International Journal of Computational Bioinformatics and In Silico Modeling
Vol. 3, No. 3 (2014): 407-411
Research Article
Open Access

ISSN: 2320-0634

Analysis of Fold Classification Technique in FSSP Database

Manish Kumar[1,*] and Ajay Prakash[2]

[1] *Shri Venkateshwara University, Uttar Pradesh.*
[2] *S.M College Chandausi, Uttar Pradesh*

*Corresponding author: Manish Kumar; e-mail: bioinfomoney@gmail.com

Received: 23 May 2014 Accepted: 07 June 2014 Online: 18 June 2014

ABSTRACT

The FSSP (Fold Classification based on Structure-Structure alignment of Proteins) is based on a structural alignment of all pairwise combinations of the proteins in DALI program. PDB has a number of redundant structures of proteins whose sequences and structures are 30% or more identical. A set of representative structures within PDB without these redundant entries was first produced by aligning all of the PDB structures with DALI. Thus each protein in the subset was then subdivided into individual domains. These domains were then aligned structurally with DALI to identify common folds. Again redundant folds were eliminated, after that a set of representative folds were chosen. Therefore there is a cluster of folds that corresponds to every representative fold type that are of the similar approximate structure. Thus the domains that have already a given cluster of folds are structurally connected, and therefore the cluster is drawn by structural alignments of these domains. Thus, fold clusters could also be organized in a hierarchical fashion with folds drawn by the most low-scoring alignments at the top of the hierarchy.

Keywords: FSSP, SCOP, DALI, CATH, MMDB, α-Helices, β-sheets

INTRODUCTION

Understanding and using proteins is a vital area of research within the ever-more important fields of biology and biotechnology [1]. The study of biological information from protein sequences is essential for the study of cellular functions and interactions, and protein fold recognition plays a vital role in the forecast of protein structures. It is unfortunate that the prediction of protein fold patterns is a challenge because of the presence of compound protein structures [2]. Proteins are thought to have a corporate fold pattern if they have the same main secondary structures with the same arrangement and topology. Fold recognition is the recognition of the structural fold of a protein founded on the given order information, and the number of possible protein folds is expected to be restricted. Thus, expectation depends on the background of 3D folds [3]. FSSP (fold classification based on structure-structure alignment of proteins) database classifies proteins based on their pairwise combinations i.e. structural alignment in the Brookhaven structural database [4]. DALI (Distance matrix ALIgnment) database can identify similar folding patterns. It uses screening program to examine the entire PDB and identify similar structures to the newly analyzed structure [5].

Protein families are known to retain the shape of the fold even when sequences have diverged below the limit of detection of significant similarities at the sequence level [6]. Whereas protein families have clear evolutionary relationships, and protein superfamilies have probable evolutionary relationships, proteins are said to have a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections. The similarities can be detected by structural comparisons that merge protein families of known 3-D structure into structural classes, and the members of which can or might not be evolutionarily related [7-10].

In classification into the homologous superfamilies, proteins are clustered according to their similarity in structure and function [11]. Thus different proteins with the similar fold often have peripheral elements of secondary structure and turn regions that differ in size and conformation. In many respects, the term fold is used synonymously with structural motif but generally refers to larger combinations of secondary structures-in some cases; a fold comprises half of a proteins total structure.

MATERIALS AND METHODS

The FSSP database presents a continuously updated structural classification of three-dimensional protein folds. It is derived victimization an automatic structure comparison program (DALI) for the all-against-all-comparison of three dimensional coordinates sets in the Protein Data Bank. Currently in FSSP database the basic structural entity used are protein chains, which are identified by the Protein Data Bank (PDB) entry code plus chain identifier [12]. In this paper we have done Statistical analysis of FSSP database through SPSS and discussed the fold classification technique used by FSSP.

RESULTS AND DISCUSSION

Out of 100 nearly 73.3% respondents feel that FSSP online protein structure classification database utilizes fold classification technique [Table 1].

Table 1. Fold classification technique.

	Frequency	Percent	Valid Percent	Cumulative Percent
SCOP	10	8.3	8.3	8.3
DALI	15	12.5	12.5	20.8
MMDB	7	5.8	5.8	26.7
FSSP	88	73.3	73.3	100.0
Total	120	100.0	100.0	

Out of 100 nearly 77.5% respondents agree to the statement that Class α/β of protein is comprised of Mainly parallel Beta sheets with intervening alpha helices, but may also have mixed β sheets [Table 2].

Table 2. Class α/β of protein

	Frequency	Percent	Valid Percent	Cumulative Percent
Mainly parallel Beta sheets with intervening alpha helices, but may also have mixed Sheets	93	77.5	77.5	77.5
Mainly segregated Alpha helices and antiparallel Beta sheets	8	6.7	6.7	84.2
Antiparallel alpha sheets, usually two sheets in close contact forming a sandwich	13	10.8	10.8	95.0
None of the above	6	5.0	5.0	100.0
Total	120	100.0	100.0	

Nearly 70.8% respondents feel that FSSP utilizes structure – structure alignment of proteins classification technique [Table 3].

