Journal of Academia and Industrial Research (JAIR)
Volume 2, Issue 4 September 2013

ISSN: 2278-5213

REVIEW ARTICLE

Principles of Protein Folding

Ali Bandehagh[1*], Pouya Motie Noparvar[1] and Ebrahim Dorani Uliaie[1]
[1]Dept. of Plant Breeding and Biotechnology, Faculty of Agriculture, University of Tabriz, Tabriz, Iran
bnandehhagh@tabrizu.ac.ir*; +98 4113392031

Abstract

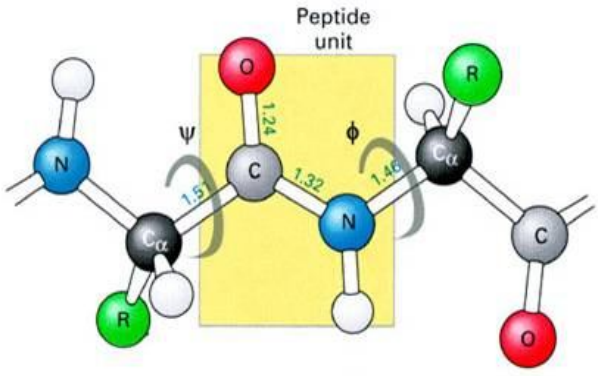
Protein folding is a process in which a polypeptide folds into a specific, stable, functional, three-dimensional structure. It is the process by which a protein structure assumes its functional shape or conformation. Proteins are comprised of amino acids with various types of side chains, which may be hydrophobic, hydrophilic or electrically charged. It is now well known that under physiological conditions, proteins normally spontaneously fold into their native conformations but there are some exterior factors which help polypeptide chain finding its natural shape. Different levels of folding a protein after amino acid sequence or primary structure consist of secondary, tertiary and quaternary structures. Protein folding pathway or mechanism is the typical sequence of structural changes; in which protein find its native structure. 3D structure of proteins is studied by scientists using different methods and there are many types of software to survey this. Many factors control protein folding, interior and exterior factors. Creation of natural folded proteins by these factors and protein translation are simultaneous. The main objective of this review is to unveil the fact; despite there are many factors controlling protein folding, mainspring is amino acids sequence which itself rises from chemo-physical laws.

Keywords: Protein folding, stability, three dimensional structure, folding factors, structure prediction.

Introduction

Amino acids are biologically important organic compounds made from amine (-NH_2) and carboxylic acid (-COOH) functional groups. What distinguishes amino acids from each other is the side chain. There are 20 different specific side chains for 20 amino acids which are encoded by genetic code (Turanov *et al.*, 2009). Although these are not the only amino acids, about 500 amino acids are known (Wagner and Musso, 1983). Amino acids can be classified in many ways. Structurally they can be classified according to the functional groups locations as alpha (α), beta (β), gamma (γ) or delta (δ) amino acids; other categories relate to polarity, pH level, and side chain group type (aliphatic, acyclic, aromatic, containing hydroxyl or sulfur, etc.). Depending on the chemical nature of the side chain, amino acids are usually divided into three different classes (Murray *et al.*, 2009) which are shown in Table 1. For synthesizing a polypeptide chain, amino acids should be coupled end to end by the formation of peptide bonds. Amino acids are linked via amide bonds which are also known as peptide bonds. To yield a peptide bond, a condensation reaction should occur and a H_2O molecule exits. These units are not rigid but have degrees of freedom and each unit could rotate around to such bonds: the C_α-C which is called psi (ψ) and the N-C_α which is called phi (ϕ) (Fig. 1). Psi and phi are the only degrees of freedom and the conformation of the whole chain of polypeptide is completely determined when these are defined for each amino acid (Ramachandran and sassiekharan, 1968).

Fig.1. Psi and Phi angles.



Source: http:// //wiki.cmbi.ru.nl/index.php/Phi-psi_angle.

The proteins we observe in nature have evolved through selection pressure to perform specific functions. The functional properties of proteins depend on their three-dimensional structures. The 3-D structures arises because of particular sequences of amino acids in polypeptide chains fold to generate, from linear chains, compact domains with specific 3-D structures (Brandon and Tooze, 1992). The amino acid sequence of a protein's polypeptide chain is called its primary structure. Interior factors such as sequence of amino acids create special structures which are called secondary structures. Properties of peptide bonds, psi and phi angles, and side groups for each amino acid are such factors.

