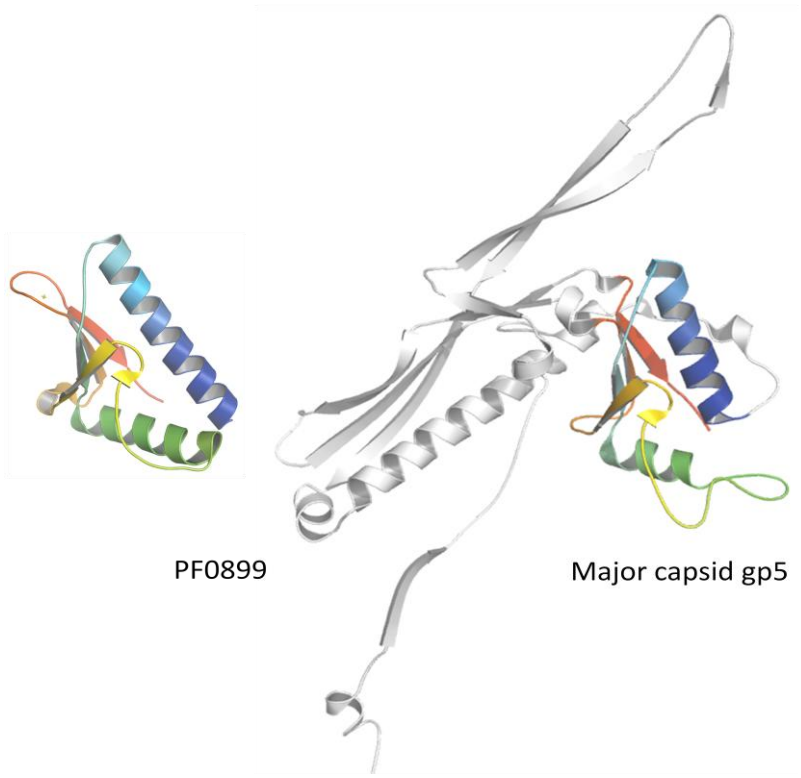


Figure 32 Representative structures in cluster 4431. The left structure is PF0899 (SCOPid: d1shea_) and the right structure is major capsid gp5 protein (SCOPid: d1ohga_). Equivalent regions have same color from blue to red. N-terminal end of PF0899 is blue and C-terminal end of PF0899 is colored in red. Insertion or deletion is white.

4.2.2 Similar in fold but unclear in homology

OB-fold (Cluster 10)				
family break; superfamily break; fold split;				
Note: Sequence/structural diversity is quite high. Generally share functional sites. MOP-like superfamily is the most divergent group.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
b	OB-fold	Hypothetical protein YgiW	1	d1nnxa_
		Bacterial enterotoxins	16	d1an8_1 d1c4qa_
		TIMP-like	5	d1br9__ d1uapa_ d1jb3a_
		Nucleic acid-binding proteins	74(81)	d1dgsa2 d1bvsa3 d1fjgl_ d1fl0a_ d1k3ra1 d1uwva1 d1jb7b_ d1je5a_ d1u5ka1 d1ltla_ d1gm5a2 d1b8aa1
		MOP-like	11	d1g2913 d1fr3a_ d1h9ka1
		Tail-associated lysozyme gp5, N-terminal domain	1	d1k28a1
		Heme chaperone CcmE	1	d1lm0a_

Viral coat and capsid proteins (Cluster 27)				
family merge; superfamily merge; fold split;				
Note: Viral coat proteins.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
B	Nucleoplasmin-like/VP (viral coat and capsid proteins)	Group II dsDNA viruses VP	6	d1p2za1 d1hx6a1 d1m3ya1
		Positive stranded ssRNA viruses	35	d1cwpa_ d1ohfa_ d1f8v.1 d1a6ca3 d1aym1_ d1ihma_ d1auya_ d1c8na_
		Satellite viruses	3	d2stv__

Cluster 59				
family merge; superfamily break; fold split;				
Note: High structural similarity but rather low sequence similarity (Dali Z: 9.4, HHsearch:0.13)				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a/b	Restriction endonuclease-like	Restriction endonuclease-like	11(33)	d1ev7a_ d1sa3a_ d1dmua_ d1sx5a_ d1xhva_ d1azo__ d3pvia_ d1gefa_ d1y1oa_
		tRNA-intron endonuclease catalytic domain-like	3	d1a79a1

Cluster 87				
family merge; superfamily merge; fold split;				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a/b	The "swiveling" beta/beta/alpha domain	Phosphohistidine domain	2	d1kbla2 d1zyma2
		LeuD-like	3	d1aco_1
		Carbamoyl phosphate synthetase, small subunit N-terminal domain	1	d1a9xb1
		Transferrin receptor ectodomain, apical domain	1	d1de4c2
		Swiveling domain of the glycerol dehydratase reactivase alpha subunit	1	d1nbwa1
		RraA-like	1	d1nxja_
		Putative cyclase	1	d1r61a_

Profilin-like proteins (Cluster 97)				
family merge; superfamily merge; fold clean;				
Note: PAS, LuxR, and sensor kinase are probably homologous (Cheng et al, 2008). SNARE, Roadblock, and GAF are probably homologous. But the relationship between the two groups is not clear.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a+b	Profilin-like	Profilin (actin-binding protein)	3	d1acf__
		GAF domain-like	7	d1stza2 d1f5ma_ d1mkma2
		PYP-like sensor domain (PAS domain)	8	d1bywa_ d1p97a_ d1v9ya_ d1nwza_ d1ll8a_ d1oj5a_
		SNARE-like	6	d1nrja_ d1gw5m2 d1h3qa_ d1ifqa_
		Pheromone-binding domain of LuxR-like quorum-sensing transcription factors	1	d1l3la2
		Sensory domain of two-component sensor kinase	2	d1ojga_
		Roadblock/LC7 domain	4	d1tgqa_

Ferredoxins (Cluster 173)				
family break; superfamily break; fold split;				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a+b	Ferredoxin-like	Acylphosphatase-like	4	d1aps__
		EF-G C-terminal domain-like	4(6)	d1t95a3 d1dar_4 d1vi7a2
		eEF-1beta-like	2	d1f60b_
		Ribosomal protein S6	2	d1loua_
		ACT-like	7	d1tdj_2 d1phza1 d1q5ya_ d1psda3 d1u8sa1
		CheY-binding domain of CheA	1(2)	d1u0sa_
		Dimeric alpha+beta barrel	5(19)	d1mli__ d1s7ia_ d1mwqa_

				d1i1ga2
		D-ribose-5-phosphate isomerase (RpiA), lid domain	2	d1o8ba2
		GlnB-like	8	d1naqa_ d1vfja_ d1o51a_ d1h3da2

