International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

# A Statistical Analysis of CATH, SCOP and FSSP Databases

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

[1*]PhD Scholar, Shri Venkateshwara University, Uttar Pradesh, India

[2]Assistant Professor, S.M College Chandausi, Uttar Pradesh. India

**Abstract:** ***There are so many databases available for protein structure classification like CATH, SCOP, and FSSP. Databases promote keyword search, sequence search, navigation, hierarchy classification, and external online links. However we have found that these databases are not consistent in determining which classes of proteins belong to the same family. Some proteins have been put in the same class despite the fact they have less robust relationship. It is essential for the available classification system to be compared and examine the classes to determine which proteins remain in the same family. Identification of the biochemical function of protein based on its structure and sequence poses several challenges in this post-genomic era. The shear amount of research being carried in the genomics field has resulted in most sequences characterized for a function, which is normally annotated as hypothetical. In this study, we have done a statistical analysis of protein structure classification databases based on its sequence, structure and function.***

**Keywords:** CATH, SCOP, FSSP, PDB, family, superfamily.

## 1. Introduction

Proteins are mainly classified in terms of structure and function [1]. To classify protein structure we have so many online classification databases available like CATH, SCOP, and FSSP. The classification schemes however evaluate proteins based on different properties [2]. Proteins can be classified in terms of structure [3]. In the SCOP database, the structures of proteins, as well as, their evolutionary relationships are used for the purpose of classification. Understanding and using proteins is a vital area of research within the ever-more important fields of biology and biotechnology. 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 [4]. As new proteins are sequenced and analysed and their description added to central databases, several problems arise. One is that the quantity data of information – already databases contain information on many thousands of additional or less related proteins. Another and maybe additional major problem is that the organization of all this data. As of these days, there is no universally accepted classification of proteins into totally different classes and subcategories. Correct categorizations are vital and can become even additional therefore, for other reasons. This problem is complicated by the fact that proteins often are divided into several distinct structures, connected by chains of unstructured amino acids. Usually these structures repeat within the totally different proteins, and may serve identical purpose. For this reason, attention late has been centered not solely on proteins as an entire however additionally on the various *protein structures*. It is important to note that all the protein databases use varying classification schemes to categorize protein into various domains [5]. As a result, it is possible to find two proteins that are placed into the same class in one database and classified into totally different categories in another. For instance, SCOP classifies proteins according to structure and evolutionary relationships while CATH classifies proteins in terms of class, domain architecture and topology, as well as homologous superfamilies. The classification schemes used in these databases are also different from that used in Pfam which categorizes proteins into families [6]. It is possible to get proteins showing close relationships in one classification scheme exhibiting less robust associations in another [7]. This is often done in circumstances wherever the alikeness at the sequence level is simply to borderline to be detected by any sequence-based similarity search program [8]. A logical beginning to the comparison of protein structures is a system of classifying these structures in order to easily identify and group similar folds and families. [9]. It would also be important to examine the classes and determine which groups of proteins remain in the same family [10].

## 2. Methods

### Comparison between CATH, SCOP and FSSP

CATH and SCOP are the two most comprehensive macromolecule structure classification resources. Each is in active development. The most recent release of SCOP (v1.75C) classifies 167,547 domains (59,514 PDB entries) compared with 173,536 (51,334 PDB entries) for CATH (v3.5) [**Table 1**]. Currently has 1313 folds classified compared with 1194 for SCOP, however comparisons at this level are problematic, as additional subjective criteria are utilized in fold classification. (**Figure 1**)

**Table 1:** Data on the content of the most current SCOP and CATH version

| | CATH 3.5 (September 2013) | SCOP 1.75 C (October 2013) | |
|---|---|---|---|
| Class | 4 | 7 | Class |
| Arch | 40 | -- | Arch |
| Topology | 1313 | 1194 | Topology |
| Superfamily | 2626 | 1961 | Superfamily |
| Family | 11,926 | 4493 | Family |
| Domains | 173,536 | 167,547 | Domains |
| PDBs | 51,334 | 59,514 | PDBs |
| PDBs in both | 48104 | | PDBs in both |

