ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

COMPUTATIONAL ANALYSIS OF SIMILAR REPEAT PATTERNS AND THEIR SECONDARY STRUCTURE VARIATION

Shikha Suman[1] ,Anurag Kulshrestha[1], Ravinder Singh[2] and Dr. Ashutosh Mishra[1]

1.Division of Applied Sciences, Indian Institute of Information Technology, Allahabad-211012, India

2.Department of Biotechnology, Sri Guru Granth Sahib World University, Fatehgarh Sahib-140406, Punjab, India.

Manuscript Info

Manuscript History:

Received: 15 June 2015
Final Accepted: 22 August 2015
Published Online: September 2015

Key words:

Repeats, Similar repeats, PDB, DSSP

Corresponding Author

Shikha Suman

Abstract

Repeats in proteins are represented by repetitive occurrence of amino acids as short sub-sequences. Protein sequences from PROTEIN DATA BANK database were analyzed for the occurrence of similar repeats. Similar repeats of varying length were identified using an algorithm. Analyzing the secondary structure of similar repeats, it appears that similar repeats attain different secondary structures. We also probe the mutation rate of amino acids, to identify those amino acids that have maximum probability of being replaced by another amino acid or amino acids that are most conserved during the course of evolution in similar repeats. Understanding these repeats may shed light on the underlying structural dynamics of repeats.

INTRODUCTION

Repeats are common iterations in the amino acid sequences. Mutation and recombination within a single gene gives rise to regions that share notable sequence similarity. In seminal paper of 2001 M.A. Andrade et al observed that these repeats occur in tandem and form aggregated assemblies when viewed in three dimensional conformations. As they have multiple binding abilities, they play important role in determining the structure of proteins (Andrade et al., 2001[1]). A study indicates that internal repetitions have been found in about 14% of all the known proteins, with higher probability of occurrence in eukaryotic proteins than prokaryotic ones (Pellegrini et al., 2011[2]). The increasing cellular functions complexity in eukaryotic organisms can be attributed from the assembly of repeats (Marcotte et al., 1999[3]). Repeats are important in understanding biological function as they occur in high frequency among sequences, with abilities to grant multiple binding and structural roles on proteins. Internal repeats sometimes cause malfunction and hence are important in the study of diseases (Heringa, 1998[4]; Dijian, 1998[5]) such as Alzheimer's disease, Creutzfeldt–Jacob disease and Gerstmann Straussler–Scheinker disease (Benvenga et al., 1999[6]; Perutz, 1999[7]). Discovering these sequence patterns aids in better understanding of evolutionary processes.

In our study we focus on a unique class of repeats that occur in a sequence called similar repeats. Similar repeats represent a considerable chunk of repetitions among amino acid sequences. Similar repeats are those repeats in which amino acids in one string differ from corresponding amino acids in the second string due to replacement with their structurally similar amino acids. The accepted replacements are

F ↔ Y, Q ↔ E, N ↔ D, K ↔ R, L ↔ I, V ↔ T, S ↔ T (Banerjee et al., 2000[8])

Like F denotes phenylalanine and can be replaced by Y, which denotes tyrosine and vice-versa and so on. Replacement of amino acids may confer new function and secondary structure to protein. The basis of replacement of amino acid is similarity in structure of each amino acid. Like phenylalanine can be replaced with tyrosine or vice versa. To summarize

Similar repeats = Identical amino acids + Structurally similar amino acids
These substitutions tend to change the secondary structure of repeats on replacement. Common regular secondary structures are the helix, β sheet conformation and coil. As an example in crystal structures of myoglobin-ligand complexes at near-atomic resolution represented by PDB Id: 1A6G Chain A, two similar repeats KVEA and KTEA of length four are present at 16 to 19 and 50 to 53 positions respectively. Here valine represented by V at 17th position in KVEA is replaced by threonine represented by T at 51th position in KTEA in the same sequence. The basis of replacement of amino acids are features like PAM 250 substitution matrix score (Dayhoff et al., 1978[9]), BLOSUM 62 substitution matrix score, molecular weight of replaceable amino acids, isoelectric point, hydrophobicity and chemistry as seen in Table 1.