SH3-like barrel proteins (Cluster 183)				
family merge; superfamily break; fold split;				
Class	Fold	Superfamily	Ndom	Representative SCOPid
b	SH3-like barrel	Cap-Gly domain	5	d1ixda_
		GW domain	2	d1m9sa2
		Chromo domain-like	1(7)	d1wgsa_
		SH3-domain	33	d1i07a_
		Electron transport accessory proteins	5	d1ugpb_ d1vie__ d1dj7b_ d1jb0e_
		Translation proteins SH3-like domain	3(9)	d1jj2p_ d1nppa2
		Tudor/PWWP/MBT	10	d1xnia2 d1h3za_ d1oi1a2

Cluster 269				
family merge; superfamily break; fold split;				
Note: They share common structural core of helical bundle.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a	SAM domain-like	SAM/Pointed domain	12	d1oxja1 d1bqv__
		RuvA domain 2-like	3(4)	d1cuk_2 d1kfta_
		C-terminal domain of RNA polymerase alpha subunit	2	d1doqa_
		Rad51 N-terminal domain-like	4	d1u9la_ d1szpa1 d1y88a1

Cluster 297				
family merge; superfamily merge; fold split;				
Note: They share common structural core and crossover loop. This cluster needs further attention.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a/b	Phosphorylase /hydrolase-like	HybD-like	2	d1c8ba_ d1cfza_
		Purine and uridine phosphorylases	11	d1b8oa_
		Peptidyl-tRNA hydrolase-like	2	d1ryba_
		Pyrrolidone carboxyl peptidase (pyroglutamate aminopeptidase)	1	d1iu8a_
		Zn-dependent exopeptidases	20	d1jwqa_ d1cg2a1 d1de4c3 d1lam_2 d1h8la2 d1obr__

Outliers of beta-Grasp fold (Cluster 400)				
family break; superfamily break; fold split;				
Note: Distinctive outliers in beta-grasp fold. Majority of beta-grasp are likely to be homologous (cluster 223). The proteins in this cluster might evolve independently or it is too distant to be detected.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a+b	beta-Grasp (ubiquitin-like)	Staphylokinase/streptokinase	1(4)	d1bmlc3
		Immunoglobulin-binding domains	2	d1hz6a_

Cluster 443				
family break; superfamily break; fold split;				
Note: There are possibly two homologous groups. This cluster requires further investigation.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a/b	Ribonuclease H-like motif	Actin-like ATPase domain	31(33)	d1g99a1 d1t6ca1 d1okja1 d1glag1 d1mwma1 d1nbwa2 d1w97l1 d1czan1 d1woqa1 d1huxa_ d1sz2a1
		Ribonuclease H-like	21(30)	d1hjra_ d1kcfa2 d1iv0a_ d1fxxa_ d1uoca_
		Translational machinery components	1(4)	d1dt9a1
		DNA repair protein MutS, domain II	2	d1e3ma3

Beta-Prism I (Cluster 488)				
family merge; superfamily merge; fold clean;				
Note: They share common structural core and same symmetry. Sequence-wise so divergent. This cluster needs further attention.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
b	beta-Prism I	Vitelline membrane outer protein-I (VMO-I)	1	d1vmoa_
		delta-Endotoxin (insecticide), middle domain	3	d1ji6a2
		Mannose-binding lectins	3	d1c3ma_

Cluster 526				
family merge; superfamily merge; fold split;				
Note: Scores between the two superfamilies; Dali Z:12.3, HHsearch:0.67. But the sequence alignment may not make much structural sense because it does not appear very similar to DALI alignment. The sequence score is erroneously high just reflecting the fact both of them have consecutive beta-strands.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
b	Supersandwich	Amine oxidase catalytic domain	5	d1n9ea1
		Galactose mutarotase-like	20	d1h54a2 d1lf6a2 d1k1xa2 d1jova_ d1nsza_ d1nkga3 d1jz8a4 d1txka2 d1cb8a3 d1qwna2 d1xsia2

Cluster 555				
family merge; superfamily merge; fold split;				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a+b	Ferredoxin-like	Bacterial exopeptidase dimerisation domain	6	d1lfwa2
		MTH1187/YkoF-like	4	d1lxja_ d1s99a_
		Hypothetical protein TT1725	1	d1j27a_
		eIF-2-alpha, C-terminal domain	1	d1q8ka2

Cluster 597				
family merge; superfamily break; fold split;				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a+b	Bacillus chorismate mutase-like	YjgF-like	3(4)	d1jd1b_
		PurM N-terminal domain-like	6	d1clia1

Cluster 668				
family merge; superfamily merge; fold split;				
Class	Fold	Superfamily	Ndom	Representative SCOPid
b	Double-stranded beta-helix	RmlC-like cupins	28	d1xe7a_ d1eyba_ d1y9qa2 d1x7na_ d1lrha_ d1v70a_ d1vj2a_ d1x8ma_ d1j1la_ d1juha_ d1sfna_ d1pmi__ d1xsqa_ d1ep0a_ d1m4oa_ d1o5ua_
		Clavaminate synthase-like	12	d1jopa_ d1e5sa_ d1dcs__ d1h2ka_ d1ds1a_ d1jr7a_ d1nx4a_ d1otja_
		cAMP-binding domain-like	12	d1cx4a1 d1ft9a2 d1omia1
		Regulatory protein AraC	1	d2arca_

N-cbl like 4 helical bundles (Cluster 778)				
family merge; superfamily merge; fold split;				
Note: 4-helical bundles. Good structural similarity score due to the structural core (Dali Z:7.3). But no detectable sequence similarity.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a	N-cbl like	N-terminal domain of cbl (N-cbl)	1	d2cbla2
		Transferrin receptor ectodomain, C-terminal domain	1	d1de4c1

Cluster 843
family break; superfamily break; fold split;
Note: They share common core of 7 beta strands in two beta-sheets packing.

Class	Fold	Superfamily	Ndom	Representative SCOPid
b	Prealbumin-like	Starch-binding domain-like	1(4)	d1nkga1
		Carboxypeptidase regulatory domain	2	d1h8la1
		Transthyretin (synonym: prealbumin)	1	d1f86a_
		Cna protein B-type domain	1(3)	d1ti6b1
		Aromatic compound dioxygenase	4	d1dmha_
		Hypothetical protein PA1324	1	d1xpna_

Cystatin-like proteins (Cluster 1182)				
family merge; superfamily break; fold split;				
Note: The proteins in this cluster share alpha-beta(x4) structural core. One protein (d1x9ya2) in cystatin superfamily was excluded from cluster and form singleton cluster (cluster 8125) because of forming swapped dimerization.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a+b	Cystatin-like	Cystatin/monellin	5(6)	d1mola_ d1kwia_ d1stfi_
		NTF2-like	19	d1gy7a_ d1m98a2 d1hkxa_ d1oh0a_ d1tp6a_ d1nwwa_ d1tuha_ d1ulib_ d1s5aa_ d1idpa_ d1sjwa_ d1mwsa1

