Journal of Proteomics & Bioinformatics - Open Access

www.omicsonline.com

Review Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000143
JPB/Vol.3 Issue 7

Classification of Protein Structural Classes using Isoluecine and Lysine Amino Acids

K. Manikandakumar[1]*, K. Gokul Raj[2], R. Srikumar[3] and S. Muthukumaran[4]

[1]*Department of Physics, Bharathidasan University College (W), Orathanadu - 614 625, Tanjavore, Tamil Nadu, India*
[2]*Department of Computer Science, Jamal Mohamed College, Tiruchirappalli – 620 020, Tamil Nadu, India*
[3]*Department of Microbiology, Bharathidasan University College (W), Orathanadu - 614 625, Tanjavore, Tamil Nadu, India*
[4]*Department of Physics, Arignar Anna Government Arts College, Attur – 636 121, Salem District, Tamil Nadu, India*

Abstract

Exploration of the structural organisation of proteins is essential for understanding of the mechanisms of protein folding and function, and for insights into protein evolution. Protein structure comparison can provide useful information on the biological function of a protein and can imply evolutionary relationships between proteins with low sequence similarity. Structural biology and structural genomics are expected to produce many three-dimensional protein structures in the near future. Each new structure raises questions about its function and evolution. Correct functional and evolutionary classification of a new structure is difficult for distantly related proteins and error-prone using simple statistical scores based on sequence or structure similarity. The objective of this study is to classifying the protein structural classes using key amino acid residues as Isoluecine (I) and Lysine (K), which is without any complicated mathematical or statistical representations. The classification of structural class is based on, grouping of 20 different amino acids with the comparison of isoluecine and lysine amino acid residues only. The combination of grouping of amino acids with the individual amino acid comparison will represent structural classes of the given protein sequences. This technique is tested over 40801 (inclusive of side chains, i.e., chain A, B, 1, 2, etc) proteins belonging to 67 different families randomly selected from All Alpha, All Beta, Alpha plus Beta and Alpha by Beta protein classes with the flat protein primary structure only. The classification rate is achieved with an accuracy of 52%. This method is alternative for experimental determination of structural classification from X-ray crystallography or NMR spectroscopy etc.

Keywords: Structural class; Protein sequence; Amino acid; Sequence analysis

Introduction

Proteins that have descended from the same ancestor retain memory of that ancestor through the sequence, structure, and function. To facilitate the understanding and access to available information for known protein structures, (Murzin et al., 1995) have constructed a Structural Classification of Proteins (SCOP) database. The concept of protein structural class was first proposed by (Levitt and Chothia, 1976). According to this concept, a globular protein can be assigned to one of the four structural classes, i.e. all alpha, all beta, alpha plus beta and alpha by beta. The all alpha and all beta proteins were defined to be composed of almost entirely alpha helices and beta strands, respectively. The alpha plus beta proteins were defined to be composed of separate segments of alpha helices and beta strands, whereas the alpha by beta proteins were defined to be composed of mixed segments of alpha helices and beta strands. Because there were very few proteins whose crystallographic structures were known in 1976, this definition of the structural classes was derived from a quite small database, i.e. 31 globular proteins only. Now the three-dimensional structures of about 50000 proteins are known. However, the definition of the protein structural classes is still accepted by the protein research community even to date. Since 1976, many definitions of the structural classes have been proposed along with the work of (Levitt and Chothia, 1976). In these studies, the classification schemes are based on the primary structure content of proteins. Only few researchers provide quantitative criteria to distinguish between the alpha plus beta and alpha by beta proteins (Chou, 1995). The mixed alpha-beta proteins should be classified as alpha by beta if the percentage of parallel strands is greater than 60%, otherwise, if the percentage of anti-parallel strands is greater than 60%, it should be an alpha plus beta protein. On the other hand, to separate the alpha plus beta and alpha by beta proteins. (Manikandakumar et al., 2009) analyzed the matrix frequency analysis of *Oryza Sativa* (japonica cultivar - group) complete genomes. The mixed alpha-beta protein (domain) is mapped onto a point in the two-dimensional plane spanned by the alternation score and percentage of parallel strands. Based on the distribution of the mapping points, a quantitative criterion was proposed to classify the mixed alpha-beta proteins (domains) into the alpha plus beta and alpha by beta classes. Obviously, the threshold of 60% adopted by (Chou, 1995) is a simple majority only. The separation between the alpha plus beta and alpha by beta classes based on the criterion is more objective and hence more reliable.

Investigation of the structural organization of proteins is significant perceptive of the mechanisms of protein folding and function, and for insights into protein evolution. Direct determination of protein structures (Chandonia and Brenner, 2006; Phillips, 1966) and comparative sequence analysis (Andreeva et al., 2004) indicate that proteins have a modular structure, i.e., a polypeptide chain may consist of several regions that can fold independently and be inherited as discrete sequence fragments, which recombine to produce novel sequence and spatial architectures. This level of protein organisation is called domain (Wetlaufer, 1973; Rossmann and Liljas, 1974; Doolittle, 1995). The concept of a structural domain of a protein may be associated with its physical compactness and thermodynamically

***Corresponding author:** K. Manikandakumar, Department of Physics, Bharathidasan University College (W), Orathanadu - 614 625, Tanjavore, Tamil Nadu, India, E-mail: bioinfokm@gmail.com

Received June 12, 2010; **Accepted** June 29, 2010; **Published** June 29, 2010

Citation: Manikandakumar K, Raj KG, Srikumar R, Muthukumaran S (2010) Classification of Protein Structural Classes using Isoluecine and Lysine Amino Acids. J Proteomics Bioinform 3: 221-229. doi:10.4172/jpb.1000143