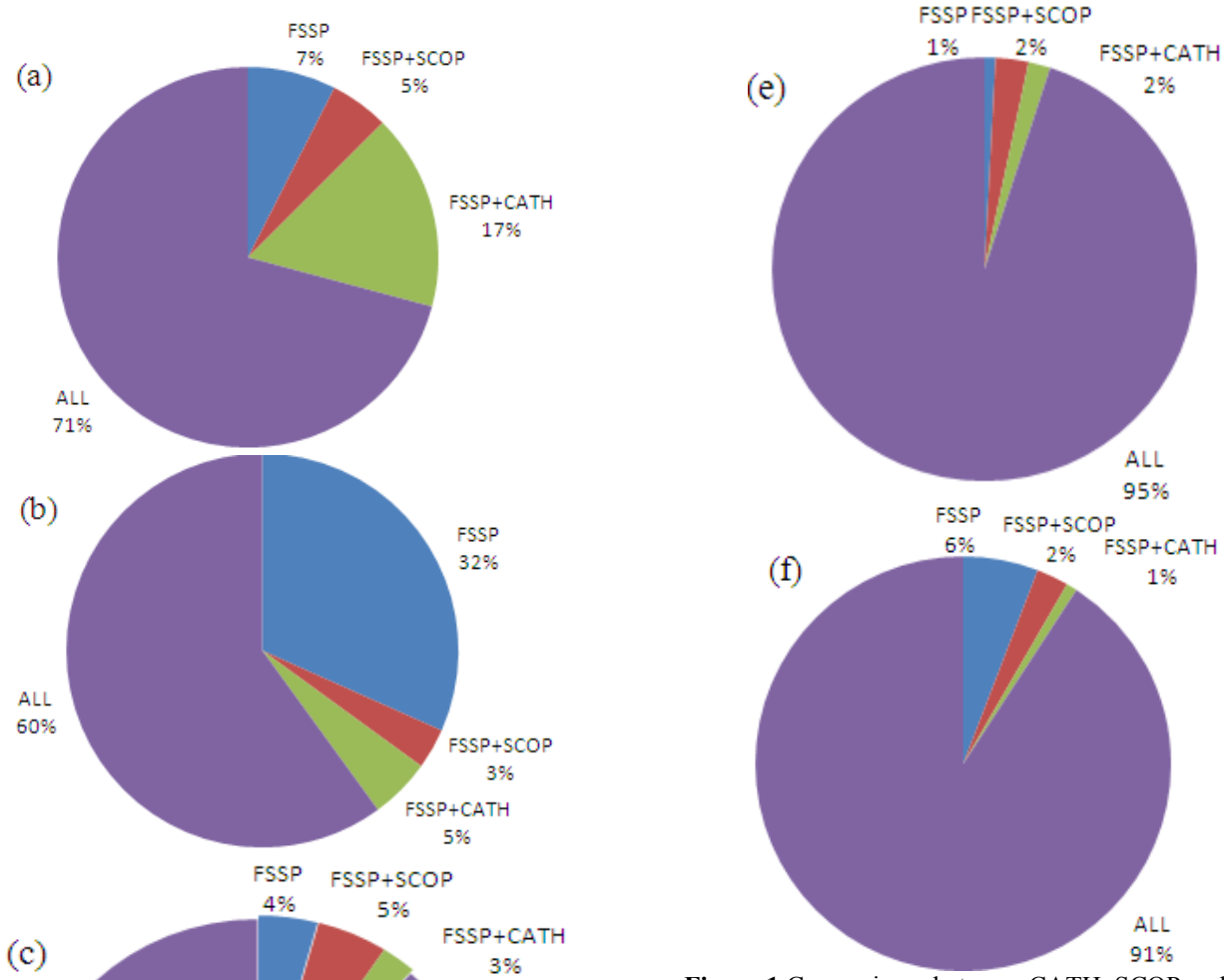
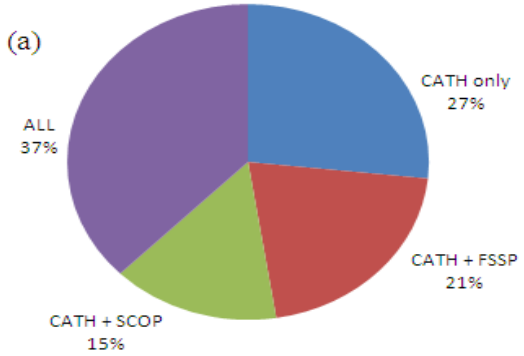



**Figure 1** Comparisons between CATH, SCOP and FSSP

**Comparing CATH pairwise matches to SCOP and FSSP.**

a) At the fold level, only 37% of the pairwise matches found in CATH are found in both SCOP and FSSP. Note that the number of CATH matches found in SCOP and FSSP is the same as the number of SCOP matches found in CATH and FSSP: the differing percentages reflect the total number of pairwise matches, which is much higher in CATH than in SCOP. A large percentage of these matches are found only in CATH **(Figure 2)**.

b) A smaller number of pairwise matches are found at the homology level, so the overall agreement between the databases is higher, and the number of pairwise matches confined solely to CATH is lower. SCOP (and FSSP) still includes additional CATH matches that the other database does not **(Figure 2)**.



