Protein Structure Comparison and Classifications into Domains

Manish Kumar[1], Kapil Govil[2]

[1]Department of Computer Science, Shri Venkateshwara University, Uttar Pradesh, India

[2]Assistant Professor, TMIMT, TMU, Moradabad, Uttar Pradesh, India

Abstract: ***Protein structure classification by using bioinformatics can involve sequence similarity searches, multiple sequence alignment, characterization of domains, involuntary protein fold recognition and constructing three-dimensional models to atomic element. In this paper, a set of protocols to classify protein structure and sequence is presented. Structure classification systems, as applied for example in the SCOP, CATH, and FSSP databases, clarify the relationship between protein folds and function and inform on the evolution of protein domains. Proteins fold into three-D structures, and protein structures are more preserved than protein sequences. Therefore, given a protein structure, it is necessary to search for geometrically comparable proteins through protein structure evaluation. This is mostly done in circumstances where the resemblance at the sequence level is too minimal to be detected by any sequence-based similarity search program. The object of protein structure comparison is to get the largest structural similarity between two structures (Reed, 2011 P 139).***

Keywords: protein structure, classification, sequence, dominion, and folds.

1. Introduction

Knowledge on proteins is essential for understanding cellular organization and function. For each new protein sequence, sequence-sequence and sequence-structure assessments are used to forecast its possible function. However, sequence-structure method of comparison is the most accurate in identifying structurally comparable proteins that lack sequence similarity. The study of three-dimensional (3D) protein structures is an efficient tool in molecular biology and cell biology.

2. Protein Fold

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. 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 (Maji, 2012 P 106).

3. Structural Domain

Proteins comprise of several structural domains. A protein domain is a preserved part of a given protein sequence and structure that can grow function, and exist individually of the rest of the protein chain. Each domain forms can be folded, and it forms a three-dimensional structure. Molecular evolution uses domains as a stepping stone and may be recombined in arrangements so as to create proteins with different functions (Liu, Wei, Li, & Global, 217 P 89). 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. One of the main challenges in this is the association of domain families with protein functions. On the other hand, a structural domain is a compact, globular substructure with added interactions within it than with the other proteins. Measures of local compactness in proteins have been used in many of the ancient methods of domain assignment and other recent ones (Maji, 2012 P 106).

4. Methods

4.1 Hierarchical Classification Framework

Structural Classification of Proteins (SCOP) was used as dataset that classifies protein structures hierarchically based on evolutionary relationships and the principles that direct their 3D structure. SCOP is a record of protein structural classification which delivers exhaustive and comprehensive description of the structural and evolutionary interactions of proteins, comprising all entries in the Protein Data Bank (PDB). The levels of protein structure are displayed below, and the protein domain is the unit of classification (Sperschneider, Sperschneider & Scheubert, L. 2008 P 47) (Table 1).

4.2 Domain Definition from Structural Coordinates

The PUU algorithm includes a harmonic model used to estimate inter domain dynamics. The original physical notion is that many rigid interfaces will occur within each domain, and loose interfaces will occur between domains. PUU algorithm is used to define domains in the FSSP domain database (Table 2).

4.3 Protein Architecture Prediction

The target proteins are first assigned CATH domains, using the standard protocol then they are assigned for the generation of the Gene3D resource. The acknowledged domain sequences in each superfamily are then scanned alongside the superfamily's family model library (Reed, 2011 P 139) (Table 3).

5. Results

Table 1: Hierarchical Classification Framework

	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	

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: Domain Definition from Structural Coordinates

	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
Total	120	100	100	

Almost 70.8% respondents feel that FSSP utilizes structure alignment of proteins classification technique.