Copyright: © 2010 Manikandakumar K, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

stability when excised or expressed independently of other domains (Veretnik et al., 2004). A formal definition of a domain, however, is still an outstanding problem. Protein structure comparison can provide useful information on the biological function of a protein (Holm and Sander, 1993) and can imply evolutionary relationships between proteins with low sequence similarity. This information is crucial in the identification of new protein folds and understanding the organisation of the known universe of protein structures. A topological algorithm for identification of structural domains of proteins developed by (Emmert-Streib and Mushegian, 2007). This method benefits from the vast collection of sequences from diverse organisms and high sensitivity of database search and protein sequence alignment. It relies on sequence similarity and thus is not applicable when the homologous sequences are not known; and, that the problem of defining the exact borders of sequence domains is itself difficult. (Jain and Hirst, 2007) by analyses for study of protein structural similarity, classification and this work is to perform protein structure analysis and comparison using structural descriptors. SCOP is a curated database which aims to provide a comprehensive description of the structural and evolutionary relationships between all protein structures. The principal levels in the SCOP hierarchy are class, fold, superfamily and family. The Automated Protein Subfamily Identification and Classification by (Brown et al., 2007) enables improved specificity of functional inference and facilitates prediction of functional shifts in new sequences. (Choi and Kim, 2006) address by evolution of protein structural classes and protein sequence families. They were report on the development of an efficient sub-graph mining technique and its application to finding characteristic sub-structural patterns within protein structural families. (Dietmann and Holm, 2001) worked for Identification of homology in protein structure classification. They were present an accurate numerical method for the identification of evolutionary relationships (homology). It is based on the principle that natural selection maintains structural and functional continuity within a diverging protein family. The Support Vector Machine algorithm was used to classify proteins from different families under the binary classification scheme. Several attempts have been made to identify the structural domains of proteins.

The previous works are developed with complex mathematical and statistical representation such as neural networks, phylogenetic analysis etc. The authoritative databases of structural domains, such as SCOP and CATH are populated in that manner. However, SCOP and CATH are not giving the undetermined protein sequences structural classification. The classification of structural classes for homologous protein sequences is not available in present days with simple representation. Now, we have suggested a simple theoretical method for classification of structural classes of proteins using few key amino acid residues and it is not incorporated complex mathematical or statistical representations.

Materials and Methods

Rule for classification of structural classes

The following rules are governed by classification of structural classes:

The all alpha structural class:

a) Hydrophobic group have higher than polar and charged group amino acid residues

b) The amino acid content of lysine(K) is greater than isoluecine (I)

i.e., All Alpha class = Hydrophobichigh when I>K

The all beta structural class:

a) Polar group have higher than hydrophobic and charged group amino acid residues

b) The amino acid content of lysine (K) is less than isoluecine (I)

i.e., All Beta class = Polarhigh when I<K

The alpha plus Beta structural class:

a) Polar group have higher than hydrophobic and charged group amino acid residues

b) The amino acid content of lysine (K) is greater than isoluecine (I)

i.e., Alpha plus Beta class = Polarhigh when I>K

The Alpha by Beta structural class:

a) Hydrophobic group have higher than polar and charged group amino acid residues

b) The amino acid content of lysine (K) is less than isoluecine (I)

i.e., Alpha by Beta class = Hydrophobichigh when I<K

Details of data used

There are four major classes of Structural Classification of Proteins (SCOP) namely, all alpha, all beta, alpha plus beta and alpha by alpha. In this study, we have classified for the structural classes for protein sequences of different families. We have taken some representative protein families and downloaded the available protein sequences belonging to the family in FASTA format from the SCOP option (http://www.rcsb.org/pdb/browse /browse.do?t=11&useMenu=no) of the Protein Data Bank (PDB) web site (http://www.rcsb.org/pdb). These sequences of different protein families are used as classifying for structural classes for the corresponding sequences.

Method of study

The proposed method is studied by the following two methods.

1) Analysis of Individual amino acid residues (separated in 20 residues)

2) Analysis of grouping of amino acid residues

Individual amino acid residues

We analysed, 20 different amino acid residues for protein sequences without any physico- chemical properties. However, all residues except Isoluecine (I) and Lysine (K) have equally take part with other neighbor residues. For example, alanine is greater than cystine residue in all alpha class, it is same for all beta and other classes also. The arginine amino acid residue is either greater or less than the serine residue in all alpha class, it is same as in all beta and other classes. But, isoluecine and lysine amino acid residues only differentiating from all classes. Therefore, these two residues have been classified for major role of classification of structural classes in protein sequences.

Grouping of amino acids

In order to find out the structural classification of protein sequences, we have grouped the amino acid residues mainly into three groups namely, Hydrophobic, Polar and Charged amino acids and denoted them as H, P and C respectively. The list of amino acids selected in each of these three groups is provided below:

1) Hydrophobic (H): Ala (A), Phe (F), Ile (I), Leu (L), Met (M), Pro (P), Val (V)

Review Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000143
JPB/Vol.3 Issue 7

2) Polar (P): Cys (C), Gly (G), His (H), Asn (N), Gln (Q), Ser (S), Thr (T), Trp (W), Tyr (Y)

3) Charged (C): Asp (D), Glu (E), Lys (K), Arg (R)

Further, we have excluded the unknown residues for our calculation denoted as 'X' and '?' which represent 'any' and 'unknown' residues respectively, from our downloaded data. We have eliminated on less than 20 amino acid residues sequences for further calculations.

Materials for calculations

The grouping of amino acid residues and separated for individual residues of the sequences are using computer C programming language. We used Microsoft Excel packages for calculations and generating figures. Then we have manually classified for all alpha, all beta, alpha plus beta and alpha by beta structural classes. However, we take only globular proteins for identification of structural classifications. This proposed method, did not classified structural classes when the content of isoluecine and lysine residues is equal. Suppose the charged group is higher than other two groups as hydrophobic or polar, the proposed method did not classified structural classes. The proposed structural classification method is compared with structures are determined from experimental techniques such as X-ray crystallography and NMR spectroscopy etc and these structural classes were deposited on PDB database of SCOP option. Therefore, we did not compare any other methods. We ignore less than 20 amino acid residues, equal content of isoluecine and lysine amino acids and error reading sequences for structural classification calculations.

Results and Discussions

Analysis of individual amino acid residues

The key residues of isoluecine (I) and lysine (K): The hydrophobic and charged groups of amino acid residues have used for identification of structural classes. The isoluecine residue is coming under hydrophobic group. The lysine residue is coming under charged group. These two residues are playing on major role for classification of structural classes. We are analysed and identified from all sequences, either the content of isoluecine residue is greater than content of lysine residue or lysine is less than isoluecine. Only few sequences are having same content of isoluecine and lysine residues. The isoluecine and lysine residues having the contents for structural classes, from all alpha proteins families are having 66.76% sequences are secured lysine is greater than isoluecine. From all beta proteins families, 40.48% sequences are having lysine is less than isoluecine, from alpha plus beta proteins families, 58.68% sequences are having isoluecine is greater than lysine and alpha by beta proteins families, 47.43% sequences are having lysine is less than isoluecine amino acid residues. In 53.62% sequences are identified for the content of lysine is greater than isoluecine amino acid residues from all alpha and all beta structural classes. In 53.05% sequences are identified for the content of lysine is less than isoluecine amino acid residues from alpha plus beta and alpha by beta structural classes. Finally, from the above four classes 53.33% sequences are identified for content of isoluecine and lysine is important for classification of structural classes.