Amino acid	Molecular weight	Isoelectric point	Hydrophobicity	Chemistry	PAM 250 score	BLOSUM 62 score
Phenylalanine(F)	147	5.5	0.951	Aromatic	7	3
Tyrosine (Y)	163	5.7	0.714	Aromatic		
Glutamine (Q)	128	5.7	0.430	Aliphatic	2	2
Glutamic acid(E)	129	3.2	0.458	Aliphatic		
Asparagine (N)	114	5.4	0.448	Aliphatic	2	1
Aspartic acid (D)	115	3.0	0.417	Aliphatic		
Lysine (K)	128	9.7	0.263	Aliphatic	3	2
Arginine (R)	156	10.8	0.000	Aliphatic		
Leucine (L)	113	6.0	0.918	Aliphatic	2	2
Isoleucine (I)	113	6.0	1.000	Aliphatic		
Valine (V)	99	6.0	0.923	Aliphatic	0	0
Threonine (T)	101	5.6	0.634	Aliphatic		
Serine (S)	87	5.7	0.601	Aliphatic	1	1
Threonine (T)	101	5.6	0.634	Aliphatic		

Table 1: Comparison of properties of replaceable amino acids

PAM 250 describes the likelihood that two amino acid residues would mutate to each other in evolutionary time. These amino acids replacement are also verified by NCBI Amino Acid Explorer.

Valine and threonine are selected as they share same shape and molecular volume. It is difficult to differentiate valine and threonine even in high resolution protein structure (Klein et al., 1998[10]). Serine and threonine share many similar properties.

METHODOLOGY

Sequence divergence and variable length of repeats presents a complicated computational task. Similar repeats were extracted from amino acid sequences present in PDB http://www.rcsb.org/ using a sliding window based algorithm. Only that PDB Ids were selected that had X-ray resolution between 0 and 2.5 and refinement factor (R) between 0 and 0.25. Around 125711 protein sequences were extracted from PDB. The algorithm for finding similar repeats in a sequence is explained as follows:
1. Read the sequence string.
2. For amino acid (*i*), evaluate against amino acid (*j*=*i*+1): if *(($a[$i]* eq *$a[$j]) || (($a[$i]* eq *'X')* AND *($a[$j]* eq *'Z'))* where $X \epsilon \{F, Q, N, K, L, V, S\}$ and $Z \epsilon (Y, E, D, R, I, T, T)$.
3. If TRUE, push the i^{th} and j^{th} element into two different arrays *a1* and *a2*.
4. $i = i+L$; $j=j+L$ where L =1, 2....
5. Perform step 2 and push in *a1* and *a2* respectively.
6. Repeat step 4 and 5 till condition 2 holds.
7. To remove subsequence of repeats:
    (a) Store start position and end position of *a1* and *a2* in four different arrays namely *b1, b2, b3, b4.*
    (b) Check for the presence of end position of *a2* in array *b4.*
    (c) If TRUE check start position and end position of *a1* in *b1* and *b2* respectively.

(d) If either of the above conditions are TRUE, compare difference of start position from *b3* and *b1* with difference of start positions of *a1* and *a2*.
(e) If condition (d) is FALSE, consider as a repeat.

8. $i = i+1$
9. Return to step 2.

At the onset we applied our similar repeat finding algorithm. This algorithm requires input in FASTA format. 541628 similar repeats pairs were extracted in total. Among these 541628 similar repeats, 83592 unique similar repeats pairs were extracted. Secondary structure was assigned with the help of DSSP (Define Secondary Structure of Proteins). The 8 structure classes were converted to three states in the following way: DSSP "H" and "G" to helix (dubbed α or H), DSSP "E" to strand (β or E), and all others to Loop (L) (Rost et al., 1994[11]). Frequency of all the possible structures attained by a particular unique repeat was computed. Secondary structure was assigned to a repeat if a particular structure occurs more than or equal to 60% for the unique repeat. The mutation rates of different replaceable amino acids among similar repeat patterns were also evaluated. The position of amino acids is found according to PDB. Similar repeats were also screened for positional conservation.

ANALYSIS

As an illustration, we present the following cases:

A. CASE I:

Chain B of Crystal structure of DNA sequence specificity subunit of a type I restriction-modification enzyme of Methanocaldococcus jannaschii DSM 2661 (PDB ID: 1YF2). The sequence contains 424 amino acid residues.