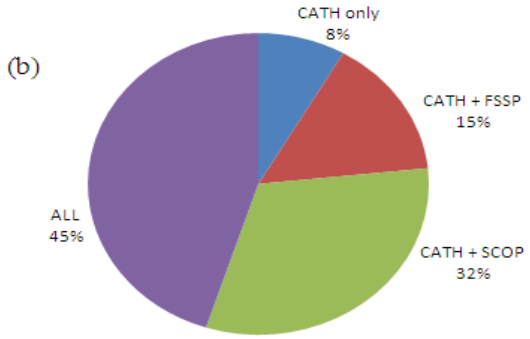



**Figure 2** Comparing CATH pairwise matches to SCOP and FSSP

**Comparing SCOP pairwise matches to CATH and FSSP**

a) At the fold level, almost two-thirds of the SCOP pairwise matches are also found in both FSSP and CATH. CATH agrees with a further 35% of the SCOP matches, whereas FSSP includes only an extra 13%. Only a small percentage of the pairwise matches are unique to SCOP **(Figure 3)**.
b) Fewer shared matches are found at the homology level in comparison to the fold level. Because of the difficulties inherent in assigning homology, there are a higher percentage of SCOP matches at this level that is not found in the other two databases **(Figure 3)**.

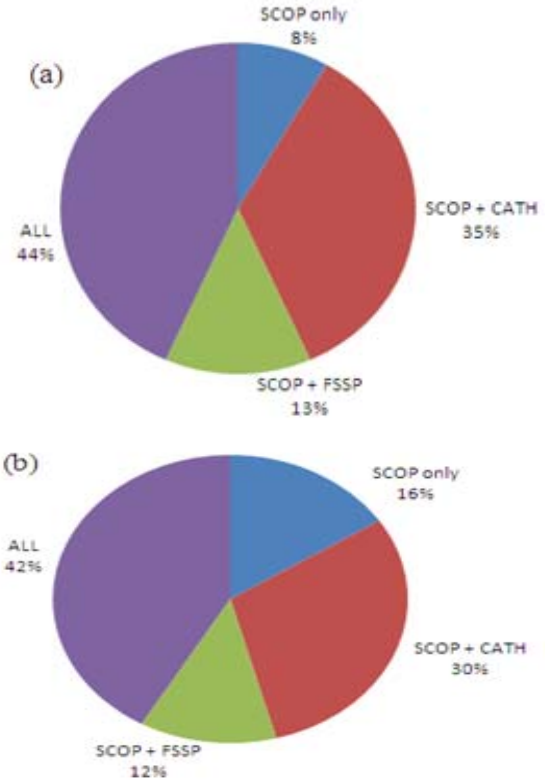



**Figure 3:** Comparing SCOP pairwise matches to CATH and FSSP

## 3. Results

Nearly 69.2% respondents agree that SCOP database classifies protein structures by a number of hierarchical levels to reflect both evolutionary and structural relationships **[Table 2].**

**Table 2**: Classification of Protein Structures

| | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|
| FFSP | 19 | 15.8 | 15.8 | 15.8 |
| SCOP | 83 | 69.2 | 69.2 | 85 |
| DALI | 12 | 10 | 10 | 95 |
| MMDB | 6 | 5 | 5 | 100 |
| Total | 120 | 100 | 100 | |

About 70.8% respondents agree that CATH protein database considers protein architecture as criteria for classification **[Table 3]**.

**Table 3:** Protein architecture as criteria for classification

| | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|
| SCOP | 17 | 14.2 | 14.2 | 14.2 |
| CATH | 85 | 70.8 | 70.8 | 85 |
| FFSP | 13 | 10.8 | 10.8 | 95.8 |
| MMDB | 5 | 4.2 | 4.2 | 100 |
| Total | 120 | 100 | 100 | |

The aim behind the comparison between different classification schemes is to identify members, which do not have robust relationship with the family (79.2%) **[Table 4, 5, 6].**

**Table 4**

| | Frequency | Percent | Valid Percent | Cumulative Percent |
|---|---|---|---|---|
| To develop perfect methods to identify folding pattern | 14 | 11.7 | 11.7 | 11.7 |
| To identify members which do not have robust relationship with the family | 95 | 79.2 | 79.2 | 90.8 |
| To identify perfect method to compare similarities among protein structures | 7 | 5.8 | 5.8 | 96.7 |
| To identify members which can be classified in different schemes at the same time | 4 | 3.3 | 3.3 | 100 |
| Total | 120 | 100 | 100 | |