Analysis of grouping of amino acids

The role of grouping of amino acids: Amino acid residue groups such as hydrophobic, polar and charged group are besides identification of structural classes. The charged group does not playing directly, but it is obliquely in performance. We identified from these three groups, either hydrophobic or polar only play for classification of structural classes. So, which one is higher than other one it may representing a key group for study of this method. Then, we compare the content of isoluecine and lysine amino acid residues for which one is greater than other one. i.e., either isoluecine is greater or lysine is greater. Finally, we have identified for the given protein sequence is obtained in which class.

Suppose we identify the hydrophobic group is higher than other groups, the result may be either all alpha class or alpha by beta class. The polar group is higher than other two groups, the result may be either all beta class or alpha plus beta class. The all alpha protein families, 75.21% sequences are having hydrophobic group of amino acid residues, from all beta protein families, 59.16% are having polar group, from alpha plus beta protein families are secured 45.71% sequences are having polar group and alpha by beta protein families are having 87.05% sequences are secured hydrophobic group of amino acid residues. From these analyses, 81.13% sequences are identified the content of hydrophobic group amino acid residue is higher than polar group for all alpha and alpha by beta structural class protein sequences. 52.43% sequences are identified the content of polar group amino acid residue is higher than hydrophobic group for all beta and alpha plus beta structural class protein sequences. Finally, from the above four classes 66.78% sequences are identified from the content of group of amino acid residue is play for classification of structural classes.

Graphical identification of structural class: From Figure 1, we have identified for the given homologous sequence is obtained in which class. Ex. The All Alpha figure shows, we have visible I residue line is less than K residue line. As well as the group of amino acid shows, the hydrophobic group is higher than other two groups. Therefore, we will confirm the given protein sequences is coming under all alpha structural class.

All Beta figure shows, we have visible I residue line is greater than K residue line. As well as the group of amino acid shows, the polar group is higher than other two groups. Therefore, we will confirm the given protein sequences is coming under all beta structural class.

Alpha plus Beta figure shows, we have visible I residue line is greater than K residue line. As well as the group of amino acid shows, the hydrophobic group is higher than other two groups. Therefore, we will confirm the given protein sequences is coming under alpha plus beta structural class.

Alpha by Beta figure shows, we have visible I residue line is less than K residue line. As well as the group of amino acid shows, the polar group is higher than other two groups. Therefore, we will confirm the given protein sequences is coming under alpha by beta structural class.

The last graph of Figure 1 is shows, the role of Isoluecine and lysine is play for the different structural classes.

Analysis of protein families

Among 67 protein families, 32 families of protein sequences are having more than 50% identity for hydrophobic group. In 19 families are having more than 50% identity for polar group. In 22 families are having more than 50% identity of lysine is greater than isoluecine amino acid contents. In 18 families are having lysine is less than isoluecine conetens. For example, Globin-like protein family, 1160 sequences out of 1338 are having high hydrophobic group and 1163 sequences are having lysine is greater than the isoluecine content.

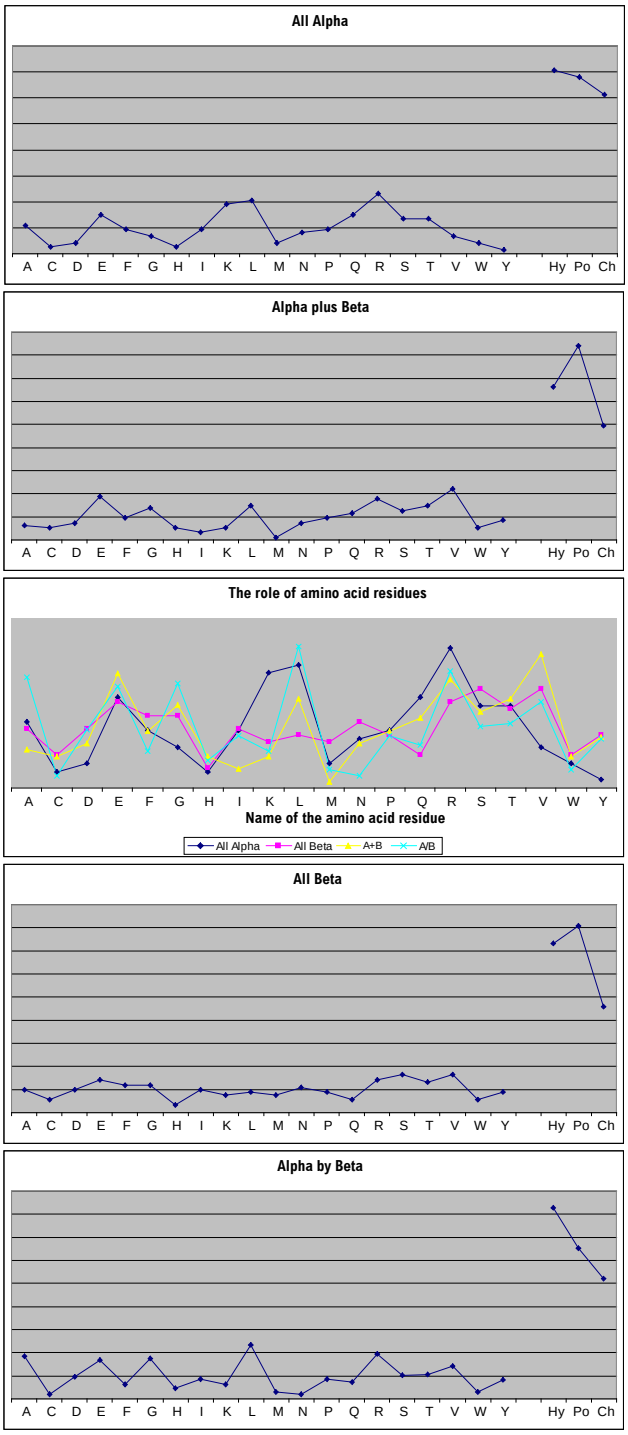

Figure 1: The representation for individual amino acid residues for different classes.

