Protein Based on its Sequence or Structure to an Existing Protein Family

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: ***To discover a new protein we have suggested some methods and also we have proposed how we can review of a protein to an existing protein family. We generally accepted that the three-dimensional structure of a peptide chain is determined by its amino acid sequence. Still, similar folds can have very different sequences. The ultimate task in sequence analysis is to predict the structure and function of a protein based on its sequence. When the protein of interest shares at least 30% amino acid identity with another protein, then these two proteins generally exhibit similar three-dimensional structure [1]. But when the proteins have the similar structure but divergent sequences, then the consensus sequence motifs can be used to assess the function of an unassigned sequences. Then these consensus motifs usually correspond to the residues interacting with the cofactors, substrate, or other proteins. In this paper we have also done statistical analysis (through SPSS) to classify identical proteins and Evolutionary Relationship and Structural similarity of the proteins.***

Keywords: Protein, PIR, SCOP, CATH, Pfam, Superfamilies.

1. Introduction

The structure of proteins refers to the bimolecular arrangement of molecules [2]. Relations between proteins sequence and structure are often analyzed by either determinative the sequence options of predefined structures [3], or by determinative structural options of preserved sequence regions. The proteins loops and their flanking regions were together found to be preserved to identical extent in an analysis of an oversized set of proteins [4]. Proteins with similar sequences adopt similar structure [5, 6]. However, similar structures can have less than 12% sequence similarity [7, 8-10]. Proteins are allotted to identical superfamily/family on condition that they share end-to-end sequence similarity, as well as common domain design (i.e. identical range, order, and all kinds of domains), and do not differ excessively in overall length (unless they are fragments or result from alternate conjunction or initiators). 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 [11]. Alternative major family databases are organized supported similarities of domain or motif regions alone, as in Pfam and PRINTS databases. There are also other databases that consist of mixtures of domain families and families of whole proteins, such as SCOP and TIGRFAMs [12]. However, in all of these, the protein to family relationship is not necessarily one-to-one, as in PIR superfamily/family, however also can be one-to-many. The PIR taxonomic category classification is that the only one that expressly includes this side, which may serve to discriminate between multi-domain proteins wherever functional variations are related to presence or absence of one or a lot of domains. An active site occurs within the tertiary (3-dimensional) or quaternary protein structure as a localized combination of amino acid side groups [13]. Families and superfamily classification frequently allow identification or probable function assignment for uncharacterized (hypothetical) sequences. To assure correct functional assignments, proteins identifications should be supported each world (Whole proteins, e.g. PIR superfamily) and native (domain and motif) sequence similarities [14].

2. Methods

We can discover a new protein with pattern recognition method. Such a method is built on the assumptions that characteristics of protein sequences or a protein structure can be used in identification of resembling traits in relevant proteins. Conserved protein's sequence regions are very significant to identify and study the function and structure of a new protein [15]. In pattern recognition method, the syntactic structure of the protein was recognized and then the algorithms were taken to detect the protein sequences. The protein was identified by studying the primitive pattern receptors. Then we made a comparison of our method with CMA algorithm to predict the residue pairs of protein's structure. Also the entropic and phylogenetic effects of the pattern recognition on the structural changes of proteins were observed.

3. Review of a Protein to an Existing Protein Family

We can review a protein to an existing protein family with the help of an expression system. These days because of the lack of post-transition modification machinery, finding and reviewing a protein in an existing protein family is an immense challenge. However, a few techniques are present which can make our task easier. The most convenient we found out of them was fusion protein technology. In a protein expression system, a subcomponent of the genes consisting of DNA and mRNA was translated into polypeptide chains. Then these chains were unfolded into proteins. Protein expression system made it easier to identify an existing protein family. Furthermore, the techniques that can help us make our task easier include reverse transcriptionase, artificial protein refolding and translocation. In fusion transcription, the protein was divided into sub-component in such a way that its measurement and abundance could be easily determined. With the help of these measurements, we also improved the protein expression system so that their polypeptide chains could be easily studied. It would also be important to examine the classes and determine which groups of proteins remain in the same family [16].

4. Better methods can be developed

Better methods can be developed by structurally identifying the protein sequence, DNA and RNA in computational biology. Homologous sequences within an existing constraint would be highly assumed to study the developmental stages. In addition, the multiple alignments of proteins' family and their domains would make it systematically possible to find out the residues in different locations. Homologous sequences targeted the particular protein group with proper genetic sequence. Then this protein group was biosynthesized to determine the exact results. Once the experiment was fully performed, we were in a position to draw the computational sequences of the protein and various developmental stages of homologous sequence. We could also observe that it were the homologous sequences, which made sure to target the appropriate proteins with their genetics.