MFYKEENFKKTEIGEIPEDWEIVELKDVCKKIKAGGTPKTSVEEYYKNGTIPFVKIEDITNSNYLTNTKIKITE
EGLNNSNAWIVPKNSVLFAMYGSIGETAINKIEVATNQAILGIIPKDNILESEFLYYILAKNKNYYSKLGMQT
TQKNLNAQIVKSFKIPLPPLEEQKQIAKILTKIDEGIEIIEKSINKLERIKKGLMHKLLTKGIGHSRFKKSEIGEI
PEDWEVFEIKDIFEVKTGTTPSTKKSEYWENGEINWITPLDLSRLNEKIYIGSSERKVTKIALEKCNLNLIKGS
IIISTRAPVGYVAVLTVESTFNQGCKGLFQKNNDSVNTEFYAYYLKFKKNLLENLSGGSTFKELSKSMLENF
KIPLPPLEEQKQIAKILSSVDKSI ELKKQKKEKLQRMKKKIMELLLTGKVRVKT

This sequence has 20 similar repeat pairs. Among these, similar repeat of length 14 is being considered.

FKKTEIGEIPEDWE from 8 to 21
FKKSEIGEIPEDWE from 216 to 229

STAMP Sequence Alignment

■ Identical Residues
■ Similar Residues

FKKTEIGEIPEDWE
FKKSEIGEIPEDWE

The Secondary structures of the above repeats were superimposed and sequence was found to be aligned with 92.857 % sequence identity. A commonly used measure to establish similarity between two oligopeptide fragments (with a given one to one correspondence between their peptide residues) is the root-mean-square deviation (RMSD). RMSD gives the least sum of square distance between corresponding residues after rigid body transformation (i.e. rotation and translation) of one fragment over the other (Collier et al., 2012[12]). RMSD value was found to be 1.628 Å, indicating high-structure similarity. STAMP (Structural alignment of multiple proteins) (Russell and Barton, 1992)[13] score of 8.069 out of 10 as seen in fig. 1. STAMP score of near about 10 also indicate high degree of structural similarity. Interestingly, the three dimensional structures adopted by these two similar repeats are almost identical.

STAMP Results

PDB ID	Chain ID	Superimposes	Sequence identity (%)	Stamp score (Max 10)	RMSD (Å)
3.PDB	B	[Fixed Molecule]	100.00	10.000	-
4.PDB	B	✓	92.86	8.069	1.628

Figure 1: STAMP alignment score of sub-sequences

It was found to be varying at position 11 in first sub-sequence and 219 in the second subsequence respectively, indicating that amino acid replacement can affect secondary structure as seen in fig. 2.

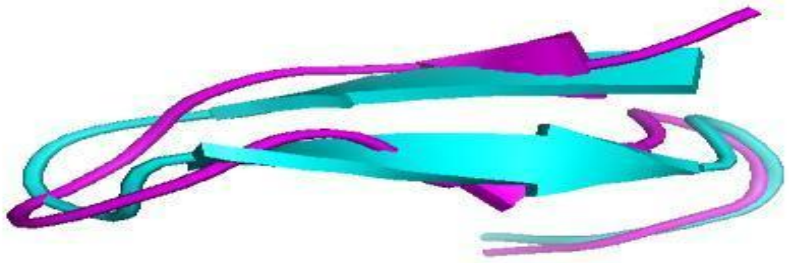

Figure 2: Visualization of the structural model of the example using PYMOL

B. CASE II:

Chain B of Crystal structure of a consensus- designed ankyrin repeat protein of Escherichia coli (PDB ID: 2BKG). Sequence has 166 amino acid residues.

MRGSHHHHHHGSDLGKKLLEAARAGQDDEVRILMANGADVNAEDTYGDTPLHLAARVGHLEIVEVLLK
NGADVNALDFSGSTPLHLAAKRGHLEIVEVLLKYGADVNADDTIGSTPLHLAADTGHLEIVEVLLKYGAD
VNAQDKFGKTAFDISIDNGNEDLAEILQ

This sequence has 3 similar repeat pairs. Among these, similar repeat of length 12 is being considered.

VGHLEIVEVLLK from 57 to 68

TGHLEIVEVLLK from 123 to 134

STAMP Sequence Alignment

■ Identical Residues
■ Similar Residues

VGHLEIVEVLLK
TGHLEIVEVLLK

The secondary structures of the above repeats were also superimposed and sequence was found to be aligned with 91.666 % sequence identity. A high degree of structure similarity was observed with RMSD value of 0.193 Å and STAMP score of 9.766 out of 10 as seen in fig. 3.

STAMP Results

PDB ID	Chain ID	Superimposes	Sequence identity (%)	Stamp score (Max 10)	RMSD (Å)
4.PDB	B	[Fixed Molecule]	100.00	10.000	-
5.PDB	B	✓	91.67	9.766	0.193

Figure 3: STAMP alignment score of sub-sequences

Here also, the three dimensional structures of the repeat patterns were almost identical.