All alpha protein families: In this method, from 19 all alpha protein families, 12 families of protein sequences are having more than 50% identity for hydrophobic group with lysine is greater than isoluecine amino acid contents except Alpha-Alpha superhelix, Alpha/Alpha toroid, Cytochrome P450, Ferritin-like, Heme-dependent peroxidases, Lambda repressor-like DNA-binding domains, Phospholipase A2, PLA2 protein families. 12 protein families are having more than 50% identity for lysine is greater than isoluecine contents except Alpha-Alpha superhelix, Alpha/Alpha toroid, Cytochrome P450, Ferritin-like, Lambda repressor-like DNA-binding domains, Nuclear receptor ligand-binding domain, SAM domain-like protein families. In 18 protein families are having more than 50% identify for hydrophobic group except Phospholipase A2, PLA2 protein family.

All beta protein families: From 19 All Beta protein families, 6-bladed beta-propeller, Acid proteases, Beta-clip, Concanavalin A-like lectins/glucanases, Galactose-binding domain-like, Glycosyl hydrolase domain, Nucleoplasmin-like/VP, Supersandwich protein families are having more than 50% identity for polar group with lysine is less than isoluecine conetents. 9 protein families except Beta-Trefoil, Carbonic anhydrase, Cupredoxin-like, Double-stranded beta-helix, Immunoglobulin-like beta-sandwich, Lipocalins, OB-fold, SH3-like barrel, Streptavidin-like, Trypsin-like serine proteases are having more than 50% identity for lysine is less than isoluecine contents. In 12 protein families of sequences except Acid proteases, Beta-clip, Cupredoxin-like, Double-stranded beta-helix, OB-fold, Prealbumin-like, SH3-like barrel are having more than 50% identity for polar group.

Alpha plus beta protein families: From 14 Alpha plus Beta protein families, Beta-Grasp, Ferredoxin-like, MHC antigen-recognition domain, Nucleotidyltransferase, Protein kinase-like, RNase A-like, SH2-like, TBP-like, Thymidylate synthase/dCMP hydroxymethylase families are having more than 50% identity of polar group with lysine is greater than isoluecine contents. 10 protein families except Cysteine proteinases, Microbial ribonucleases, Thymidylate synthase/dCMP hydroxymethylase, Zincin-like Zincin-like protein families are having more than 50% identity of lysine is greater than isoluecine contents. In 7 protein families except Beta-Grasp, Ferredoxin-like, Glyceraldehyde-3-phosphate dehydrogenase-like, Nucleotidyltransferase, C-terminal domain, Protein kinase-like, TBP-like, Thymidylate synthase/dCMP hydroxymethylase protein families are having more than 50% identity of polar group.

Alpha by beta protein families: From 15 Alpha by Beta protein families, Dihydrofolate reductases, Flavodoxin-like, UDP-Glycosyltransferase/glycogen phosphorylase, Alpha/beta-Hydrolases, Phosphorylase/hydrolase-like, PLP-dependent transferases, Subtilisin-like, P-loop containing nucleoside triphosphate hydrolases protein families are having more than 50% identity of hydrophobic group with lysine is less than isoluecine contents. 9 protein families except FAD/NAD(P)-binding domain, NAD(P)-binding Rossmann-fold domains, Periplasmic binding protein-like II, S-adenosyl-L-methionine-dependent methyltransferases, Thioredoxin fold, P-loop containing nucleoside triphosphate hydrolases protein families are having more than 50% identity of lysine is less than isoluecine contents, 14 protein families except Subtilisin-like protein families are having more than 50% identity of hydrophobic group.

Classification of all alpha structural class: From Table 2 and Figure 2, we critically analysed All Alpha proteins, we have taking 10775 sequences for this study. In 971 (9.01%) sequences are identified for less than 20 residues sequences out of 10775 sequences. So, we eliminate these sequences. Finally, we analysed 9804 all alpha sequences. Of the 9804, 7239 (73.84%) sequences are identified for hydrophobic group compare than other groups. Out of 9804, 6372 (64.99%) sequences are identified for the content of lysine is greater than isoluecine amino acid. Of the 7239 sequences, 4681 (64.66%) sequences are identified for hydrophobic group with content of lysine is greater than isoluecine residues. In 559 (5.70%) sequences are identified for the content of isoluecine is equal to lysine amino

Review Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000143
JPB/Vol.3 Issue 7

acid residue out of 9804 sequences. In 166 (1.69%) sequences are error-reading out of 9804 sequences.

Classification of all beta structural class: From All Beta proteins, we have taking 14423 sequences for this study. In 1120 (7.77%) sequences are identified for less than 20 residues sequences. Finally, we analysed 13303 all beta sequences. In 7460 (56.08%) sequences out of 13303 sequences are identified for polar group compare than other two groups. In 5142 (38.65%) sequences are identified for the content of lysine is less than isoluecine amino acid out of 13303 sequences. In 2810 (37.67%) sequences are identified for polar group with the content of lysine is less than isoluecine amino acid residues for out of 7460 sequences. In 469 (3.53%) sequences are identified for the content of isoluecine is equal to lysine amino acid residues for out of 13303 sequences. In 173 (1.30%) sequences are error-reading sequences out of 13303 sequences (Table 2 and Figure 2).

Identification of alpha plus beta structural classification: From Alpha plus Beta structural proteins, we have taking 8755 sequences for this study. In 975 (11.14%) sequences are identified for less than 20 amino acid residues sequences. Finally, we analysed 7780 alpha plus beta sequences. In 3106 (39.92%) sequences out of 7780 sequences are identified for polar group compare than other two groups. In 4420 (56.81%) sequences are identified for the content of lysine is greater than the isoluecine amino acid for out of 7780 sequences. In 1812 (58.34%) sequences are identified for polar group with the content of lysine is greater than isoluecine amino acid residues out of 3106 sequences. In 1025 (13.17%) sequences are identified for the content of isoluecine is equal to the content of lysine amino acid residue for out of 7780 sequences. In 88 (1.13%) sequences are error-reading sequences out of 7780 sequences (Table 2 and Figure 2).