Activator of Hsp90 ATPase & Bacterial permeability-increasing protein 1 (Cluster 1216)				
family merge; superfamily merge; fold clean;				
Note: Striking structural similarity (DaliZ 9.4), but sequence similarity is not high. The relationship should be studied further.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a+b	Aha1/BPI domain-like	Bactericidal permeability-increasing protein, BPI	2	d1ewfa1
		Activator of Hsp90 ATPase, Aha1	1	d1usub_

Two alpha hairpins: Prefoldin & tRNA binding arm (Cluster 1561)				
family merge; superfamily break; fold split;				
Note: Good sequence similarity (HHsearch 0.82) and high structural similarity (DaliZ 9.6). But they are mere two long helices connected by a loop. Also the loops are quite different in the two superfamilies.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
A	Long alpha-hairpin	Prefoldin	2	d1fxka_
		tRNA-binding arm	1(4)	d1seta1

Open-sided beta-meanders and the chimera (Cluster 1961)				
family merge; superfamily merge; fold clean;				
Note: Moderately good structural similarity score (DaliZ 5.4) but no sequence similarity detected. d1ospo_ can be considered as hybrid of histone methyltransferase (d1h3ia1) and the other outer surface protein (d1rjlc_). The relationship needs further study.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
b	open-sided	Outer surface protein	2	d1ospo_

	beta-meander	Histone H3 K4-specific methyltransferase SET7/9 N-terminal domain	1	d1h3ia1

Canonical IF3-like domains (Cluster 2859)				
family merge; superfamily merge; fold split;				
Note: (beta-alpha)x2-beta-beta core. This cluster might have homologous proteins but not sure because of high structural/sequence diversities. iSCG splitted IF3-like fold proteins into 5 different clusters due to their oligomeric status and other structural features.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a+b	IF3-like	Translation initiation factor IF3, C-terminal domain	1	d1tig__
		YhbY-like	2	d1jo0a_
		AlbA-like	3	d1nh9a_ d1vm0a_

Secretion chaperone & Arp2/3 complex subunit (Cluster 2947)				
family break; superfamily break; fold split;				
Note: Moderate structural similarity but no significant shared sequence motif was found. The modes of dimerization are different in two superfamilies. First helices have opposite directions with one strand deletion in Arp2/3 complex subunit superfamily.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a+b	Secretion chaperone-like	Type III secretary system chaperone	7(8)	d1ry9a_
		Arp2/3 complex subunits	3	d1k8kd1

Table 6 Clusters containing different superfamilies within same fold in "Fold similar; Homology unclear" category. The legend is the same as Table 2.

4.2.2.1 Cluster 10: OB-fold proteins

OB-fold proteins share the topology of 5 stranded beta-barrels. OB-fold proteins are very diverse in function but shows relatively well conserved structural fold (Arcus, 2002). OB-fold proteins are classified into 10 different superfamilies in SCOP. Among those 10 superfamilies, iSCG clustered 7 superfamilies into cluster 10. The superfamilies clustered into this cluster are followings; Nucleic acid-binding proteins, MOP-like, TIMP-

like, Bacterial enterotoxins, Heme chaperon CcmE, Hypothetical protein YgiW, and N-terminal domain of Tail-associated lysozyme gp5.

Majority of proteins (72 proteins out of 109 total proteins in cluster 10) belong to nucleic acid-binding superfamily. Nucleic acid-binding proteins are diverse in their molecular functions from anticodon-binding in tRNA synthetase to double stranded DNA binding in cold shock proteins. But their evolutionary relationship is clear with high structural similarity with same general function at the same active site with conserved sequence motif related to protein fold stability (Theobald et al, 2003b). Other superfamilies in this cluster generally do not share the function of nucleic acid binding, but they share structural similarity along with sequence motif related to fold stability (Arcus, 2002; Ginalski et al, 2004; Mitton-Fry et al, 2002; Theobald et al, 2003a). MOP-like superfamily has specialized function in binding small molecule, i.e. molybdenum (Wagner et al, 2000). TIMP-like superfamily proteins developed their function in binding and inhibiting metalloproteases (Williamson et al, 1994). Thus the general function of TIMP-like superfamily is binding proteins. Bacterial enterotoxins are subdivided into two groups, AB5-like toxins and superantigens. AB5-like toxins bind to oligosaccharides on mammalian cell surface and puncture them (Merritt & Hol, 1995). Superantigens bind to MHC II complex and T-cell receptors and make many different T-cells activated (C. Bachert, 2002). Thus the general function of superantigen proteins is binding to protein. Heme chaperon CcmE forms intermediately covalent bond with heme on histidine residue close to C-terminal end (Enggist et al, 2003). The surface for ligand binding is

hydrophobic and relatively flat to bind heme (Enggist et al, 2002). The function of Hypothetical protein YgiW is not known yet, but this protein is very similar to single stranded DNA binding proteins in Nucleic acid binding protein superfamily, which suggest the function of YgiW (Ginalski et al, 2004). N-terminal domain of Tail-associated lysozyme gp5 is a protein insert phage T4 double stranded DNA through the pore generated by lysozyme gp5 (Kanamaru et al, 2002). This N-terminal domain of gp5 has putative function of double stranded DNA binding. The superfamilies in this cluster are similar in topology but it is hard to definitely conclude their homology given the simple topology of OB-fold and vast functional diversity of those OB-fold proteins. Detailed analysis with more recent structures and sequences will give us definite answer about conclusive relationships between the proteins in this cluster.

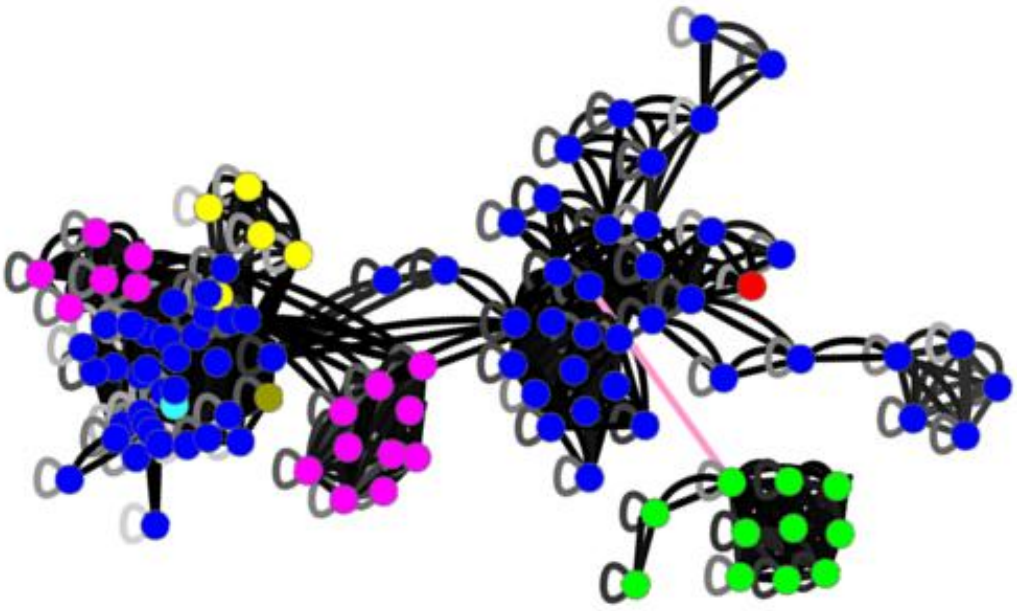


Figure 33 Network view of Cluster 10. Each domains or proteins represent as colored circles (node). Lines connecting domains (edges) represent significant similarity between domains