It was found to be varying at position 57 in first sub-sequence and 123 in the second subsequence respectively as seen in fig.4.

Figure 4: Visualization of the structural model of the example using PYMOL.

Secondary structure was assigned to a repeat if the relative occurrence of that particular structure for a unique repeat is more than 60% as evident from Table 2. This approach may also help in assigning structure to new protein sequences.

Repeat	Structure	Occurrence (%)	Repeat	Structure	Occurrence (%)
DLLK	HHHH	81.250	DILK	HHHH	78.125
LTGD	LLLL	93.103	LTGN	SSSL	89.655
DSIQ	LLHH	66.667	NSLQ	LLHH	100.00
SIAY	LHHH	66.667	TLAF	SSSS	66.667
KELG	HHLL	62.500	KEIG	HHLL	62.500
ETLLN	LLHHH	100.00	ESILN	HHHHH	100.00
KRLE	LSSL	71.428	KKLQ	HHHH	100.00
LSSS	LLLL	63.636	ISST	SSLL	63.636
SLSS	LSLL	67.384	TLSS	SSSS	99.692
DALT	HHHH	100.00	NALS	HHLH	89.878
SGSGS	SSSSS	90.459	SGSGT	SSSLL	87.563
GTRTI	LLLLS	95.652	GSRSL	LSLLS	69.565
LNPS	SSLL	100.00	LDPT	LLLL	100.00
TVLD	HHHH	100.00	SVID	SSSS	100.00
NTSG	SSLL	100.00	DVTG	SSSL	100.00
TDTL	HHHH	100.00	VDSL	HHHH	100.00
DLQV	HHHH	100.00	NIEV	LSSS	100.00
SLYG	SSSS	100.00	SIFG	SLLL	100.00
DSTL	LLLL	83.333	DSSL	SHHH	83.333
SVGK	LLLH	66.667	TVGK	LLSS	66.667
KLDD	LLSL	83.333	KLNN	HHHH	100.00
KQGVE	HLLLL	100.00	KEGVE	HHHHH	100.00
TGGG	LLLH	66.667	VGGG	SSSS	66.667
RVAV	LSSS	100.00	RTAV	LLLL	90.909

Table 2: Secondary structure assignment with occurrence (%)

RELATIVE OCCURRENCE OF SIMILAR REPEATS

Similar repeats of variable length were obtained. On exploring, relative occurrence of similar repeats of length four was found to be more as compared to others as seen in fig. 5. This indicates that repeats of smaller length are more conserved during the evolutionary process.

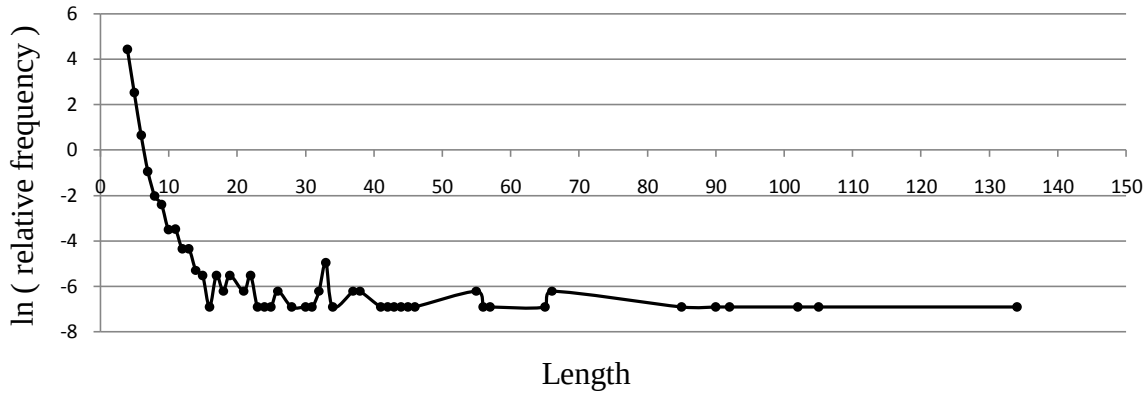

Figure 5: Relative frequency of similar repeats of different length

MUTATION RATE OF CONSTITUTING AMINO ACIDS

For predicting mutation rate, mutation rate of each replaceable amino acid of first string with respect to corresponding amino acid of second string is represented by

$$M_{ij} = \frac{count_i(j)}{\sum_{j=1}^{14} count_i(j)}$$

for $i = F, Y, Q, E, N, D, K, R, L, I, V, T, S, T$ and $j = Y, F, E, Q, D, N, R, K, I, L, T, V, T, S$
$count_i(j)$ denotes the times that amino acid i is replaced by j.
The highest relative mutation rate was elucidated of isoleucine to leucine that was found to be 11.948% as evident from fig. 6. This means there is maximum probability of conversion of isoleucine to leucine or least conserved during the evolutionary process. Conversely, there is least probability of conversion of Phenylalanine to Tyrosine or most conserved during the evolutionary process. As it has least relative mutation rate of 3.215% among all amino acids.