Identification of alpha by beta structural class: From Alpha by Beta structural proteins, we have taking 10312 sequences for this study. In 398 (3.86%) sequences are identified for less than 20 amino acid residues. Finally, we analysed 9914 alpha by beta sequences. In 8578 (86.52%) sequences out of 9914 are identified for hydrophobic group compare than other two groups. In 4679, (47.20%) sequences are identified for the content of lysine is greater than isoluecine amino acid residues for out of 9914 sequences. In 3952 (46.07%) sequences are identified for hydrophobic group with the content of lysine is less than isoluecine amino acid residues for out of 8578 sequences. In 433 (4.37%) sequences are identified for the content of isoluecine is equal to lysine amino acid residue for out of 9914 sequences. In 166 (1.67%) sequences are error-reading sequences out of 9914 sequences (Table 2 and Figure 2).

Discussions

Analysis of individual protein families

All alpha protein families: The 4-helical cytokines protein family is having 204 sequences. In 115 sequences are identified for hydrophobic group and 161 sequences are identified for lysine is greater than isoluecine amino acid residue content. In 90 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. The Alpha-Alpha superhelix protein family is having 857 sequences. 510 sequences are identified for hydrophobic group and 348 sequences are identified for lysine is greater than isoluecine amino acid residue content. In 226 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. The Alpha/Alpha toroid protein family is having 455 sequences. 226 sequences are identified for hydrophobic group and 151 sequences are identified for lysine is greater than isoluecine amino acid residue content. 26 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, we suggest this family of protein sequences is not maintaining their class. The Cytochrome c protein family is having 594 sequences. 372 sequences are identified for hydrophobic group and 480 sequences are identified for lysine is greater than isoluecine amino acid residue content. 301 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. The Cytochrome P450 protein family is having 171 sequences. 170 sequences are identified for hydrophobic group and 71 sequences are identified for lysine is greater than isoluecine amino acid residue content. 71 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is not maintaining their class. The DNA/RNA-binding 3-helical bundle protein family is having 1746 sequences. 1003 sequences are identified for hydrophobic group and 899 sequences are identified for lysine is greater than isoluecine amino acid residue content. 602 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. EF Hand-like protein family is having 607 sequences. 444 sequences are identified for hydrophobic group and 411 sequences are identified for lysine is greater than isoluecine amino acid residue content. 301 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. Ferritin-like protein family is having 586 sequences. 513

S. No.	Name of the Structural Class	No. of Pro. Chains	Total No. AA	H (%)	P (%)	C %)
1.	All Alpha	10775	2210235	39.66	36.33	24.01
2.	All Beta	14423	3412650	37.84	41.17	20.99
3.	Alpha plus Beta	8755	1893591	37.44	39.04	23.52
4.	Alpha by Beta	10312	3258422	41.38	35.54	23.09

Table 1: Representation of grouping of amino acid residues for different classes.

Sl. No.	Proteins	No. of protein chains	No. of Hy/Po seq.	No. of I<K & I>K seq.	No. of Hy/Po with I<K & I>K seq.	No. of I=K seq.	No. of error seq.
1.	All Alpha	9804	7239	6372	4681	559	166
2.	All Beta	13303	7460	5142	2810	469	173
3.	Alpha plus Beta	7780	3106	4420	1812	1016	88
4.	Alpha by Beta	9914	8578	4679	3952	433	166
		40801	26383	20613	13255	2477	593

(Hy - hydrophobic group, Po - polar group)

Table 2: Representation of grouping of amino acids and individual amino acid residues calculation of different classes.

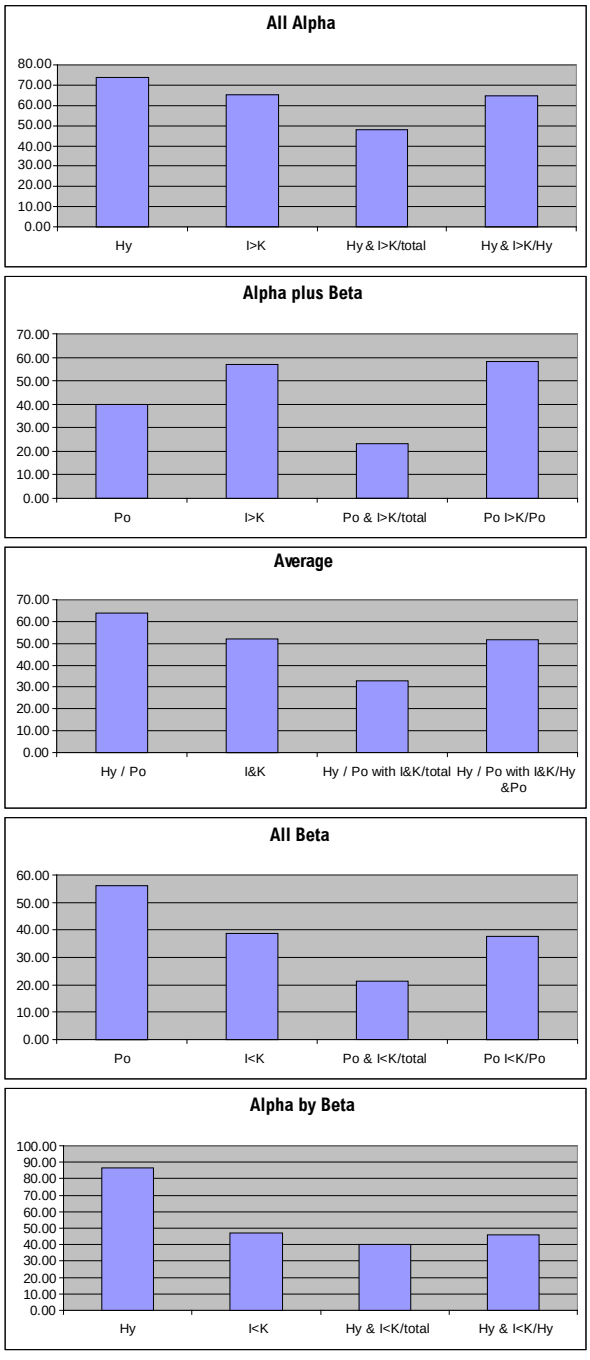

Figure 2: The representation for group of amino acid with individual amino acid residues in different structural classes.