Table 3. Classification Technique used by FSSP.

	Frequency	Percent	Valid Percent	Cumulative Percent
Number of hierarchical level	28	23.3	23.3	23.3
Hierarchical level plus architect and fold	7	5.8	5.8	29.2
Structure-structure alignment of proteins	85	70.8	70.8	100.0
Total	120	100.0	100.0	

Regarding the uses of RVP – net online program, nearly 78.3% respondents feel that RVP-net is online program for identifying solvent accessibility [Table 4].

Table 4. RVP-net online program

	Frequency	Percent	Valid Percent	Cumulative Percent
Solvent accessibility	94	78.3	78.3	78.3
Protein folding	4	3.3	3.3	81.7
Three-dimensional structure of protein	19	15.8	15.8	97.5
Sequence of amino-acids in protein	3	2.5	2.5	100.0
Total	120	100.0	100.0	

Out of 100 nearly 75% respondents who feel FSSP online protein structure classification database utilizes fold classification technique agree to the statement that as protein evolves, it 'Retain function and specificity', 'Retain function but alter specificity', 'Change to a related function; or a similar function in a different metabolic context' and 'Change to a completely unrelated function' [Table 5, 6] [Figure 1] (Chi Square test statistic = 18.750, p – value = 0.027< 0.05) [Test 3].

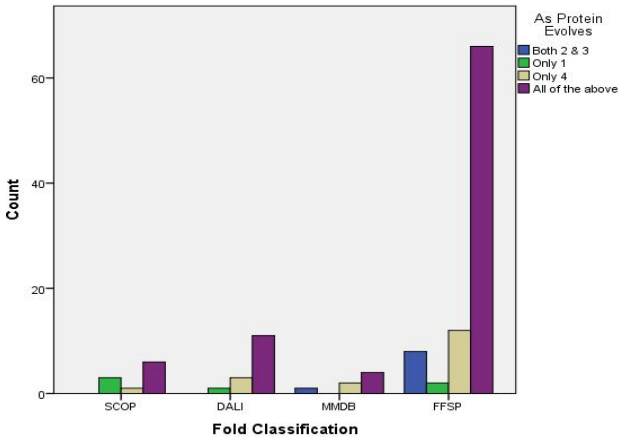

Figure 1. Fold Classification as protein evolves

Out of 100 about 79.6% respondents agree that Class α/β of protein is comprised of mainly parallel Beta sheets with intervening alpha helices, but may also have mixed β sheets. They also said that Pfam is a protein structure database and NATASA is sever used for identifying solvent accessibility [Table 7, 8] [Figure 2] (Chi Square test statistic = 21.271, p – value = 0.011 < 0.05) [Test 4].

Table 5. Fold Classification **(Test – 3)**

			As Protein Evolves				Total
			Both 2 & 3	Only 1	Only 4	All of the above	
Fold Classification	SCOP	Count	0	3	1	6	10
		% within Fold Classification	0.0%	30.0%	10.0%	60.0%	100.0%
	DALI	Count	0	1	3	11	15
		% within Fold Classification	0.0%	6.7%	20.0%	73.3%	100.0%
	MMDB	Count	1	0	2	4	7
		% within Fold Classification	14.3%	0.0%	28.6%	57.1%	100.0%
	FSSP	Count	8	2	12	66	88
		% within Fold Classification	9.1%	2.3%	13.6%	75.0%	100.0%
Total		Count	9	6	18	87	120
		% within Fold Classification	7.5%	5.0%	15.0%	72.5%	100.0%

Table 6. Chi-Square Tests of Fold Classification **[Test -3]**

Chi-Square Tests			
	Value	**Degree of Freedom**	**Asymptotic Significance (2-sided)**
Pearson Chi-Square	18.750[a]	9	.027
Likelihood Ratio	14.295	9	.112
Linear-by-Linear Association	.181	1	.671
N of Valid Cases	120		

** Superscript 'a' represents 0 cells (.0%) have expected count less than 5.

Table 7. Class α/β **[Test – 4]**

			Protein structure classification				Total
			Only Pfam	Only NATASA	Both	None	
Class α/β	Mainly parallel Beta sheets with intervening alpha helices, but may also have mixed β sheets	Count	74	2	12	5	93
		% within Class α/β	79.6%	2.2%	12.9%	5.4%	100.0%
	Mainly segregated Alpha helices and antiparallel Beta sheets	Count	7	0	0	1	8
		% within Class α/β	87.5%	0.0%	0.0%	12.5%	100.0%
	Antiparallel alpha sheets, usually two sheets in close contact forming a sandwich	Count	8	1	4	0	13
		% within Class α/β	61.5%	7.7%	30.8%	0.0%	100.0%
	None of the above	Count	4	2	0	0	6
		% within Class α/β	66.7%	33.3%	0.0%	0.0%	100.0%
Total		Count	93	5	16	6	120
		% within Class α/β	77.5%	4.2%	13.3%	5.0%	100.0%

Frequency: Frequency represents the number of times a specific category is repeated in that variable. From the above frequency distribution table, we see that the frequency of the category 'Only Pfam' is 93. This indicates that this category have occurred 93 times in Protein structure classification.