5. Results

Regarding the responsible for the development of new protein with new functionality and structure, nearly 72.5% respondents feel that 'Gene duplication', 'Genetic Rearrangement', and 'Development of all new gene copies' are responsible for the development of new protein with new functionality and structure **[Table 1] [Figure 1]**.

Table 1: Responsible for the development of new protein

	Frequency	Percent	Valid Percent	Cumulative Percent
Gene Duplication	15	12.5	12.5	12.5
Genetic Rearrangement	12	10	10	22.5
Development of new gene copies	6	5	5	27.5
All of the above	87	72.5	72.5	100
Total	120	100	100	

Nearly 73.6% respondents who feel that 'Gene duplication', 'Genetic Rearrangement', and 'Development of all new gene copies' are responsible for the development of new protein with new functionality and structure agree that CATH protein database considers protein architecture as a criteria for classification **[Table 2, 3]** (Chi Square test statistic = 20.767, p – value = 0.014 < 0.05). **[Test-1]**

Table 2: Development of a new protein

			Protein Classification				Total
			SCOP	CATH	FFSP	MMDB	
New Protein Development	Gene Duplication	Count	6	6	5	0	17
		% within New Protein Development	38.3%	50.7%	11.0%	0.0%	100.0%
	Genetic Rearrangement	Count	0	8	3	5	16
		% within New Protein Development	9.3%	73.7%	10.3%	6.7%	100.0%
	Development of new gene copies	Count	4	6	0	4	10
		% within New Protein Development	21.3%	35.0%	0.0%	43.7%	100.0%
	All of the above	Count	4	68	11	8	91
		% within New Protein Development	13.8%	73.6%	10.3%	2.3%	100.0%
Total		Count	14	88	19	17	134
		% within New Protein Development	10.2%	65.2%	18.2%	6.4%	100.0%

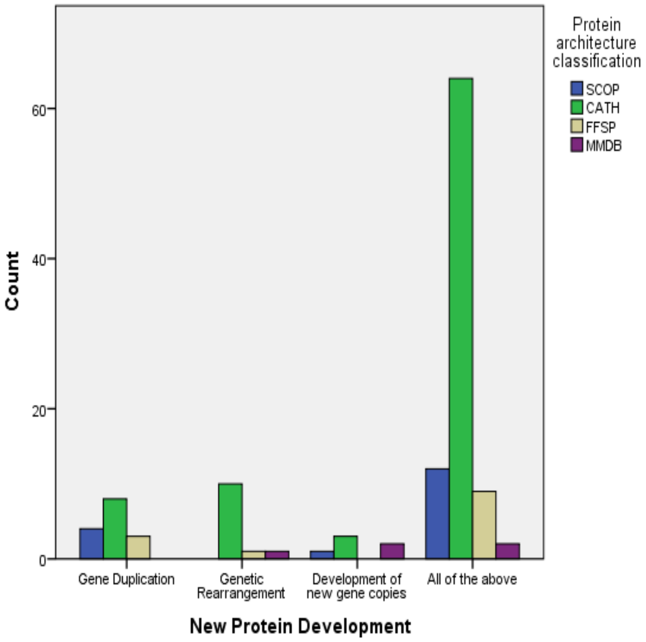


Figure 1: Development of a new protein

Table 3: Chi-Square Tests of New Protein Developments **(Test – 1)**

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	20.767^a	9	.014
Likelihood Ratio	15.853	9	.070
Linear-by-Linear Association	.085	1	.770
N of Valid Cases	120		

Nearly 56.7% respondents feel that 50% of evolutionary relationship between two proteins can be predicted easily if they have structural similarity **[Table 4, 5, 6] [Figure 2] [Test 2]**.