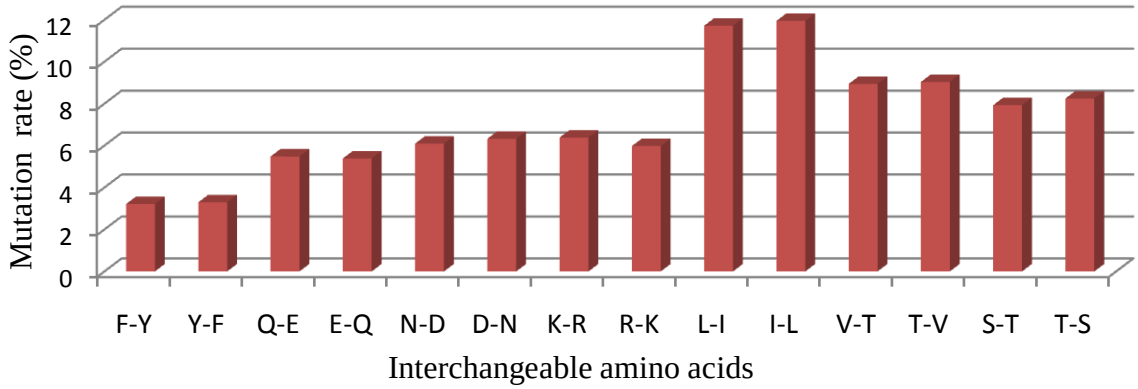

Figure 6: Mutation rate of different interchangeable amino acids

It can also be concluded that conversion of isoleucine to leucine plays a significant role in decreasing structural conservation of protein due to its maximum probability for conversion.

POSITIONAL CONSERVATION OF REPEATS

Similar repeats were explored for positional conservation. It can also be emphasized with high confidence that repeats have positional conservation in majority of cases. This means that repeats are found at particular positions only, among different proteins as evident from Table 3.

PDB Id	Repeat	Position	Repeat	Position
1A06_A	VALD	79 to 82	TALD	280 to 283
4FG8_B	VALD	79 to 82	TALD	280 to 283
4FG8_A	VALD	79 to 82	TALD	280 to 283
4FG9_B	VALD	79 to 82	TALD	280 to 283
4FG9_A	VALD	79 to 82	TALD	280 to 283
1A09_A	DSIQ	144 to 147	NSLQ	224 to 227
1A1A_A	DSIQ	144 to 147	NSLQ	224 to 227
4F59_A	DSIQ	144 to 147	NSLQ	224 to 227
4F5A_A	DSIQ	144 to 147	NSLQ	224 to 227
4F5B_A	DSIQ	144 to 147	NSLQ	224 to 227
1Y57_A	DSIQ	144 to 147	NSLQ	224 to 227
1CGT_A	LTGD	580 to 583	LTGN	604 TO 607
1CGU_A	LTGD	580 to 583	LTGN	604 TO 607
3CGT_A	LTGD	580 to 583	LTGN	604 TO 607
5CGT_A	LTGD	580 to 583	LTGN	604 TO 607
8CGT_A	LTGD	580 to 583	LTGN	604 TO 607
9CGT_A	LTGD	580 to 583	LTGN	604 TO 607
1AIS_A	LLIFSSGKLV	57 to 66	ILLFSSGKIV	148 to 157
1D3U_A	LLIFSSGKLV	57 to 66	ILLFSSGKIV	148 to 157
1PCZ_B	LLIFSSGKLV	57 to 66	ILLFSSGKIV	148 to 157

Table 3: Position conservation of similar repeats

DISTRIBUTION IN VARIOUS ORGANISMS

Distribution of similar repeats in various organisms was analyzed. Homo sapiens constitute maximum percentage of about 17.646 % of similar repeats found. The proportion of similar repeats in various organisms is depicted in fig. 7.