Percent: Percent represents the percentage terms of the relative weight of each of the categories. From the above frequency distribution table, we see that the frequency of the category 'Only Pfam' is 93. Therefore, the percentage of occurrence of only Pfam category is 93/120 = 0.775 or 77.5. Percentage calculation takes the entire samples into account (includes both missing and non missing observations).

Valid Percent: Valid percent represents the percentage terms of the relative weight of each of the "valid" categories only. That is, it computes the percentage values based on the non missing observations.

Cumulative Percent: Cumulative percent is obtained by adding the valid percent column together row by row.

Table 8 Chi-Square Tests of Class α/β **[Test – 4]**

Chi-Square Tests			
	Value	**Degree of Freedom**	**Asymptotic Significance (2-sided)**
Pearson Chi-Square	21.271[a]	9	.011
Likelihood Ratio	16.065	9	.066
Linear-by-Linear Association	.133	1	.715
N of Valid Cases	120		

Out of 100 about 68.2% respondents feel that FSSP utilizes structure-structure alignment of proteins classification technique. They also said that Pfam is a protein structure database and NATASA is sever used for identifying solvent accessibility [Table 9, 10] [Figure 3] (Chi Square test statistic = 14.345, p – value = 0.026< 0.05) [Test 5].

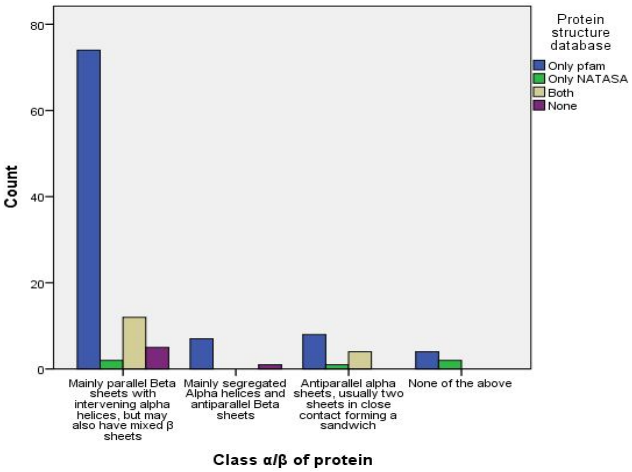

Figure 2. Class α/β of protein

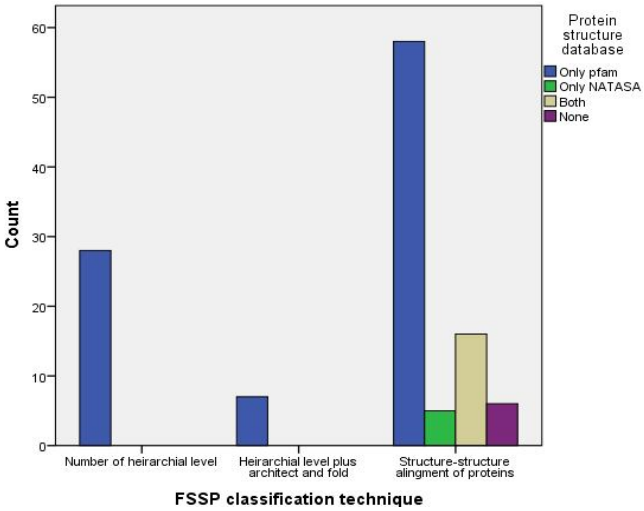

Figure 3. FSSP Classification Techniques

Table 9 FSSP Classification [**Test – 5**]

			Protein structure classification				Total
			Only Pfam	Only NATASA	Both	None	
FSSP Classification	Number of hierarchical level	Count	28	0	0	0	28
		% within FSSP Classification	100.0%	0.0%	0.0%	0.0%	100.0%
	Hierarchical level plus architect and fold	Count	7	0	0	0	7
		% within FSSP Classification	100.0%	0.0%	0.0%	0.0%	100.0%
	Structure-structure alignment of proteins	Count	58	5	16	6	85
		% within FSSP Classification	68.2%	5.9%	18.8%	7.1%	100.0%
Total		Count	93	5	16	6	120
		% within FSSP Classification	77.5%	4.2%	13.3%	5.0%	100.0%