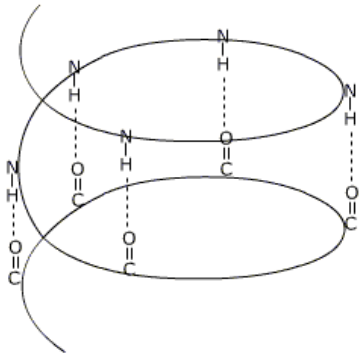
Table 1. Twenty amino acids classified into different categories.

Category	Name	Three-letter code	One-letter code	Molecular formula	Helical propensity
With aliphatic side chains	Glycine	Gly	G	$C_2H_5NO_2$	1
	Alanine	Ala	A	$C_3H_7NO_2$	0.0
	Valine	Val	V	$C_5H_{11}NO_2$	0.61
	Leucine	Leu	L	$C_6H_{13}NO_2$	0.21
	Isoleucine	Ile	I	$C_6H_{13}NO_2$	0.41
With side chains containing hydroxylic groups	Serine	Ser	S	$C_3H_7NO_3$	0.5
	Threonine	Thr	T	$C_4H_9NO_3$	0.66
	Tyrosine	Tyr	Y	$C_9H_{11}NO_3$	0.53
With side chains containing sulfur atoms	Cysteine	Cys	C	$C_3H_7NO_2S$	0.68
	Methionine	Met	M	$C_5H_{11}NO_2S$	0.24
With side chains containing acidic groups of their amides	Aspartic acid	Asp	D	$C_4H_7NO_4$	0.69
	Aspargine	Asn	N	$C_4H_8N_2O_3$	0.65
	Glutamic acid	Glu	E	$C_5H_9NO_4$	0.40
	Glutamine	Gln	Q	$C_5H_{10}N_2O_3$	0.39
With side chains containing basic groups	Arginine	Arg	R	$C_6H_{14}N_4O_2$	0.21
	Lysine	Lys	K	$C_6H_{14}N_2O_2$	0.26
	Histidine	His	H	$C_6H_9N_3O_2$	0.61
Containing aromatic rings	tryptophan	Trp	W	$C_{11}H_{12}N_2O_2$	0.49
	Histidine	His	H	$C_6H_9N_3O_2$	0.61
	Phenylalanine	Phe	F	$C_9H_{11}NO_2$	0.54
	Tyrosine	Tyr	Y	$C_9H_{11}NO_3$	0.53
Imino acid	proline	Pro	P	$C_5H_9NO_2$	3.16

Source: Classification from Murray (2009), molecular formula from "http://www.webqc.org/aminoacids.php" and helical propensity from Pace *et al.* (1998).

For example, interior of most proteins has almost exclusively hydrophilic side chains (e.g. myoglobin) (Kendrew *et al.*, 1958, 1960). Maybe this is not a regular shape, thus there are stable structures like helices and sheets. Alpha helices and beta sheets are most common secondary structures (Cuomo *et al.*, 1984) which are characterized by having the main chain NH and CO groups participating in hydrogen bonds to each other (Fig. 2). These are formed when a number of consecutive residues have same psi and phi angles. α-helix is a right handed coiled or spiral conformation and the helix has 3.6 residues per turn (Neurath, 1940). This repeated i+4→i hydrogen bonding is the most prominent characteristic of α-helix. Some other prevalent helices are π-helix (i+5→i) and 3.0_{10} helix (i+3→i) (Dieter *et al.*, 2005).

Fig. 2. H-Bonds in α-helix.

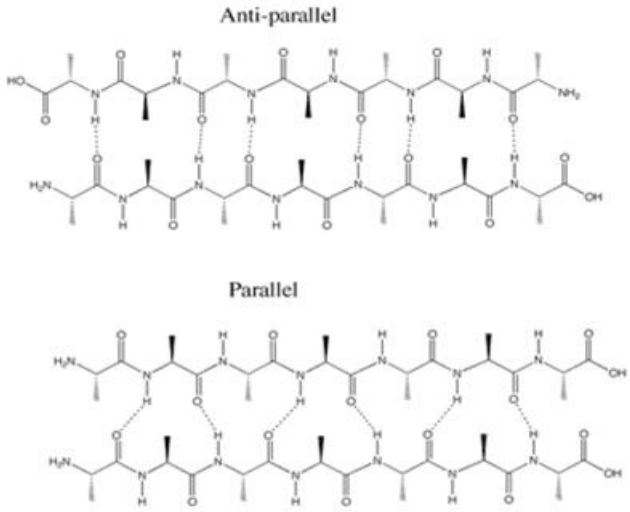


Source: http://www.goiit.com/posts/list/organic-chemistry-what-is-bita-helix-structure-of-protein-molecules-912015.htm.