sequences are identified for hydrophobic group and 266 sequences are identified for lysine is greater than isoluecine amino acid residue content. 210 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. Four-helical up-and-down bundle protein family is having 395 sequences. 312 sequences are identified for hydrophobic group and 311 sequences are identified for lysine is greater than isoluecine amino acid residue content. 257 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. Globin-like protein family is having 1395 sequences. 1338 sequences are identified for hydrophobic group and 1163 sequences are identified for lysine is greater than isoluecine amino acid residue content. 1160 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. Glutathione S-transferase protein family is having 419 sequences. 386 sequences are identified for hydrophobic group and 338 sequences are identified for lysine is greater than isoluecine amino acid residue content. 338 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. Heme-dependent peroxidases protein family is having 297 sequences. 189 sequences are identified for hydrophobic group and 154 sequences are identified for lysine is greater than isoluecine amino acid residue content. 54 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is not maintaining their class. Lambda repressor-like DNA-binding domains protein family is having 235 sequences. 128 sequences are identified for hydrophobic group and 82 sequences are identified for lysine is greater than isoluecine amino acid residue content. 62 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is not maintaining their class. Long alpha-hairpin protein family is having 857 sequences. 474 sequences are identified for hydrophobic group and 590 sequences are identified for lysine is greater than isoluecine amino acid residue content. 309 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. Multiheme cytochromes protein family is having 159 sequences. 89 sequences are identified for hydrophobic group and 133 sequences are identified for lysine is greater than isoluecine amino acid residue content. 63 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. Nuclear receptor ligand-binding domain protein family is having 360 sequences. 295 sequences are identified for hydrophobic group and 182 sequences are identified for lysine is greater than isoluecine amino acid residue content. 134 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. Phospholipase A2, PLA2 protein family is having 256 sequences. 9 sequences are identified for hydrophobic group and 211 sequences are identified for lysine is greater than isoluecine amino acid residue content. There is non sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is not maintaining their class. SAM domain-like protein family is having 647 sequences. 303 sequences are identified for hydrophobic group and 182 sequences are identified for lysine is greater than isoluecine amino acid residue content. 172 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. Spectrin repeat-like protein family is having 535 sequences. 455 sequences are identified for hydrophobic group and 345 sequences are identified for lysine is greater than isoluecine amino acid residue content. 305 sequences are identified for hydrophobic with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class.

All beta protein families: The 6-bladed beta-propeller protein family is having 244 sequences. In 199 sequences are identified for polar group and 168 sequences are identified for lysine is less than isoluecine amino acid residue content. In 139 sequences are identified

Review Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000143
JPB/Vol.3 Issue 7

for polar group with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The Acid proteases protein family is having 730 sequences. In 17 sequences are identified for polar group and 456 sequences are identified for lysine is less than isoluecine amino acid residue content. In 12 sequences are identified for polar group with the lysine content is less than isoluecine content. So, this family of protein sequences is not maintaining their class. The Beta-clip protein family is having 283 sequences. 19 sequences are identified for polar group and 229 sequences are identified for lysine is less than isoluecine amino acid residue content. In 7 sequences are identified for polar group with the lysine content is less than isoluecine content. So, this family of protein sequences is not maintaining their class. The Beta-Trefoil protein family is having 310 sequences. 236 sequences are identified for polar group and 102 sequences are identified for lysine is less than isoluecine amino acid residue content. In 73 sequences are identified for polar group with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The Carbonic anhydrase protein family is having 198 sequences. 197 sequences are identified for polar group and 4 sequences are identified for lysine is less than isoluecine amino acid residue content. In 4 sequences are identified for polar with the lysine content is less than isoluecine content. So, this family of protein sequences is not maintaining their class. The Concanavalin A-like lectins/glucanases protein family is having 855 sequences. In 698 sequences are identified for polar group and 562 sequences are identified for lysine is less than isoluecine amino acid residue content. In 472 sequences are identified for polar group with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The Cupredoxin-like protein family is having 710 sequences. In 271 sequences are identified for polar group and 133 sequences are identified for lysine is less than isoluecine amino acid residue content. 21 sequences are identified for polar with the lysine content is less than isoluecine content. So, this family of protein sequences is not maintaining their class. The Double-stranded beta-helix protein family is having 570 sequences. 262 sequences are identified for polar group and 195 sequences are identified for lysine is less than isoluecine amino acid residue content. In 50 sequences are identified for polar with the lysine content is less than isoluecine content. So, this family of protein sequences is not maintaining their class. The Galactose-binding domain-like protein family is having 298 sequences. 254 sequences are identified for polar group and 242 sequences are identified for lysine is less than isoluecine amino acid residue content. 225 sequences are identified for polar with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The Glycosyl hydrolase domain protein family is having 297 sequences. 217 sequences are identified for polar group and 205 sequences are identified for lysine is less than isoluecine amino acid residue content. 168 sequences are identified for polar with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The Immunoglobulin-like beta-sandwich protein family is having 2224 sequences. 1755 sequences are identified for polar group and 497 sequences are identified for lysine is less than isoluecine amino acid residue content. 351 sequences are identified for polar with the lysine content is less than isoluecine content. So, this family of protein sequences is not maintaining their class. The Lipocalins protein family is having 283 sequences. 191 sequences are identified for polar group and 31 sequences are identified for lysine is less than isoluecine amino acid residue content. 27 sequences are identified for polar with the lysine content is less than isoluecine content. So, this family of protein sequences is not maintaining their class. The Nucleoplasmin-like/VP protein family is having 592 sequences. 401 sequences are identified for polar group and 315 sequences are identified for lysine is less than isoluecine amino acid residue content. 222 sequences are identified for polar with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The OB-fold protein family is having 2405 sequences. 762 sequences are identified for polar group and 703 sequences are identified for lysine is less than isoluecine amino acid residue content. 189 sequences are identified for polar with the lysine content is less than isoluecine content. So, this family of protein sequences is not maintaining their class. The Prealbumin-like protein family is having 554 sequences. 217 sequences are identified for polar group and 280 sequences are identified for lysine is less than isoluecine amino acid residue content. 71 sequences are identified for polar with the lysine content is less than isoluecine content. So, this family of protein sequences is not maintaining their class. The SH3-like barrel protein family is having 1197 sequences. 308 sequences are identified for polar group and 252 sequences are identified for lysine is less than isoluecine amino acid residue content. 22 sequences are identified for polar with the lysine content is less than isoluecine content. So, this family of protein sequences is not maintaining their class. The Streptavidin-like protein family is having 374 sequences. 355 sequences are identified for polar group and 20 sequences are identified for lysine is less than isoluecine amino acid residue content. 15 sequences are identified for polar with the lysine content is less than isoluecine content. So, this family of protein sequences is not maintaining their class. The Supersandwich protein family is having 329 sequences. 242 sequences are identified for polar group and 180 sequences are identified for lysine is less than isoluecine amino acid residue content. 140 sequences are identified for polar with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The Trypsin-like serine proteases protein family is having 1970 sequences. 1269 sequences are identified for polar group and 811 sequences are identified for lysine is less than isoluecine amino acid residue content. 602 sequences are identified for polar with the lysine content is less than isoluecine content. Therefore, this family of protein sequences is maintaining their class.