Table 4: Evolutionary Relationship and Structural similarity

	Frequency	*Percent*	*Valid Percent*	*Cumulative Percent*
90%	6	5	5	5
75%	15	12.5	12.5	17.5
50%	68	56.7	56.7	74.2
35%	31	25.8	25.8	100
Total	120	100	100	

Table 5: Evolutionary Relationship and Structural similarity **(Test-2)**

			As protein evolves				Total
			Both 2 & 3	Only 1	Only 4	All of the above	
Evolutionary Relationship and Structural similarity around	90%	Count	1	0	0	5	6
		% within Evolutionary Relationship	16.7%	0.0%	0.0%	83.3%	100.0%
	75%	Count	0	4	3	8	15
		% within Evolutionary Relationship	0.0%	26.7%	20.0%	53.3%	100.0%
	50%	Count	7	2	9	50	68
		% within Evolutionary Relationship	10.3%	2.9%	13.2%	73.5%	100.0%
	35%	Count	1	0	6	24	31
		% within Evolutionary Relationship	3.2%	0.0%	19.4%	77.4%	100.0%
Total		Count	9	6	18	87	120
		% within Evolutionary Relationship	7.5%	5.0%	15.0%	72.5%	100.0%

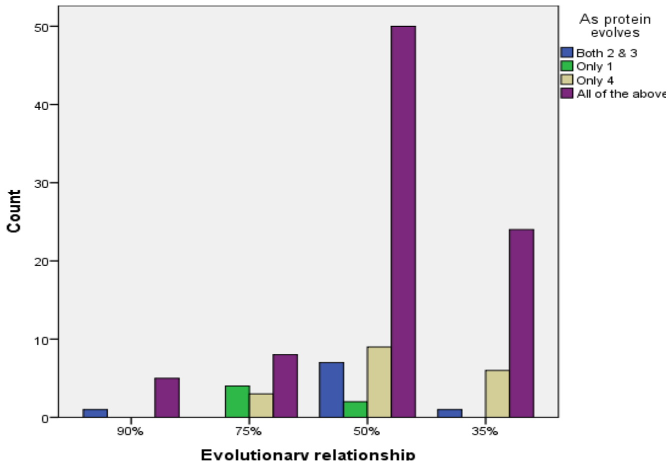


Figure 2 Evolutionary Relationship as protein evolves

Table 6: Chi-Square Tests of Evolutionary Relationship and Structural similarity **(Test-2)**

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	22.424[a]	9	.008
Likelihood Ratio	19.185	9	.024
Linear-by-Linear Association	1.740	1	.187
N of Valid Cases	120		

Regarding the classification of identical proteins, about 71.7% respondents agree that two proteins can be considered highly identical if they have similar sequences **[Table 7, 8, 9] [Figure 3] [Test 3]**.

Table 7: Classification of Identical Proteins

	Frequency	Percent	Valid Percent	Cumulative Percent
Similar sequences	86	71.7	71.7	71.7
Same biological action	11	9.2	9.2	80.8
Same type of active sites	19	15.8	15.8	96.7
All of the above	4	3.3	3.3	100
Total	120	100	100	

About 84.9% respondents said that two proteins can be considered highly identical if they have similar sequences **[Table 8]**. They also said that NATASA server is extensively used for identifying solvent accessibility (Chi Square test statistic=19.686, p – value =0.019<0.05) **[Test-3]**.

Table 8: Identical Proteins classification

			NATASA Server				Total
			Solvent accessibility	Protein function	Folding	Protein sequencing	
Identical Proteins classification	Similar sequences	Count	73	9	4	0	86
		% within Identical Proteins classification	84.9%	10.5%	4.7%	0.0%	100.0%
	Same biological action	Count	7	0	3	1	11
		% within Identical Proteins classification	63.6%	0.0%	27.3%	9.1%	100.0%
	Same type of active sites	Count	13	4	1	1	19
		% within Identical Proteins classification	68.4%	21.1%	5.3%	5.3%	100.0%
	All of the above	Count	2	1	1	0	4
		% within Identical Proteins classification	50.0%	25.0%	25.0%	0.0%	100.0%
Total		Count	95	14	9	2	120
		% within Identical Proteins classification	79.2%	11.7%	7.5%	1.7%	100.0%