Helices observed in proteins can range from four to over forty residues long, but a typical helix contains about ten amino acids (about three turns) (Lupas and Gruber, 2005). The More it is long, the more instability there is. In general, the backbone hydrogen bonds of alpha-helices are considered slightly weaker than those found in beta-sheets and are readily attacked by the ambient water molecules (Fain *et al.*, 2008). As Pace (1998) clarifies, there are scales of propensity for each amino acid as mentioned in Table 1.

The second frequent structural element found in proteins is the beta (β) sheet. Beta sheets consist of beta strands connected laterally by at least two or three backbone hydrogen bonds, forming a generally twisted, pleated sheet (Pauling and Corey, 1951). A beta strand is built up from polypeptide chain typically 3 to 10 amino acids long. The majority of β-strands are arranged adjacent to other strands and form an extensive hydrogen bond network with their neighbors in which the N-H groups in the backbone of one strand establish hydrogen bonds with the C=O groups in the backbone of the adjacent strands (Voet and Voet, 1990). The beta sheets that are formed from several such beta strands are "pleated" with C_α atoms successively a little above and below the beta sheet (Brandon and Tooze, 1992). When amino acids in one beta strand run in the same biochemical direction with another strand, N-terminal to C-terminal, this is called parallel. Whereas in antiparallel sheets, amino acids in successive strands can have alternative directions, N-terminal to C-terminal followed by C-terminal to N-terminal followed by N-terminal to C-terminal and so on (Fig. 3).

Fig. 3. Antiparallel and parallel β-sheets.



Source: http://chemed.chem.wisc.edu/chempaths/GenChem-Textbook/Secondary-Protein1027.html.

Other levels of protein folding

Alpha helices and beta sheets are the simplest structures for proteins whereas there are supersecondary structures too. Alpha helices and beta strands are connected by loop regions of various lengths and different shapes. A combination of secondary structure elements forms the stable constructions which are generally called motifs. For example, when two antiparallel beta strands are connected via loop regions, they are in fact hairpin loops. Some motifs have particular functions such as DNA binding, others could be part of a larger assemble and they could associated with specific functional sites of a protein, such as protein-protein interaction sites, small molecule binding sites etc. (Chiang *et al.*, 2007). The most common motifs found in protein are shown in Table 2.

Table 2. Motifs and their structure.

Motif	Structure
Beta hairpin	Two antiparallel beta strands connected by a tight turn of a few amino acids between them.
Greek key	Four beta strands folded over into a sandwich shape.
Omega loop	A loop in which the residues that make up the beginning and end of the loop are very close together.
Helix-loop-helix	Consists of alpha helices bound by a looping stretch of amino acids. This motif is seen in transcription factors.
Zinc finger	Two beta strands with an alpha helix end folded over to bind a zinc ion. Important in DNA binding proteins.
Helix-turn-helix	2 α-helices joined by a short strand of amino acids.

Source: Krebs *et al.* (2008).

Several motifs usually combine to form compact structures, which are called domains. A polypeptide chain may have one or more domains but each domain have an independent function, in which a domain could act independently and its job separate from original peptide. A fundamental unit of tertiary structure is the domain. In fact, domains are biological system's hands; it means that domains do the leading functions in a cell or even out of a cell (i.e. intracellular space). For example, in the lambda repressor protein, there is one domain at the N-terminal of the polypeptide chain that binds DNA and a second C-terminal domain holds tow polypeptide chains together into a dimeric repressor molecule (Anderson *et al.,* 1981). Levitt *et al.* (1977) presented taxonomy of protein structures and they could show that combination of motifs build up the core of most domain structures and also form the basis of a classification into three main groups: alpha domains, beta domains and alpha/beta domain. Today Structural Classification of Proteins (SCOP) database is a large manual classification of protein structural domains based on similarities of their structures and amino acid sequences.

Proteins with the same shapes but having little sequence or functional similarity are placed in different "superfamilies" and are assumed to have only a very distant common ancestor. Proteins having the same shape and some similarity of sequence and/or function are placed in "families" and are assumed to have a closer common ancestor. The shapes of domains are called "folds" in SCOP. In this way levels of SCOP are as follows (Lo Conte *et al.*, 2002).