(above combined score 5). Different node colors were mapped for different SCOP superfamilies; Nucleic acid binding domains: blue, Bacterial enterotoxins: purple, TIMP-like: yellow, Heme chaperon CcmE: light blue, Hypothetical protein YgiW: dark yellow, MOP-like: green, N-terminal domain of lysozyme gp5: red. The edge colors represent the combined score values; the higher the value the lighter the color (from black to white). One exception is the light purple edge connecting MOP-like domains (green circles) and Nucleic acid-binding domains (blue circles). This edge is below the significant score but added to show the connectivity.

While most similarity links are confident with significant score, one link colored in light pink connecting MOP-like domains and Nucleic-acid binding domains are below the significant score threshold. MOP-like domains are indeed quite different from other proteins in the OB-fold. MOP-like domains bind small molecule and use different binding site different from other members. The binding site for small molecule is at the interface between the homo-dimer.

One other interesting feature of this cluster is that the domains form two big groups within the cluster as shown in Figure 33. Those two groups are functional groups; the nucleic acid-binding proteins (denoted as blue circles) in the left group are majorly single strand binding proteins and the nucleic acid-binding proteins in the right group are majorly double strand binding domains. Also the left group has more diverse functions than the lower group. From this observation, it is tempting to argue that the evolutionary origin of OB-fold might be in the single strand binding domains. Again, it needs more extensive study to have definite conclusion.

Within the upper group the bacterial enterotoxins also form two distinct groups; the left group consist of AB5 toxins that binds oligosaccharides on the cell surface and the right group is superantigen that bind T-cell receptors and MHC complexes. This

might indicate that those two bacterial enterotoxins evolved independently from nucleic acid binding OB-fold proteins.

4.2.2.2 Canonical IF3-like domains (Cluster 2859)

This cluster contains proteins of core structure beta-alpha-beta-alpha-beta(x2) (Figure 34). Three superfamilies are clustered; IF3 C-terminal domain, AlbA-like, and YhbY-like superfamilies. IF3 C-terminal domain binds to ribosome on the two helices (Pioletti et al, 2001). AlbA proteins probably bind DNA on the same surface (Chou et al, 2003). YhbY proteins also probably bind RNA on the same helical surface (Liu & Wyss, 2004). Given the functional information of those proteins, the proteins in this cluster might be homologous. But the sequence conservation was hard to detect.

IF3-like fold proteins in defined in SCOP database are split into 5 different clusters due to structural features, such as trimerization (cluster 1119, RNA 3'-terminal phosphate cyclase superfamily), relatively parallel helix interation (cluster 762, SirA-like superfamily), missing first strand (cluster 3822, R3H domain), and wider distance between two helices (cluster 2062, C-terminal domain of ProRS). Notably, all the proteins in the IF3-like fold are implicated to interact with DNA or RNA except SirA-like proteins. SirA proteins are implicated in disulfide bond formation. However the functional site for RNA 3'-terminal phosphate cyclases and R3H proteins probably different from other DNA/RNA interacting proteins in this fold.

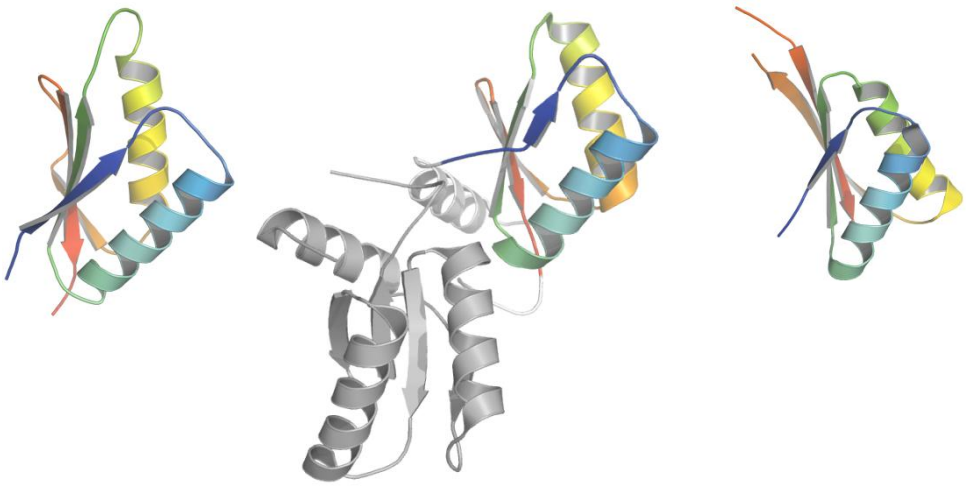


Figure 34 Representative structures in cluster 2859. The left structure is IF-3 protein and the middle structure is AlbA protein. The right structure is YhbY protein. All proteins in color blue to red color are equivalent domains. Also all equivalent secondary structural elements are in the same color. The middle structure (AlbA protein) was shown in homo-dimeric status. The dimer partner is shown in gray.

4.2.3 Partial similarity, unlikely in homology

Four helical bundles (Cluster 319)				
family merge; superfamily merge; fold break;				
Note: Four helical bundle forms the structural core of proteins in the cluster. Interestingly, bromodomain-like fold and four-helical up-and-down bundles are mirror images. They are connected through I/LWEQ domains. Partial alignments between fold groups; 3 helices aligned.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a	A middle domain of Talin 1	A middle domain of Talin 1	1	d1sj7a1
	I/LWEQ domain	I/LWEQ domain	2	d1r0da_
	Four-helical up-and-down bundle	Apolipoprotein	1	d1gs9a_
		Histidine-containing phosphotransfer domain, HPT domain	6	d1y6da_ d1c02a_ d2a0b__ d1sr2a_ d1i5na_
		Bacterial GAP domain	3	d1g4us1
		Outer surface protein C (OspC)	1	d1f1ma_
		FAT domain of focal adhesion kinase	1	d1k04a_
		Oxygen-evolving enhancer protein 3,	1	d1nzea_
		Flagellar export chaperone FliS	2	d1orja_
		Aspartate receptor, ligand-binding domain	1	d2liga_
		Nickel-containing superoxide dismutase,	1	d1t6ua_