Table 10. Chi-Square Tests of FSSP Classification **[Test-5]**

Chi-Square Tests			
	Value	Degree of Freedom	Asymptotic Significance (2-sided)
Pearson Chi-Square	14.345[a]	6	.026
Likelihood Ratio	21.695	6	.001
Linear-by-Linear Association	11.808	1	.001
N of Valid Cases	120		

Table 11. As Protein Evolves

As protein Evolves				
	Frequency	Percent	Valid Percent	Cumulative Percent
Both 2 & 3	9	7.5	7.5	7.5
Only 1	6	5.0	5.0	12.5
Only 4	18	15.0	15.0	27.5
All of the above	87	72.5	72.5	100.0
Total	120	100.0	100.0	

Out of 100 about 72.5% respondents feel that as protein evolves, it 'Retain function and specificity', 'Retain function but alter specificity', 'Change to a related function; or a similar function in a different metabolic context' and 'Change to a completely unrelated function [Table 11] (Chi Square test statistic = 147, p – value = 0.000< 0.05) [Test – 6].

Test 6

Test Statistics	
Chi-Square	147.000
Degree of Freedom	3
Asymptotic Significance	.000

Questionnaire Used

1. Which of this online protein structure classification database utilizes fold classification technique?
   - SCOP
   - DALI
   - MMDB
   - FFSP*

2. Class α/β of protein is comprised of:
   - Mainly parallel Beta sheets with intervening α helices, but may also have mixed β sheets.*
   - Mainly segregated Alpha helices and antiparallel Beta sheets.
   - Antiparallel alpha sheets, usually two sheets in close contact forming a sandwich.
   - None of the above.

3. RVP-net is online program for identifying:
   - Solvent accessibility.*
   - Protein folding.
   - Three-dimentional structure of protein.
   - Sequence of amino-acids in protein.

4. FSSP utilises which classification technique:
   - Number of heirarchial level.
   - Heirarchial level plus architect and fold.
   - Structure-structure alingment of proteins.*
   - Similar arrangment of secondary structure.

CONCLUSION

Out of 100 about 68.2% respondents feel that FSSP utilizes structure-structure alignment of proteins classification technique. Class α/β of protein is comprised of mainly parallel Beta sheets with intervening alpha helices, but may also have mixed β sheets. As protein evolves, it 'Retain function and specificity', 'Retain function but alter specificity', 'Change to a related function; or a similar function in a different metabolic context' and 'Change to a completely unrelated function'. In this paper we have analyzed the fold classification technique used by FSSP. It covers some basics of the protein structure such as domain, and folds databases. FSSP Classify protein domain structures using all-against-all comparison mechanism. Considering this fact can increase the sensitivity of protein function prediction approaches. Therefore, it should be likely to improve any method that is based on protein sequence evaluations by performing these comparisons on the domain level instead of incorporating the results obtained for all domains.

REFERENCES

1. Kumar, Manish, and Prakash, Ajay. (2014). A Statistical Analysis of CATH, SCOP and FSSP Databases. International Journal of Science and Research 3(5): 1586-1589.
2. Kumar, Manish, and Govil, Kapil. (2013). Protein Structure Comparison and Classifications into Domains. International Journal of Science and Research 2(10): 20-22.
3. P. Maji, & S. K. Pal, Rough-fuzzy pattern recognition: Applications in bioinformatics and medical imaging. Hoboken, N.J: John Wiley & Sons, 2012.
4. Kumar M (2013). Proposed Enhanced Proteins Classification Databases, International Journal for Pharmaceutical Research Scholars, 2(4): 160-163.
5. Holm L, Rosenstrom, P (2010). Dali server: conservation mapping in 3D, Nucleic Acids Research, 38: W545-W549.
6. Kumar, Manish, and Govil, Kapil (2013). The FSSP database: Fold Classification based on Structure–Structure alignment of Proteins. International Journal of Science and Research, 2(10): 23-25.
7. A.G. Murzin, S.E. Brenner, Hubbard, T., and C. Chotia, (1994) J. Mol. Biology, 247: 536-540.
8. J. Overington, M.S. Johnson, A. Sali, and T.L. Blundell, (1990) Proc. Royal Soc. Lond., B241, 132-145.
9. C.A. Orengo, T.P. Flores, J.M Thomton, and W.R.Taylor (1993) Protein Eng., 6: 485-500.
10. L. Holm and C. Sander (1994). Proteins, 19: 165-173.
11. Kumar, Manish, Kapil Govil, and Chanchal Chawla (2013). Comparison Between The Various Protein Classification Schemes. Journal of Engineering Computers & Applied Sciences 2(8): 59-61.
12. Liisa Holm, Chris Sander (1996). The FSSP Database: Fold Classification Based on Structure-Structure Alignment of Proteins. Nucl. Acids Res. 24 (1): 206-209.

*****