1. *Class:* Types of folds, e.g., beta sheets.
2. *Fold:* The different shapes of domains within a class.
3. *Superfamily:* The domains in a fold are grouped into superfamilies, which have at least a distant common ancestor.
4. *Family:* The domains in a superfamily are grouped into families, which have a more recent common ancestor.
5. *Protein domain:* The domains in families are grouped into protein domains, which are essentially the same protein.
6. *Species:* The domains in "protein domains" are grouped according to species.
7. *Domain:* part of a protein. For simple proteins, it can be the entire protein.

The folds are grouped into "classes". The classes are the top level, or "root" of the SCOP hierarchical classification. The classes are displayed something like this:

a. *All alpha proteins:* Domains consisting of α-helices
b. *All beta proteins:* Domains consisting of ß-sheets
c. Alpha and beta proteins: Mainly parallel beta sheets (beta-alpha-beta units).

d. *Alpha and beta proteins (a+b):*Mainly antiparallel beta sheets (segregated alpha and beta regions).
e. *Multi-domain proteins (alpha and beta):* Folds consisting of two or more domains belonging to different classes
f. *Membrane and cell surface proteins and peptides:* Does not include proteins in the immune system
g. *Small proteins:* Usually dominated by metal ligand, heme, and/or disulfide bridges.
h. *Coiled-coil proteins:* Not a true class.
i. *Low resolution protein structures:* Peptides and fragments. Not a true class.
j. *Peptides:* Peptides and fragments. Not a true class.
k. *Designed proteins:* Experimental structures of proteins with essentially non-natural sequences. Not a true class.

Proteins are not always monomeric, it means they don't have only one polypeptide chain. Several identified proteins have more than one chain and these kinds of proteins are called quaternary structures. These subunits can function either independently of each other (Berg *et al.,* 2002). Subunits could be identical or different. Common shorthand for describing such proteins is to use Greek letters for each type of subunit and subscribe numeral to specify numbers of subunits (Petsko and Ringe, 2004). There are two major categories of proteins with quaternary structure, fibrous and globular. Silk is an example for fibrous proteins and insulin is globular one. A variety of bonding interactions including hydrogen bonding, salt bridges and disulfide bonds hold the various chains into a particular geometry. Quaternary structures add stability by decreasing the surface/volume ratio of smaller subunit. Surface/volume ratio decreases with size and reduces the tendency to aggregate (Shirota *et al*., 2008). Large structure provides rigidity necessary to orient the substrate and key amino acids to enable catalysis e.g. extremely small proteins require metals or disulfides for stability (Klapper *et al.*, 1986).

One of the most important problems in molecular biology is the protein structure prediction problem. Unfortunately, determining the three-dimensional fold of a protein is very difficult. Moreover, same structures may have different functions (Berger *et al.*, 1995) because the relationship between primary structure and tertiary structure is not straight forward, two biopolymers may share the same motif yet lack appreciable primary structure similarity. Nonetheless, it has been shown that protein structures are three to ten times more conserved than the amino acid sequence (Illergard *et al.,* 2009). Thus, a particular motif, i.e., a zinc-binding domain of very similar or virtually identical structure, can be found in many different proteins, which could also be unrelated to each other when function is concerned (Sousounis *et al.*, 2012). Anyway, scientists have developed different programs using different algorithms and vast amounts of data to predict 3-D structures for proteins.

There are programs for homology modeling (RaptorX, 3D-JIGSAW, Biskit, CABS and so on), fold recognition (3D-PSSM, Bioingbu, HHpred, LOOPP etc.), initio structure prediction (EVfold, QUARK, I-TASSER, ROBETTA, Abalone and PEO-FOLD), secondary structure prediction (NetSurfP, GOR, Jpred, Meta-PP, PREDATOR and PSSpred), transmembrane helix predict (HMMTOP, MEMAST, TMHMM and SVMTop2) and signal peptide prediction (SignalP).