		NiSOD		
		Mannose-6-phosphate receptor binding protein 1 (Tip47), C-terminal domain	1	d1szia_
		Domain from hypothetical 2610208m17rik protein	1	d1ug7a_
		TrmE connector domain	1	d1xzpa1
		Cytochromes	7	d1bbha_ d256ba_
		alpha-catenin/vinculin	11	d1t01a1
	Bromodomain-like	Acyl-CoA dehydrogenase C-terminal domain-like	8	d1w07a2 d1ivha1
		alpha-ketoacid dehydrogenase kinase, N-terminal domain	2	d1gkza1
		Plant invertase/pectin methylesterase inhibitor	2	d1rj1a_
		Mob1/phocein	1	d1pi1a_
	STAT-like	CAPPD, an extracellular domain of amyloid beta A4 protein	1	d1rw6a_

Coiled coil proteins (Cluster 964)				
family merge; superfamily merge; fold break;				
Note: Proteins in this cluster share coiled coil structural core similarity. The most similar proteins between superfamilies are d1avo.1 and d1i4da_ (Dali Z-score 9 and HHsearch 0.9). DNA repair protein MutS domain III are quite different from the other two superfamilies.				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a	DNA repair protein MutS, domain III	DNA repair protein MutS, domain III	2	d1e3ma1
	BAR/IMD domain-like	BAR/IMD domain-like	2	d1i4da_ d1urua_
	Four-helical up-and-down bundle	Proteasome activator reg(alpha)	1	d1avo.1

Hsp33 & GFP (Cluster 1729)				
family merge; superfamily merge; fold merge;				
Note: Surprisingly high structure similarity score between barrel protein and two beta-sheets with a helix inside. (Dali Z-score 6).				
Class	Fold	Superfamily	Ndom	Representative SCOPid
a+b	Hsp33 domain	Hsp33 domain	3	d1hw7a_
	GFP-like	GFP-like	3	d1ggxa_ d1gl4a1

Table 7 Summary for clusters in "partial similarity, unlikely homology" category. The legend is same as Table 2.

4.2.3.1 Coiled coil proteins (Cluster 964)

This cluster contains many different coiled coil proteins. They are BAR/IMD domain-like proteins, DNA repair protein MutS domain III, and proteasome activator reg (alpha).

Since they are all alpha helical proteins, they share structural similarities came from this alpha helical packing. The structural similarity between them measured by DaliLite Z score is close to 9 that is quite high score. This high score, however, just reflect the fact that their structural alignment is quite long over 120 or more residues were aligned by matching long helices (Figure 35). One other thing should be mentioned is that HHsearch probability is also quite high 0.9 or 90% but the alignments from HHsearch are very short (10 residues) and meaningless. This high HHsearch score came from the secondary structure matching and amino acid composition bias. In general significant structural similarity and sequence similarity are quite strong evidence for homology but those similarity scores need to be considered based on the quality (or correctness) of alignments.

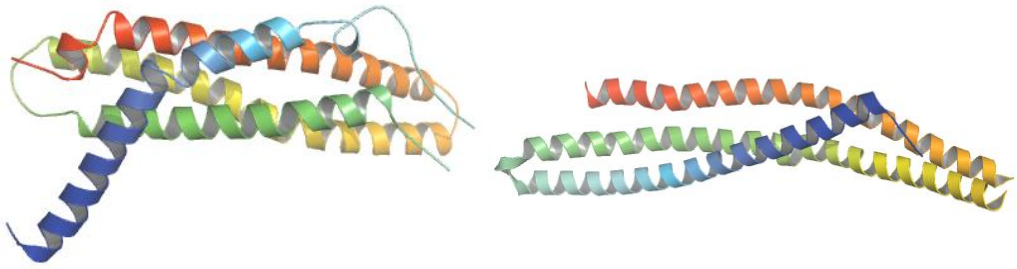


Figure 35 Representative structures from Cluster 964. The left structure is BAR/IMD domain like and the right structure is Four-helical up-and-down bundle structure. Both structures are colored from blue to red. Blue ends are N-termini and red ends are C-termini.

4.2.3.2 Cluster 1729: Hsp33 and GFP-like proteins

This cluster contains Hsp33 (Heat Shock Protein 33) and GFP (Green Fluorescent Protein)-like fold proteins. As the name implies, Hsp33 is a molecular chaperone. Hsp33 is different from other chaperon in the regulation of chaperone function since Hsp33 is induced by redox potential change by the oxidizing environment (Vijayalakshmi et al, 2001). GFP proteins emit fluorescent light and extensively used in laboratories.

Hsp33 and GFP fold are different in their three dimensional shape. The architecture of Hsp33 is two beta sheet that flanked by alpha helices packed against each other with a long helix inside. The architecture of GFP is a perfect beta barrel with a helix inside of the barrel. The DaliLite Z-score 6 between GFP and Hsp33 is very surprising score given their architectural and topological differences (Figure 36). However there is no sequence, structural or functional evidence that the two proteins share ancestors, since this structural similarity detected by structure comparison programs are spurious and random similarity.

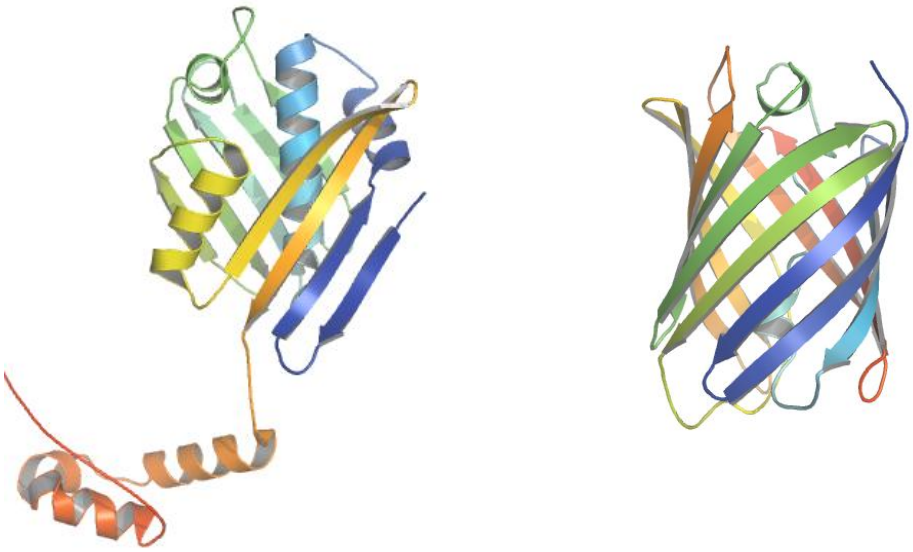


Figure 36 Representative structures from cluster 1729. The left structure is Hsp33 and the right structure is Green Fluorescent Protein (GFP). Structures are colored from blue to red. The N-terminal ends are colored in blue and the C-terminal ends are colored in red.