Analysis of alpha plus beta families: The Beta-Grasp protein family is having 937 sequences. 328 sequences are identified for polar group and 471 sequences are identified for lysine is greater than isoluecine amino acid residue content. 270 sequences are identified for polar with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. The Cysteine proteinases protein family is having 314 sequences. 193 sequences are identified for polar group and 128 sequences are identified for lysine is greater than isoluecine amino acid residue content. 83 sequences are identified for polar with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. The Ferredoxin-like protein family is having 2132 sequences. 367 sequences are identified for polar group and 1036 sequences are identified for lysine is greater than isoluecine amino acid residue content. 200 sequences are identified for polar with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. The Glyceraldehyde-3-phosphate dehydrogenase-like, C-terminal domain protein family is having 361 sequences. There is non of sequences are identified for polar group and 231 sequences are identified for lysine is greater than isoluecine amino acid residue content. There is non of sequences are identified for polar with the lysine

content is greater than isoluecine content. So, this family of protein sequences is not maintaining their class. The Lysozyme-like protein family is having 1109 sequences. 648 sequences are identified for polar group and 623 sequences are identified for lysine is greater than isoluecine amino acid residue content. 180 sequences are identified for polar with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. The MHC antigen-recognition domain protein family is having 1045 sequences. 731 sequences are identified for polar group and 726 sequences are identified for lysine is greater than isoluecine amino acid residue content. 613 sequences are identified for polar with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. The Microbial ribonucleases protein family is having 295 sequences. 260 sequences are identified for polar group and 42 sequences are identified for lysine is greater than isoluecine amino acid residue content. 39 sequences are identified for polar with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. The Nucleotidyl transferase protein family is having 405 sequences. 169 sequences are identified for polar group and 154 sequences are identified for lysine is greater than isoluecine amino acid residue content. 6 sequences are identified for polar with the lysine content is greater than isoluecine content. So, this family of protein sequences is not maintaining their class. The Protein kinase-like protein family is having 626 sequences. 58 sequences are identified for polar group and 478 sequences are identified for lysine is greater than isoluecine amino acid residue content. 16 sequences are identified for polar with the lysine content is greater than isoluecine content. So, this family of protein sequences is not maintaining their class. The RNase A-like protein family is having 273 sequences. 237 sequences are identified for polar group and 217 sequences are identified for lysine is greater than isoluecine amino acid residue content. 199 sequences are identified for polar with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. The SH2-like protein family is having 314 sequences. 169 sequences are identified for polar group and 155 sequences are identified for lysine is greater than isoluecine amino acid residue content. 105 sequences are identified for polar with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. The TBP-like protein family is having 325 sequences. 166 sequences are identified for polar group and 112 sequences are identified for lysine is greater than isoluecine amino acid residue content. 39 sequences are identified for polar with the lysine content is greater than isoluecine content. So, this family of protein sequences is not maintaining their class. The Thymidylate synthase/dCMP hydroxymethylase protein family is having 188 sequences. 29 sequences are identified for polar group and 63 sequences are identified for lysine is greater than isoluecine amino acid residue content. 23 sequences are identified for polar with the lysine content is greater than isoluecine content. So, this family of protein sequences is maintaining their class. The Zincin-like protein family is having 431 sequences. 201 sequences are identified for polar group and 119 sequences are identified for lysine is greater than isoluecine amino acid residue content. 39 sequences are identified for polar with the lysine content is greater than isoluecine content. Therefore, this family of protein sequences is maintaining their class.

Analysis of alpha by beta protein families: The Dihydrofolate reductases protein family is having 160 sequences. 149 sequences are identified for hydrophobic group and 100 sequences are identified for lysine is less than isoluecine amino acid residue content. 100 sequences are identified for hydrophobic with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The FAD/NAD(P)-binding domain protein family is having 444 sequences. 406 sequences are identified for hydrophobic group and 190 sequences are identified for lysine is less than isoluecine amino acid residue content. 182 sequences are identified for hydrophobic with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The Flavodoxin-like protein family is having 1368 sequences. 1220 sequences are identified for hydrophobic group and 676 sequences are identified for lysine is less than isoluecine amino acid residue content. 646 sequences are identified for hydrophobic with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The UDP-Glycosyltransferase/glycogen phosphorylase protein family is having 181 sequences. 167 sequences are identified for hydrophobic group and 104 sequences are identified for lysine is less than isoluecine amino acid residue content. 104 sequences are identified for hydrophobic with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The Alpha/beta-Hydrolases protein family is having 610 sequences. 387 sequences are identified for hydrophobic group and 360 sequences are identified for lysine is less than isoluecine amino acid residue content. 217 sequences are identified for hydrophobic with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The NAD(P)-binding Rossmann-fold domains protein family is having 1753 sequences. 1689 sequences are identified for hydrophobic group and 856 sequences are identified for lysine is less than isoluecine amino acid residue content. 830 sequences are identified for hydrophobic with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The Periplasmic binding protein-like II protein family is having 468 sequences. 343 sequences are identified for hydrophobic group and 51 sequences are identified for lysine is less than isoluecine amino acid residue content. 41 sequences are identified for hydrophobic with the lysine content is less than isoluecine content. So, this family of protein sequences is not maintaining their class. The Phosphorylase/hydrolase-like protein family is having 473 sequences. 390 sequences are identified for hydrophobic group and 354 sequences are identified for lysine is less than isoluecine amino acid residue content. 290 sequences are identified for hydrophobic with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The PLP-dependent transferases protein family is having 543 sequences. 539 sequences are identified for hydrophobic group and 272 sequences are identified for lysine is less than isoluecine amino acid residue content. 272 sequences are identified for hydrophobic with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The Ribonuclease H-like motif protein family is having 448 sequences. 208 sequences are identified for hydrophobic group and 157 sequences are identified for lysine is less than isoluecine amino acid residue content. 73 sequences are identified for hydrophobic with the lysine content is less than isoluecine content. So, this family of protein sequences is not maintaining their class. The S-adenosyl-L-methionine-dependent methyltransferases protein family is having 352 sequences. 286 sequences are identified for hydrophobic group and 97 sequences are identified for lysine is less than isoluecine amino acid residue content. In 97 sequences are identified for hydrophobic with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The Subtilisin-like protein family is having 209 sequences. 49 sequences