Natural shape formation

Having natural folded proteins require lowest level of energy. During protein synthesis, residues waggle to fine the best conformation with the lowest level of energy (i.e. co-translational folding) (Alexey *et al*., 1997). Protein folding forms energetically favorable structures stabilized by hydrophobic interactions clumping, hydrogen bonding and Van der Waals forces between amino acids. Besides hydrogen bindings, covalent bonding may also occur during the folding to a tertiary structure, through the formation of disulfide bridges or metal clusters. According to Roger Pain's "Mechanisms of Protein Folding" (2000), molecules also often pass through an intermediate "molten globule" state formed from a hydrophobic collapse (in which all hydrophobic side-chains suddenly slide inside the protein or clump together) before reaching their native confirmation. However, this means all the main chain NH and CO groups are buried in a non-polar environment, but they prefer an aqueous one, so secondary structures must fit together very well, so that the stabilization through hydrogen bonding and Van der Waals forces interactions overrides their hydrophilic tendencies. Water soluble proteins fold into compact structures with non-polar, hydrophobic cores. The inside of protein contains non-polar residues in center (i.e. leucine, valine, methionine and phenylalanine), while the outside contains primarily polar, charged residues (i.e. aspatate, glutamate, lysine and arginine). This way the polar, charged molecules can interact with the surrounding water molecules while the hydrophobic molecules are protected from the aqueous surroundings. Many factors affect proteins stability. As discussed before the major factors affecting protein stability are hydrophobic interactions, hydrogen bonds and conformational entropy but there are other intermediates too, which interact with folding factors such as chaperones, cochaperones, peptidyl prolyl cis/trans isomerases (PPIase), oxidoreductases, glycan-binding protein and client- specific folding factors (Booth and Curnow, 2009). Many folding factors are great in that they are multi-functional. One folding factor can take care of different areas of the folding pathway. Unfortunately, this leads to redundancy due to different classes of proteins carrying out overlapping functions. This functional redundancy complicates the understanding of the specific roles of individual folding factors in aiding maturation of client proteins.

Folding factors also prefer to act in concert during the maturation process, which further obscures the individual roles of each factor. Since these roles are not clear, it is difficult to confirm that even if one folding factor deals with a particular reaction in one protein, that same folding factor will carry out the same function in another. Amino acid sequences itself with all these factors contribute each other to form the protein in its natural shape with correct activity.

Conclusion

In biological environment, proteins get their natural folding or their native state. In retrospect this behavior can be seen to emerge as a consequence of three self-obvious principles-the cooperativity of foldon units, their sequential stabilization and the ubiquity and chance nature of folding errors (Englander *et al.*, 2008). The independently cooperative nature of the foldon building blocks that compose native proteins is well known. Much information now demonstrates their continued unfolding refolding behavior in structured proteins. Thermodynamic principle requires that their continued unfolding and refolding under native conditions must recapitulate the protein's natural folding pathway. Pre-existing structure guides and stabilizes the formation of complementary structure. So in total, three main principles are beyond protein folding, including amino acid sequences, favorable thermodynamically tendency and finally intermediates.