4.2.4 No Similarity, wrong cluster

Mirror images (Cluster 363)				
family merge; superfamily merge; fold break;				
Note: mirror image				
Class	Fold	Superfamily	Ndom	Representative SCOPid
b	Single-stranded right-handed beta-helix	Pectin lyase-like	19	d1rwra_ d1daba_ d1ofla_ d1ru4a_ d1tyv__ d1bhe__ d1gq8a_ d1bn8a_ d1qcxa_ d1ogmx2 d1h80a_
		Alpha subunit of glutamate synthase, C-terminal domain	1	d1ofda1
		C-terminal domain of adenylylcyclase associated protein	2	d1k4za_
		Stabilizer of iron transporter SufD	1	d1vh4a_
	Single-stranded left-handed beta-helix	Trimeric LpxA-like enzymes	11	d1fxja1 d1j2za_ d1qrea_ d1ocxa_ d1tdta_ d1ssqa_
		An insect antifreeze protein	1	d1l0sa_
		Adhesin YadA, collagen-binding domain	1	d1p9ha_

Table 8 Summary of cluster in "No similarity, wrong cluster" category. Table legend is the same as in Table 2.

4.2.4.1 Mirror images (Cluster 363)

Cluster 363 contains right handed single stranded beta-helix and left handed single stranded beta-helices. Since the two folds do not have any structural similarity, this cluster is clearly wrong. This cluster is formed because the wrongly high score of the most influencing DaliLite score. The program DaliLite uses contact matrix similarity to quantify the similarity between structures. And the chirality information is lost when the contact matrix is compared. In general, this loss of information does not cause any problem, but this regular left-handed beta-helix and right-handed beta-helix was not distinguished by the contact matrix. This problem will be fixed by checking the chirality.

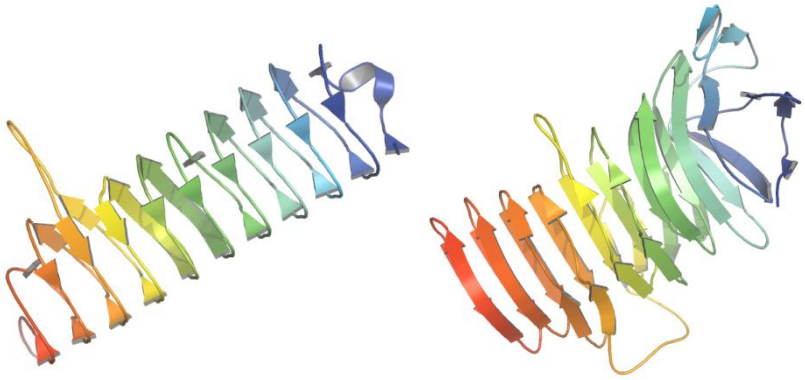


Figure 37 The representative structures in cluster 363. The left structure is single-stranded left handed beta-helix and the right structure is single-stranded right handed beta-helix. The structures are colored from blue to red. Blue color represents N-terminus and red color represents C-terminus.

CHAPTER 5

Concluding Remarks

Homology inference is one of classical topic in computational biology but still an ongoing quest for computational biologist. Many homology inference methods are developed in searching sequence similarity or structural similarities. But after proteins had diverged and found their own functional niche in the ecological environments (i.e. the molecular, cellular, or organismal context) and changed into quite different sequences and structures, it is very challenging task to elucidate their true evolutionary history.

The work described in this dissertation is can viewed on the perspective of finding homologous proteins without consulting previous knowledge or gold standard. This procedure of extracting new rules without consulting gold standard (commonly known as clustering analysis) is, however, very challenging task and one of most extensively studied area. In bigger view of scientific activities, this clustering analysis is also one of versatile technique used by researchers in vast array of expertise, since the core idea of extracting information without previous knowledge is indeed a very appealing to researchers in many fields of studies. Also it should be emphasized that the work described in this dissertation is not the first one using clustering in protein classification. Already a large body of research was done on this particular subject.

One might ask for the validity of the basic assumption at this point; why do we need to develop an automatic classification procedure (or clustering analysis procedure) to classify proteins? One generally accepted answer is that the number of newly found sequences and structures are beyond human ability to make good classification. In another point of view, the answer might be that the automated classification procedure might simply filter likely homologs out of vast number of non-homologs. This will reduce the effort of experts which can be already a big success. However, this modest goal will be naturally achieved if we try to make a automated classification procedure that is comparable (not necessarily better than experts) to experts in the field like SCOP database or CATH (in the sense this is semi-curated by experts not entirely automatic).

The main focus in previous research of the field of protein classification was on developing better algorithms to fix problems in similarity measures. Even though similarity measures are generally good to detect strong similarities, especially sequence similarity measures are very accurate by the endeavors of many brilliant researchers in the field of remote sequence similarity search, the similarity measure inevitably makes some spurious errors because these similarity measures are statistical in nature. Even if the error rate is very low, extensive comparison of proteins, i.e. the dimension of the similarity matrix for 7000 proteins is 7000 times 7000 ~=49,000,000, naturally increases the number of errors by sheer large number of comparisons. This is why in this study we combined many different similarity measures. Another reason to combine is that it would be beneficial to combine information from different sources, i.e. sequence and

structure. Indeed combining information help to have better clustering, as expected. This combination of information and building better classification was largely neglected in the field.

Beside combining and making better similarity measure, to achieve more accurate classification (i.e. comparable to manual classification or low errors in classification) we developed a stringent clustering method. This clustering method developed and termed as self consistency grouping (SCG) is an appealing methodology in not making errors because it requires strict criterion that all proteins should be more similar to form a cluster (details in section 3.2). Indeed this requirement is very strict and the clusters were very small. Since the methodology is very sensitive to the quality of similarity measure, SCG was used as a tool to combine different information. SCG is now became one of few tools to combine information without consulting previous known examples (or training set). It should be noted that after the combination of scores, the quality of the "combined similarity measure" was still low. There are possibly several reasons; 1) the combination was done in linear way. Since the main focus was to show that the combination works, the simplest way was chosen. 2) the normalization might be inadequate. As shown in the over-splitting in small or abundant proteins, like HTH or Rossmann-like folds, the simple normalization by Z-score without removing homologous protein made the same normalized score have different significance in different kind of proteins. 3) The number of scores tried might not enough to achieve higher quality.

One of the ambitious future goals is to establish better reference or gold standard database. Currently, SCOP database is regarded as a gold standard. Even though SCOP is one of best available resource, it is not perfect, as shown by the deep analysis done in this work. It is possible that the quality of automated classifications were graded lower than the actual quality because of the imperfect reference. The work described in this thesis possibly contributes to the field by iteratively improving reference database. This work started from SCOP domains and the first version of automated classification is finished. Based on the manual analysis we can start second round of refined domains (domain definitions are sometimes wrong as we found during analyses) with improved reference. This improved reference can be built by combining first version of automatic classification result with SCOP database result. After each iteration, the reference database can be closer to the true evolutionary classification and apart from initial starting point SCOP database.