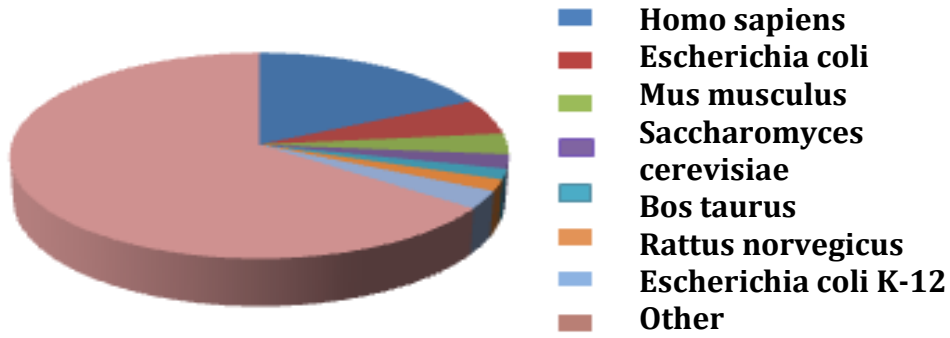

Figure 7: Distribution of similar repeats in various organisms

CONCLUSION

Our study presents a comprehensive picture of new pattern among repeats. Although internal repeats have been identified in various proteins, not much is known about their role in protein structure and function. Similar repeats form a considerable proportion of repeats among amino acid sequences. 60.233% of similar repeats depicted variation in secondary structure. Thereby inferring that even a change in single amino acid confers changes in secondary structure of proteins without altering the physiochemical properties. Furthermore, similar repeats are conserved at their position in different proteins. The high level of conservation of similar repeats at their positions indicate possible role in structural and functional implications. We hope that the range of similar repeat patterns as identified by our studies will be helpful for further analysis of internal similar repeats with respect to their evolution and their implications on protein function.

ACKNOWLEDGEMENT

Computing facilities at Indian Institute of Information Technology, Allahabad, India are gratefully acknowledged. Special thanks to Prof. K. Sekar for his encouragement.

REFERENCES

[1] Andrade, M. A., Iratxeta, C. P. and Ponting, C. (2001): Protein repeats: structures, functions, and evolution. J. Struct. Biol., **134(2-3)**:117-131.

[2] Pellegrini, M., Renda, M. E. and Vecchio, A. (2011): Ab initio detection of fuzzy amino acid tandem repeats in protein sequences. BMC Bioinform., **13**: S3.

[3] Marcotte, E. M., Pellegrini, M., Yeates T. O. and Eisenberg, D. (1999): A census of protein repeats. J. Mol. Biol., **293**:151-160.

[4] Heringa, J. (1998): Detection of internal repeats: How common are they? Curr. Opin. Struct. Biol., **8**:338–345.

[5] Djian, P. (1998): Evolution of simple repeats in DNA and their relation to human diseases. Cell, **94**:155–160.

[6] Benvenga, S., Campenni, A. and Facchiano, A. (1999): Internal repeats of prion protein and A beta PP, and reciprocal similarity with the amyloid-related proteins. Amyloid, **6**:250–255.

[7] Perutz, M. F. (1999): Glutamine repeats and neurodegenerative diseases: Molecular aspects. Trends Biochem. Sci., **24**:58–63.

[8] Banerjee, N., Chidambarathanu, N., Sabarin, R., Daliah, M., Vasuki, C., Rajani, B. N. and Sekar, K. (2000): An algorithm to find similar internal sequence repeats. Curr. Sci., **97**:1345– 1349.

[9] Dayhoff, M. O., Schwartz, R. M. and Orcutt, B. C. (1978): A model of evolutionary change in proteins. Atlas of protein sequence and structure, **5**:345-352.

[10] Klein O., Polack, G .W., Surti, T., Kegler-Ebo, D., Smith, S. O. and Dimaio, D. (1998): Role of Glutamine 17 of the bovine Papilloma virus E5 Protein in Platelet derived growth factor β receptor activation and cell transformation. Jour. Virology, 1**72**:8921-8932.

[11] Rost, B. and Sander, C. (1994): Combining evolutionary information and neural networks to predict protein secondary structure. PROTEINS: Struct. Funct. Genet., **19**:55-72.

[12] Collier, J. H., Lesk, A. M., Banda, M. G. and Konagurthu, A. S. (2012): Super: a web server to rapidly screen superposable oligopeptide fragments from the protein data bank. Nucleic Acids Res., **40**:334-339.

[13] Russell, R .B. and Barton, G .J. (1992): Multiple protein sequence alignment from tertiary structure comparison. PROTEINS: Struct. Funct. Genet., **14**:309-323.