References

1. Alexey, N., Baldwin, F.O. and Baldwin, T.O. 1997. Cotranslational protein folding. *J. Biol. Chem.* 272: 32715-32718.
2. Anderson, W.F., Ohlendorf, D.H., Takeda, Y. and Matthews, B.W. 1981. Structure of the cro repressor from bacteriophage lambda and its interaction with DNA. *Nature.* 290: 754-758.
3. Berg, J.M., Tymoczko, J.L. and Stryer, L. 2002. Biochemistry, 5th edition. New York: W H Freeman. pp.280-348.
4. Berger, B., Wilson, D.B., Tonchev, T., Milla, M. and Kim, P.S. 1995. Predicting coiled coils using pair-wise residue correlations. *Proc. Natl. Acad. Sci. USA*. 92: 8259-8263.
5. Booth, P.J. and Curnow, P. 2009. Folding scene investigation: membrane proteins. *Curr. Opin. Struc. Biol.* 19: 8-13.
6. Brandon, C. and Tooze, J. 1992. Introduction to protein structure. Garland publishing company, NewYork. pp.1-76.
7. Chiang, Y.S., Gelfand, T.I., Kister, A.E. and Gelfand, I.M. 2007. New classification of supersecondary structures of sandwich like proteins uncovers strict patterns of strand assemblage. *Proteins*. 68: 915-921.
8. Cuomo, V., Macchiato, M.F. and Tramontano, A. 1984. A statistical method for predicting alpha-helical and beta-sheet regions in proteins from their amino acidic sequences. *Ilnuovo cimento D.* 3: 421-435.
9. Dieter, S., Hook, D.F. and Glattli, A. 2005. Helices and other secondary structures of β- and γ-Peptides. *Wiley Inter Sci.*DOI 10.1002/bip.20391.
10. Englander, S.W., Mayne, L. and Krishna, M.M.G. 2008. Protein folding and misfolding: Mechanism and principles. *Quart. Rev. Biophy.* 40: 287-326.
11. Fain, A.V., Ukrainskii, D.L., Dobkin, S.A., Galkin, A.V. and Esipova, N.G. 2008. Arrangement of single, double and bifurcated hydrogen bonds in protein α-helices: Statistical analysis of PDB-select. *Biophy.* 53: 125-133.
12. Illergard, K., Ardell, D.H. and Elofsson, A. 2009. Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. *Proteins.* 77: 499-508.
13. Kendrew, J.C., Bodo, G., Dintzis, H.M., Parrish, R.G., Wyckoff, H. and Phillips, D.C. 1958. A three-dimensional model of the Myoglobin molecule obtained by X-Ray analysis. *Nature.* 181: 662-666.
14. Kendrew, J.C., Dickerson, R.E., Strandberg, B.E., Hart, R.G., Davies, D.R., Phillips, D.C. and Shore, V.C. 1960. Structure of myoglobin: A three-dimensional fourier synthesis at 2 Å resolution. *Nature.* 185: 422-427.
15. Klapper, I., Hagstrom, R., Fine, R., Sharp, K. and Honig, B. 1986. Focusing of electric fields in the active site of Cu-Zn superoxide dismutase: Effects of ionic strength and amino-acid modification. *Proteins.* 1: 47-59.
16. Krebs, M.R., Domike, K.R., Cannon, D. and Donald, A.M. 2008. Common motifs in protein self-assembly. *Faraday Discussions.* 139: 265-274.
17. Levitt, M., Chothia, C. and Richardson, D. 1977. Structure of proteins: Packing of alpha-helices and pleated sheets. *Proc. Natl. Acad. Sci. USA*. 74: 4130-4134.
18. Lo Conte, L., Brenner, S.E., Hubbard, T.J.P., Chothia, C. and Murzin, A. 2002. SCOP database in 2002: Refinements accommodate structural genomics. *Nuc. Acids Res.* 30: 264-267.
19. Lupas, A.N. and Gruber, M. 2005. The structure of α-helical coiled coils. *Adv. Prot. Chem.* 70: 37-78.
20. Murray, R.K., Rodwell, V.W., Bender, D., Botham, K.M., Weil, P.A. and Kennelly, P.J. 2009. Harper's Illustrated Biochemistry, 28th Edition. McGraw-Hill Companies, Inc. pp.10-60.
21. Neurath, H. 1940. Intramolecular folding of polypeptide chains in relation to protein structure. *J. Phys. Chem*. 44: 296-305.
22. Pace, C., Scholtz, N. and Martin, J. 1998. A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. J.* 75: 422-427.
23. Pain, R.H. 2000. Frontiers in molecular biology, 32: Mechanisms of protein folding. Oxford university press, Oxford, New York. pp.26-55.
24. Pauling, L. and Corey, R.B. 1951. The pleated sheet, a new layer configuration of polypeptide chains. *Proc. Natl. Acad. Sci. USA.*37: 251-256.
25. Petsko, G.A. and Ringe, D. 2004. Protein structure and function. New science press. pp.20-40.
26. Ramachandran, G.N. and Sassiekharan, V. 1968. Conformation of polypeptides and proteins. *Adv. Prot. Chem*. 28: 283-437.
27. Shirota, M., Ishida, T. and Kinishita, K. 2008. Effects of surface to volume ratio of proteins on hydrophilic residues: Decrease in occurrence and increase in buried fraction. *Prot. Sci*. 17: 1596-1602.
28. Sousounis, K., Haney, C.E., Cao, J., Sunchu, B. and Tsonis, P.A. 2012. Conservation of the three-dimensional structure in non-homologous or unrelated proteins. *Human Genomics.* 6: 10-20.
29. Turanov, A.A., Lobanov, A.V., Fomenko, D.E., Morrison, H.G., Sogin, M.L., Klobutcher, L.A., Hatfield, D.L. and Gladyshev, V.N. 2009. Genetic code supports targeted insertion of two amino acids by one codon. *Sci.* 323: 259-261.
30. Voet, D. and Voet, J.G. 1990. Biochemistry. John Wiley and Sons, New York. pp.250-410.
31. Wagner, I. and Musso, H. 1983. New naturally occurring amino acids. *Angewandte Chemie Int. Edn.* 22: 816-828.