Besides protein classification, there are many different future research directions. Since databases of millions of pairwise sequence comparisons and structural comparisons were produced, those databases can be very useful to any large scale sequence-structure-function relationship studies. Those databases already used in improving profile-profile alignment methods and remote similarity search methods. One interesting new research might be studying the power of sequence positional correlation in improving the quality of protein structure predictions.

Bibliography

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215**(3)**:** 403-410

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**(17)**:** 3389-3402

Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJP, Chothia C, Murzin AG (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Research* **36:** D419-D425

Anna R. Panchenko YIW, Larisa A. Panchenko, Thomas Madej, (2005) Evolutionary plasticity of protein families: Coupling between sequence and structure variation. *Proteins: Structure, Function, and Bioinformatics* **61**(3)**:** 535-544

Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM (2005) The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol Rev* **29**(2)**:** 231-262

Aravind L, Anantharaman V, Koonin EV (2002) Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA. *Proteins* **48**(1)**:** 1-14

Arcus V (2002) OB-fold domains: a snapshot of the evolution of sequence, structure and function. *Current Opinion in Structural Biology* **12**(6)**:** 794-801

Bandarian V, Pattridge KA, Lennon BW, Huddler DP, Matthews RG, Ludwig ML (2002) Domain alternation switches B12-dependent methionine synthase to the activation conformation. *Nat Struct Mol Biol* **9**(1)**:** 53-56

Brenner SE, Koehl P, Levitt M (2000) The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res* **28**(1)**:** 254-256

Bruno WJ, Socci ND, Halpern AL (2000) Weighted Neighbor Joining: A Likelihood-Based Approach to Distance-Based Phylogeny Reconstruction. *Mol Biol Evol* **17**(1)**:** 189-197

C. Bachert PG, P. van Cauwenberge, (2002) *Staphylococcus aureus* enterotoxins: a key in airway disease? *Allergy* **57**(6)**:** 480-487

Chaudhuri I, Soding J, Lupas AN (2008) Evolution of the beta-propeller fold. *Proteins* **71**(2)**:** 795-803

Cheng H, Kim BH, Grishin NV (2008) Discrimination between distant homologs and structural analogs: lessons from manually constructed, reliable data sets. *J Mol Biol* **377**(4)**:** 1265-1278

Chou CC, Lin TW, Chen CY, Wang AHJ (2003) Crystal structure of the hyperthermophilic archaeal DNA-binding protein Sso10b2 at a resolution of 1.85 angstroms. *Journal of Bacteriology* **185**(14)**:** 4066-4073

Coles M, Hulko M, Djuranovic S, Truffault V, Koretke K, Martin J, Lupas AN (2006) Common evolutionary origin of swapped-hairpin and double-psi beta barrels. *Structure* **14**(10)**:** 1489-1498

Corbett KD, Shultzaberger RK, Berger JM (2004) The C-terminal domain of DNA gyrase A adopts a DNA-bending beta-pinwheel fold. *Proc Natl Acad Sci U S A* **101**(19)**:** 7293-7298

de Hoon MJL, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. *Bioinformatics* **20**(9)**:** 1453-1454

Dietmann S, Park J, Notredame C, Heger A, Lappe M, Holm L (2001) A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. *Nucleic Acids Res* **29**(1)**:** 55-57

Dobzhansky T (1964) Biology, Molecular and Organismic. *Am Zool* **4:** 443-452

Dokholyan NV (2005) The architecture of the protein domain universe. *Gene* **347**(2)**:** 199-206

Dokholyan NV, Shakhnovich B, Shakhnovich EI (2002) Expanding protein universe and its origin from the biological Big Bang. *Proceedings of the National Academy of Sciences of the United States of America* **99**(22)**:** 14132-14136

Engel M, Hoffmann T, Wagner L, Wermann M, Heiser U, Kiefersauer R, Huber R, Bode W, Demuth H-U, Brandstetter H (2003) The crystal structure of dipeptidyl peptidase IV (CD26) reveals its functional regulation and enzymatic mechanism. *Proceedings of the National Academy of Sciences of the United States of America* **100**(9)**:** 5063-5068

Enggist E, Schneider MJ, Schulz H, Thony-Meyer L (2003) Biochemical and Mutational Characterization of the Heme Chaperone CcmE Reveals a Heme Binding Site. *J Bacteriol* **185**(1)**:** 175-183

Enggist E, Th?y-Meyer L, G?tert P, Pervushin K (2002) NMR Structure of the Heme Chaperone CcmE Reveals a Novel Functional Motif. *Structure* **10**(11)**:** 1551-1557

Everitt B (1974) *Cluster analysis*, New York: John Wiley & Sons.

Felsenstein J (1989) PHYLIP—Phylogeny Inference Package (version 3.2). *Cladistics* **5**(2)**:** 164-166

Fowlkes EB, Mallows CL (1983) A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association* **78**(383)**:** 553-569

Fraser ME, James MNG, Bridger WA, Wolodko WT (1999) A detailed structural description of Escherichia coli succinyl-CoA synthetase. *Journal of Molecular Biology* **285**(4)**:** 1633-1653

Gilks WR, Audit B, De Angelis D, Tsoka S, Ouzounis CA (2002) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics* **18**(12)**:** 1641-1649

Ginalski K, Kinch L, Rychlewski L, Grishin NV (2004) BOF: a novel family of bacterial OB-fold proteins. *FEBS Letters* **567**(2-3)**:** 297-301

Grishin NV (1999) Phosphatidylinositol phosphate kinase: a link between protein kinase and glutathione synthase folds. *J Mol Biol* **291**(2)**:** 239-247

Grishin NV (2001) Fold change in evolution of protein structures. *J Struct Biol* **134**(2-3)**:** 167-185

Heger A, Holm L (2003) Exhaustive Enumeration of Protein Domain Families. *Journal of Molecular Biology* **328**(3)**:** 749-767

Holm L, Sander C (1993) Protein structure comparison by alignment of distance matrices. *J Mol Biol* **233**(1)**:** 123-138

Iyer L, Burroughs AM, Aravind L (2006) The prokaryotic antecedents of the ubiquitin-signaling system and the early evolution of ubiquitin-like beta-grasp domains. *Genome Biol* **7**(7)**:** R60

Kanamaru S, Leiman PG, Kostyuchenko VA, Chipman PR, Mesyanzhinov VV, Arisaka F, Rossmann MG (2002) Structure of the cell-puncturing device of bacteriophage T4. *Nature* **415**(6871)**:** 553-557

Kikuchi S, Tominaga D, Arita M, Takahashi K, Tomita M (2003) Dynamic modeling of genetic networks using genetic algorithm and S-system. *Bioinformatics* **19**(5)**:** 643-650

Kinch LN, Cheek S, Grishin NV (2005) EDD, a novel phosphotransferase domain common to mannose transporter EIIA, dihydroxyacetone kinase, and DegV. *Protein Sci* **14**(2)**:** 360-367

Kinch LN, Grishin NV (2002) Expanding the nitrogen regulatory protein superfamily: Homology detection at below random sequence identity. *Proteins* **48**(1)**:** 75-84