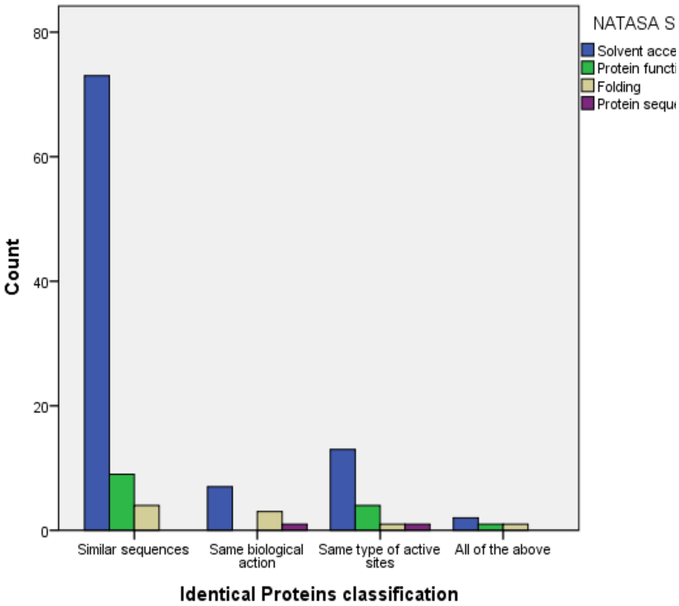


Figure 3: Identical Proteins Classification

Table 9: Test Statistics of Identical Proteins classification **(Test-3)**

Chi-Square Tests			
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	19.868^a	9	.019
Likelihood Ratio	17.098	9	.047
Linear-by-Linear Association	6.433	1	.011
N of Valid Cases	120		

6. Discussion

Conserved protein's sequence regions are very significant to identify and study the function and structure of a new protein. Sequence similarity, is exclusive in providing comprehensive and non-overlapping bunch of proteins sequences into a stratified order to replicate their biological process relationships. Proteins are allotted to identical superfamily/family on condition that they share end-to-end sequence similarity, as well as common domain design (i.e. identical range, order, and all kinds of domains), and do not differ excessively in overall length (unless they are fragments or result from alternate conjunction or initiators). Families and superfamily classification frequently allow identification or probable function assignment for uncharacterized (hypothetical) sequences. To assure correct functional assignments, proteins identifications should be supported each world (Whole proteins, e.g. PIR superfamily) and native (domain and motif) sequence similarities [11].

References

[1] Doolittle, R.F. 1986. *Of URFs and ORFs: A primer on how to analyse derived amino acid sequences,* University Science Books, Mill Valley, CA.

[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] Bystroff, C.; Simons, K.T.; Han, K.F.; Baker D. Local sequence structure correlations in proteins. *Curr. Opin. Biotechnol.*, 1996, **7**, 417-421.

[4] Liu, J.; Tan, H; Rost, B. Loopy proteins appear conserved in evolution. *Journal of Molecular Biology*, 2002, **322**, 53-64.

[5] Chothia, C.; Lesk, A.M. The relation between the divergence of sequence and structure in proteins. *EMBO J.*, 1986, **5**, 823-826.

[6] Doolittle, R.F. Similar amino acid sequences: chance or common ancestry? *Science*, 1981, **214**, 149-159.

[7] Hubbard, T.J.; Murzin, A.G.; Brener, S.E.; Chothia. SCOP: a structural classification of proteins database. C. *Nucleic Acids Research*, 1997, **25**, 236-239.

[8] Holm, L.; Sander, C. The FSSP database: fold classification based on structure-structure alignment of proteins. *Nucleic Acids Research*, 1996, **24**, 206-209.

[9] Brenner, S.E.; Chothia, C.; Hubbard, T.J.; Murzin A.G. Understanding protein structure: using SCOP for fold interpretation. *Methods Enzymol.,* 1996, **266**, 635-643.

[10] Rost, B. Protein structures sustain evolutionary drift. *Fold Des.*, 1997, **2**, S19-24.

[11] 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).

[12] Haft, D. H.; Loftus, B.J.; Richardson, D.L.; Yang, F.; Eisen, J.A.; Paulsen, I.T.; White, W.TIGRFAMs: "a protein family resource for the functional identification of proteins", *Nucleic Acids Research,* **29**, 41-43, 2001.

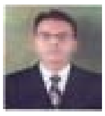
[13] Kumar, Manish, Govil, Kapil. "Protein Structure Comparison and Classifications into Domains," *International Journal of Science and Research*, Volume 2, Issue 10, 20-22, 2013.

[14] Wu, C.H.; Huang, H.; Yeh, L.L.; Barker, W.C., "Protein family classification and functional annotation", *Comp. Biol. Chem.*, **27**, 37-47, 2003.

[15] Sandhya R. Shenoy and B. Jayaram, "Proteins: Sequence to Structure and Function – Current Status", *Current Protein and Peptide Science*, *2010, 11, 498-514.*

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

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.