Review Article
OPEN ACCESS Freely available online
doi:10.4172/jpb.1000143
JPB/Vol.3 Issue 7

are identified for hydrophobic group and 156 sequences are identified for lysine is less than isoluecine amino acid residue content. 22 sequences are identified for hydrophobic with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The Nucleotide-diphospho-sugar transferases protein family is having 313 sequences. 264 sequences are identified for hydrophobic group and 196 sequences are identified for lysine is less than isoluecine amino acid residue content. 83 sequences are identified for hydrophobic with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The Thioredoxin fold protein family is having 838 sequences. In 729 sequences are identified for hydrophobic group and 98 sequences are identified for lysine is less than isoluecine amino acid residue content. 86 sequences are identified for hydrophobic with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class. The P-loop containing nucleoside triphosphate hydrolases protein family is having 2152 sequences. 1804 sequences are identified for hydrophobic group and 1035 sequences are identified for lysine is less than isoluecine amino acid residue content. 909 sequences are identified for hydrophobic with the lysine content is less than isoluecine content. So, this family of protein sequences is maintaining their class.

Conclusions

We know the method of the protein sequence is must for analysis for the function of protein sequences. Ideally, the content for grouping of amino acids with the individual amino acid contents based structural classification of proteins to different levels in SCOP hierarchy should be distinctive. In this result we will try proved to, a human expert bases the most straightforward approach on visual identification of structural classes for protein sequences based on the organization of amino acid residues. Our results show some bias towards this. Efforts to improve this are ongoing. Individual amino acid residues carry this research work of classification of protein structural classes using isoluecine and lysine amino acid residues and grouping of amino acids are used for classifying the homologous protein structural classes. This research work has paved way for the trainers to work with the protein sequence analysis for structural classification and structure determination in easy method for the existing methods. The problems of identification of unknown protein sequences may be overcome this research study. The distinction between the alpha plus beta and alpha by beta proteins is not only possible but also necessary. As shown in this study, the difference of the secondary structure content between the two classes is of statistical significance. This means that the distinction between the two classes is objective, rather than subjective. Furthermore, it will be worthwhile for the structure and function prediction of proteins based on the differentiation of alpha plus beta and alpha by beta classes. This study may be considered a first step towards distinguishing between the alpha plus beta and alpha by beta proteins with a reliable quantitative criterion. It is hoped that our work will be useful for the development of protein classification database. It is hoped that the new quantitative criterion will be useful for the development of protein classification databases. Theoretical method of classification of protein structural classes that identify the four structural classes defined in SCOP provide up to 52% accuracy for the datasets in which sequence identity of any pair of sequences. The main contribution of this method is shown to uncover several relations between the predicted secondary structural classes. We show that the main source of the information that allows for successful predictions of structural classes is the secondary structure predicted methods. Therefore, investigations into improving predictions for the mixed classes would constitute an interesting subject for future work. This is for alternative method for experimental determination of structural classification for X-ray crystallography or NMR spectroscopy etc.

References

1. Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, et al. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32: D226-D229.
2. Brown DP, Krishnamurthy N, Sjolander K (2007) Automated Protein Subfamily Identification and Classification. PLoS Comput Biol 3: e160.
3. Chandonia JM, Brenner SE (2006) The Impact of Structural Genomics: Expectations and Outcomes. Science 311: 347-351.
4. Choi IG, Kim SH (2006) Evolution of protein structural classes and protein sequence families. Proc Natl Acad Sci USA 103: 14056-14061.
5. Chou KC (1995) A novel Approach to predicting protein structural classes in a (20–1)-D amino acid composition space. Proteins 21: 319-344.
6. Dietmann S, Holm L (2001) Identification of homology in protein structure Classification. Nat Struct Biol 8: 953-957.
7. Doolittle RF (1995) The multiplicity of domains in proteins. Annu Rev Biochem 64: 287-314.
8. Emmert-Streib F, Mushegian A (2007) A topological algorithm for identification of structural domains of Proteins. BMC Bioinformatics 8: 237.
9. Holm L, Sander C (1993) Protein structure comparison by alignment of distance matrices. J Mol Biol 233: 123-138.
10. Jain P, Hirst JD (2007) Study of Protein Structural Descriptors: Towards Similarity and Classification. NIC Series 36: 165-167.
11. Levitt M, Chothia C (1976) Structural patterns in globular proteins. Nature 261: 552-558.
12. Manikandakumar K, Muthu Kumaran S, Srikumar R (2009) Matrix Frequency Analysis of Oryza Sativa (japonica cultivar-group) Complete Genomes. J Comput Sci Syst Biol 2: 159-166.
13. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536-540.
14. Phillips DC (1966) The three-dimensional structure of an enzyme molecule. Sci Am 215: 78-90.
15. Protein Data Bank (PDB) web site : http://www.rcsb.org/pdb.
16. Rossmann MG, Liljas A (1974) Letter: Recognition of structural domains in globular proteins. J Mol Biol 85: 177-181.
17. Veretnik S, Bourne PE, Alexandrov NN, Shindyalov IN (2004) Toward consistent assignment of structural domains in proteins. J Mol Biol 339: 647-678.
18. Wetlaufer DB (1973) Nucleation, rapid folding, and globular intrachain regions in proteins. Proc Natl Acad Sci USA 70: 697-701.