Table 3: Protein Architecture Prediction

			Protein architecture classification				*Total*
			SCOP	*CATH*	*FFSP*	*MMDB*	
New Protein Develo-pment	Gene Duplication	Count	4	8	3	0	15
		% within New Protein Development	26.70%	53.30%	20.00%	0.00%	100%
	Genetic Rearrangement	Count	0	10	1	1	12
		% within New Protein Development	0.00%	83.30%	8.30%	8.30%	100.00%
	Development of new gene copies	Count	1	3	0	2	6
		% within New Protein Development	16.70%	50.00%	0.00%	33.30%	100.00%
	All of the above	Count	12	64	9	2	87
		% within New Protein Development	13.80%	73.60%	10.30%	2.30%	100.00%
Total		Count	17	85	13	5	120
		% within New Protein Development	14.20%	70.80%	10.80%	4.20%	100.00 %

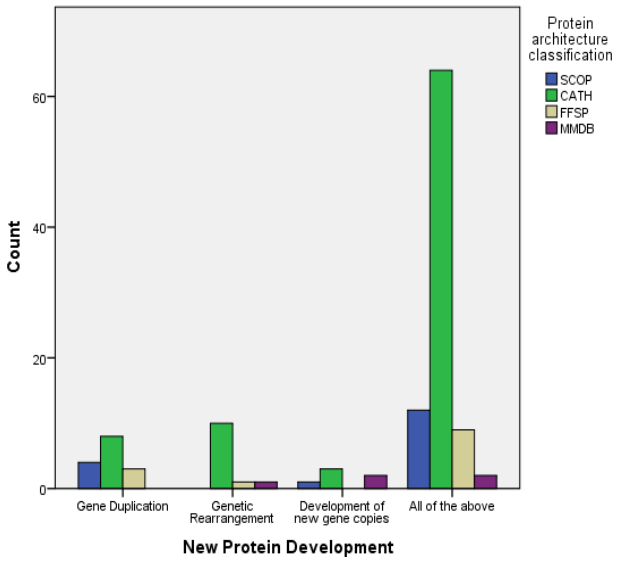

Figure 1: Protein Architecture Classification

Chi-Square Tests

	Value	*df*	*Asymp. Sig. (2-sided)*
Pearson Chi-Square	20.767[a]	9	0.014
Likelihood Ratio	15.853	9	0.07
Linear-by-Linear Association	0.085	1	0.77
N of Valid Cases	120		

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 (Chi Square test statistic = 20.767, p – value = 0.014 < 0.05).

6. Discussion

Proteins fold into three-D structures, and protein structures are more preserved than protein sequences. Therefore, given a protein structure, it is necessary to search for geometrically comparable proteins through protein structure evaluation. This is mostly done in circumstances where the resemblance at the sequence level is too minimal to be detected by any sequence-based similarity search program. The object of protein structure comparison is to get the largest structural similarity between two structures (Reed, 2011 P 139). Gene duplication, genetic rearrangement and development of a new gene copies are some extent responsible for development of new protein. From our data, one of the most common enzyme folds is the central α/β-barrel substrate binding domain. It is observed in various enzyme families completely catalyzing unrelated reactions.

7. Conclusion

This paper provides a guide to protein structure and function, with various aspects of bioinformatics. It covers some basics of the protein structure such as domain, folds databases, and the three-dimensional structure. As discussed in this paper, the relationship between computer science and biology is a natural one because of various reasons. To begin with, the phenomenal rate of biological data being produced

poses a challenge. This is because large amounts of data have to be stored, analyzed, and made accessible. Second, the biological data is presented statistically, and hence computation, is necessary. This applies precisely to the information on the building plans of proteins and the three-dimensional organization of their expression in the cell encoded by the DNA.

References

[1] W.K. Sung, Algorithms in bioinformatics: A practical introduction. Boca Raton: Chapman & Hall/CRC Press, 2010.
[2] P. Maji, & S. K. Pal, Rough-fuzzy pattern recognition: Applications in bioinformatics and medical imaging. Hoboken, N.J: John Wiley & Sons, 2012.
[3] V. Sperschneider, J. Sperschneider, & L. Scheubert, Bioinformatics: Problem solving paradigms. Berlin: Springer, 2008.
[4] D. Reed, A balanced introduction to computer science. Boston: Prentice Hall, 2011.
[5] L. A. Liu, D. Wei, Y. Li, & IGI Global, Computational biology and environmental sciences. Hershey, Pa: IGI Global (701 E. Chocolate Avenue, Hershey, Pennsylvania, 17033, USA, 2011.

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.