Kundrat M (2004) When did theropods become feathered?--evidence for pre-Archaeopteryx feathery appendages. *J Exp Zoolog B Mol Dev Evol* **302**(4)**:** 355-364

Liu DJ, Wyss DF (2004) Letter to the Editor: Solution structure of the hypothetical protein SAV1595 from Staphylococcus aureus, a putative RNA binding protein. *Journal of Biomolecular Nmr* **29**(3)**:** 391-394

Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, Ke Z, Krylov D, Lanczycki CJ, Liebert CA, Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Thanki N, Yamashita RA, Yin JJ, Zhang D, Bryant SH (2007) CDD: a conserved domain database for interactive domain family analysis. *Nucleic Acids Res* **35**(Database issue)**:** D237-240

Marsden RL, McGuffin LJ, Jones DT (2002) Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci* **11**(12)**:** 2814-2824

Merritt EA, Hol WGJ (1995) AB5 toxins. *Current Opinion in Structural Biology* **5**(2)**:** 165-171

Mitton-Fry RM, Anderson EM, Hughes TR, Lundblad V, Wuttke DS (2002) Conserved Structure for Single-Stranded Telomeric DNA Recognition. *Science* **296**(5565)**:** 145-147

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) Scop - a Structural Classification of Proteins Database for the Investigation of Sequences and Structures. *Journal of Molecular Biology* **247**(4)**:** 536-540

Nagano N, Orengo CA, Thornton JM (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J Mol Biol* **321**(5)**:** 741-765

Orengo CA, Jones DT, Thornton JM (1994) Protein Superfamilies and Domain Superfolds. *Nature* **372**(6507)**:** 631-634

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH--a hierarchic classification of protein domain structures. *Structure* **5**(8)**:** 1093-1108

Park J, Holm L, Heger A, Chothia C (2000) RSDB: representative protein sequence databases have high information content. *Bioinformatics* **16**(5)**:** 458-464

Pei J, Kim B-H, Grishin NV (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. *Nucl Acids Res***:** gkn072

Pioletti M, Schlunzen F, Harms J, Zarivach R, Gluhmann M, Avila H, Bashan A, Bartels H, Auerbach T, Jacobi C, Hartsch T, Yonath A, Franceschi F (2001) Crystal structures of complexes of the small ribosomal subunit with tetracycline, edeine and IF3. *Embo Journal* **20**(8)**:** 1829-1839

Ponting CP, Russell RB (2000) Identification of distant homologues of fibroblast growth factors suggests a common ancestor for all beta-trefoil proteins. *Journal of Molecular Biology* **302**(5)**:** 1041-1047

Qian J, Luscombe NM, Gerstein M (2001) Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model. *Journal of Molecular Biology* **313**(4)**:** 673-681

Reeves GA, Dallman TJ, Redfern OC, Akpor A, Orengo CA (2006) Structural Diversity of Domain Superfamilies in the CATH Database. *Journal of Molecular Biology* **360**(3)**:** 725-741

Russell RB, Saqi MAS, Sayle RA, Bates PA, Sternberg MJE (1997) Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *Journal of Molecular Biology* **269**(3)**:** 423-439

Rychlewski L, Jaroszewski L, Li W, Godzik A (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci* **9**(2)**:** 232-241

Sadreyev R, Grishin N (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol* **326**(1)**:** 317-336

Sadreyev RI, Baker D, Grishin NV (2003) Profile-profile comparisons by COMPASS predict intricate homologies between protein families. *Protein Science* **12**(10)**:** 2262-2272

Sadreyev RI, Tang M, Kim BH, Grishin NV (2007) COMPASS server for remote homology inference. *Nucleic Acids Research* **35:** W653-W658

Salzberg SL, White O, Peterson J, Eisen JA (2001) Microbial genes in the human genome: Lateral transfer or gene loss? *Science* **292**(5523)**:** 1903-1906

Sam V, Tai C-H, Garnier J, Gibrat J-F, Lee B, Munson P (2006) ROC and confusion analysis of structure comparison methods identify the main causes of divergence from manual protein classification. *BMC Bioinformatics* **7**(1)**:** 206

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research* **13**(11)**:** 2498-2504

Soding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research* **33:** W244-W248

Sonnhammer EL, Eddy SR, Durbin R (1997) Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**(3)**:** 405-420

Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng* **12**(5)**:** 387-394

Tatusov RL, Koonin EV, Lipman DJ (1997) A Genomic Perspective on Protein Families. *Science* **278**(5338)**:** 631-637

Theobald DL, Cervantes RB, Lundblad V, Wuttke DS (2003a) Homology Among Telomeric End-Protection Proteins. **11**(9)**:** 1049-1050

Theobald DL, Mitton-Fry RM, Wuttke DS (2003b) NUCLEIC ACID RECOGNITION BY OB-FOLD PROTEINS. *Annual Review of Biophysics and Biomolecular Structure* **32**(1)**:** 115-133

Van Rijsbergen CJ (1979) Information retrieval. 2nd edition. *Information retrieval 2nd edition*

Vens C, Struyf J, Schietgat L, Džeroski S, Blockeel H (2008) Decision trees for hierarchical multi-label classification. *Machine Learning* **73**(2)**:** 185-214

Vijayalakshmi J, Mukhergee MK, Graumann J, Jakob U, Saper MA (2001) The 2.2 Å Crystal Structure of Hsp33: A Heat Shock Protein with Redox-Regulated Chaperone Activity. *Structure* **9**(5)**:** 367-375

Wagner UG, Stupperich E, Kratky C (2000) Structure of the Molybdate/Tungstate Binding Protein Mop from Sporomusa ovata. *Structure* **8**(11)**:** 1127-1136

Wikoff WR, Conway JF, Tang JH, Lee KK, Gan L, Cheng NQ, Duda RL, Hendrix RW, Steven AC, Johnson JE (2006) Time-resolved molecular dynamics of bacteriophage HK97 capsid maturation interpreted by electron cryo-microscopy and X-ray crystallography. *Journal of Structural Biology* **153**(3)**:** 300-306

Williamson RA, Martorell G, Carr MD, Murphy G, Docherty AJP, Freedman RB, Feeney J (1994) Solution structure of the active domain of tissue inhibitor of metalloproteinases-2. A new member of the OB fold protein family. *Biochemistry* **33**(39)**:** 11745-11759

Yu L, Gunasekera AH, Mack J, Olejniczak ET, Chovan LE, Ruan X, Towne DL, Lerner CG, Fesik SW (2001) Solution structure and function of a conserved protein SP14.3 encoded by an essential Streptococcus pneumoniae gene. *Journal of Molecular Biology* **311**(3)**:** 593-604

Zhang Y, Skolnick J (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* **33**(7)**:** 2302-2309

Zhu J, Weng Z (2005) FAST: a novel protein structure alignment algorithm. *Proteins* **58**(3)**:** 618-627