**Table 5:** Comparison between different Classification Schemes **[Test – 7]**

| Comparison Between Different Classification Schemes | | | | |
|---|---|---|---|---|
| | Frequency | Percent | Valid Percent | Cumulative Percent |
| To develop perfect methods to identify folding pattern | 14 | 11.7 | 11.7 | 11.7 |
| To identify members which do not have robust relationship with the family | 95 | 79.2 | 79.2 | 90.8 |
| To identify perfect method to compare similarities among protein structures | 7 | 5.8 | 5.8 | 96.7 |
| To identify members which can be classified in different schemes at the same time | 4 | 3.3 | 3.3 | 100 |
| Total | 120 | 100 | 100 | |

**Table 6:** Test Statistics of comparison between different classifications schemes **[Test-7]**

| Test Statistics | |
|---|---|
| Chi-Square | 189.533[a] |
| df | 3 |
| Asymp. Sig. | .000 |

## 4. Discussion

Protein function in nature is a function of the inner dynamics of folds, the surface properties that give binding specificity, and the global architecture. This makes computational methods incorporating information on all the three levels to be way superior than sequence derived methods. The analysis has used a general approach for elucidating protein function in terms of both local and global structural similarity. Concerns have always been raised on whether the predicted structures can help in function prediction because the methods used mainly predict the protein core while the function of a protein depends on the surface properties. The methods analyzed, and the statistical findings show that some aspects of the protein core can be associated to function [11]. Structure based predictions complement predictions derived from sequence and that correct predictions can be made when no sequence similarity exists. The tests carried out have provided substantial support for the viability of structural genomics reducing the number of functional uncharacterized proteins.

## 5. Conclusion

Organization of protein structures according to folding pattern imposes a very useful logical structure on the entire in the Protein Data Bank. It affords a basis of structure-oriented information retrieval. Several databases derived from the PDB are built around classifications of protein structure. They offer useful features for exploring the protein structure world, including search for a keyword or sequence, navigation, among similar structures at various levels of the classification hierarchy, presentation of structure, and links to other sites.

## References

[1] Barton, Geoffrey. "SCOP: Structural classification of Proteins domains." Trends in Biochemical Science 19.13 (1994): 554-555. Print

[2] Kumar, Manish, Kapil Govil, and Chanchal Chawla. "Comparison Between The Various Protein Classification Schemes." *Journal of Engineering Computers & Applied Sciences* 2.8 (2013): 59-61.

[3] Getz, Gad, Michele Vendrusco, David Sach, and Eytan Domany. "The Automated Assigning of SCOP and CATH Protein Structure Classifications from FSSP." Protein Structure and Functions 46.4 (2002): 405-415. Print.

[4] Kumar, Manish, and Govil, Kapil. "The FSSP database: Fold Classification based on Structure--Structure alignment of Proteins" *International Journal of Science and Research (IJSR),* Volume 2, Issue 10, 23-25, (2013).

[5] Finkelstein, Alexei, and O. B. Ptitsym. Protein physics a course of lectures. Amsterdam: Academic Press, 2003. Print.

[6] McCall, Darralyn, David Stock, and Phillip Achey. Introduction to microbiology. Malden, MA: Blackwell Science, 2001. Print.

[7] Taylor, W. R., and AndrasAszodi. Protein, classification, geometry, symmetry, and topology. Bristol: Institute of Physics Pub., 2005. Print.

[8] Kumar, Manish, and Govil, Kapil. "Protein Structure Comparison and Classifications into Domains," *International Journal of Science and Research (IJSR)*, Volume 2, Issue 10, 20-22, (2013).

[9] Caroline Hadley and David T Jones. A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure* September 1999, **7**:1099–1112.

[10] Kumar M, Proposed Enhanced Proteins Classification Databases*, International Journal for Pharmaceutical Research Scholars,* 2013, 2(4), 160-163.

[11] Baker D, Sali A (2001) Protein structure prediction and structural genomics. *Science,* **294**: 93–96.

## Author Profile

**Manish Kumar** is pursuing PhD in Bioinformatics, from Shri Venkateshwara University, Uttar Pradesh. He has also completed M. Sc (Bioinformatics) and B.Sc (Biosciences) from Jamia Millia Islamia University, New Delhi. He has three years of teaching and research experience. He has been earlier associated with Guru Nanak Dev University, Amritsar, in area of Computer Aided Drug Design and Sequence Analysis. He has published number of research papers in national and international journals. He has also attended number of conferences, workshops and refresher course within India. His areas of interest are Computer Aided Drug Design, Sequence Analysis and Computational & Structural Biology.