IOSR Journal of Pharmacy and Biological Sciences (IOSR-JPBS)
e-ISSN: 2278-3008, p-ISSN:2319-7676. Volume 9, Issue 5 Ver. IV (Sep -Oct. 2014), PP 01-06
www.iosrjournals.org

Comparison of evolutionary characteristics of orthologous sets from halophiles, thermophiles and mesophiles

Shyamashree Banerjee
(Department of Biotechnology, The University of Burdwan, Burdwan, 713104, West Bengal, India.)

Abstract: *Molecular evolution is the change of amino acid composition at locus specific positions of proteins for maintaining structural and functional integrity over long periods of time. In orthologous protein sets, adapted in different environmental conditions such as normal, high temperature and high ionic condition might have different or identical evolutionary characters. To understand the basis of this fact and to assess the differentials, if any are there, I perform detailed computational analysis on clusters of orthologous protein sequences procured from halophiles, thermophiles and mesophiles using authentic web based and our laboratory-developed programs. Results reveal domain of lives and orthologous protein specific variation in these sets. Unlike thermophiles which show higher usage of hetero-pairs with dominant pair from hydrophobic ones, halophiles, show lower usage with dominant hetero-pairs as* ***ED****. Maximally diverse residue is from bulky hydrophobic class in thermophiles and that in the halophiles is acidic ones. Overall, the study demonstrates domain of lives and orthologous proteins specific evolutionary characteristics, the information of which has potential application in biological evolution of homologous proteins under different solvent conditions.*
Keywords: *Halophiles, Mesophiles, Molecular Evolution, Orthologous protein, Thermophiles.*
Abbreviation: *DP-dominating pair; MDR-maximally diverse residue.*

I. Introduction

Orthologous proteins have emerged due to speciation event for their evolution from common ancestor. Because of this ancestral relationship, orthologs represent the evolutionary history of species most accurately and also may perform similar function in different domains of lives such as halophiles, thermophiles and mesophiles. While proteins from mesophiles work under normal environmental conditions, bio-molecules from halophiles need extreme of ionic strength (≥2.5M monovalent salt) [1, 2, 3, 4, 5] and that in thermophiles need a temperature of 45-122 °C [6] for proper functioning. For example, the optimum specific activity for D-glyceraldehyde-3-phosphate-dehydrogenase from thermophiles and mesophiles was found to be 85 °C and 35 °C [7] respectively. What makes the difference in these two domains? Sequences of proteins are made up of twenty amino acids. The main chains of all these amino acids are identical. The difference in the side chain in these 20 amino acids incorporates diversity in structure-function relationship in proteins adapted in different environments. Substitution, deletion and insertion are the available mechanisms that tune sequences for survival in their respective environment [8]. In comparison to the insertion and deletion processes, amino-acid substitutions are well studied as its effect is comparable due to position specific changes of amino acids. Dayhoff et al. [9, 10] proposed the most influential model of amino-acid replacement. Substitution may be conservative or non conservative in nature, of which replacement of one amino acid residue with residues of similar physico-chemical properties has a far greater chance of being, accepted [11]. Sometimes, non conservative residue substitutions can also be tolerated with no loss or alteration of activity and three dimensional structures at many residue positions [12].

Generally, which position and the degree to which an amino acid site is free to vary are strongly dependent on its structural and functional importance because within a given protein, each position is under a different type and magnitude of selection pressure [13, 14]. For example, a buried position which maintains the protein's configurational stability would be under strong selection; a surface residue with no functional role would be under no, or weak selection while a residue present on active-site is hardly to accept mutations [15]. According to neutral theory of molecular evolution, amino acid positions that are under stringent selective constraints evolve more slowly and expected to be highly conserved than positions with weak constraints[16] because stringent negative selection pressure limits the number of replacement or substitutions[17]. In literature many examples of this constraint-rate relationship have been described [18]. In turn, Graur (1985) [19] claimed and illustrated that the substitution rate of a protein is mainly determined by its amino acid composition and the changeabilities of amino acids with examples of cytochrome c, cytochrome b5, ras-related genes, the calmodulin protein family, and fibrinopeptides. However, all these aspects of the evolutionary process are not easily studied with controlled experiment. Further, an evolutionary biologist is faced with the task of extracting as much information as possible from a data set that was generated under an uncontrolled natural process. In these cases,

study of computational evolutionary biology and data analysis can be crucial to gain insight into aforementioned evolutionary details.

In this work comparative analyses of evolutionary details of substitutions on three sets of orthologous proteins (each set contains 6 candidates) that are procured from halophiles, thermophiles and mesophiles are presented. Observed hetero pairs frequency, types, their usage limit, diversity and others are compared among these domains of lives. In particular, the present study investigates the evolutionary patterns and estimated the extent of divergence among functionally identical but independently evolving orthologous proteins in relation to different environmental conditions.

II. Materials And Methods

1.1. Data set and FASTA file

Each of six orthologous proteins from each of extreme halophilic, thermophilic and mesophilic domain are studied in this work. Representative sequences (total 18 i.e. 3*6) were retrieved in FASTA format from **UNIPROT** (www.expasy.org/sprot/). **BLAST** (http://web.expasy.org/blast/) was performed against each non-equivalent sequence for each domain of lives (6 per domain). A total of 30 sequences were selected for each non-equivalent sequence from halophiles, thermophiles and mesophiles for analysis. Thus a total of 540 sequences are studied [i.e. 30 (candidates sequences per non-equivalent protein)*6 (non-equivalent proteins per domain)*3 (domains of lives)].

1.2. Preparation of BLOCK-FASTA file and analysis

Each raw FASTA file of 30 candidate sequences was subjected for BLOCK FASTA file. The raw FASTA file is aligned using Clustal Omega program (http://www.ebi.ac.uk) and INDELs are manually removed.

APBEST software (http://sourceforge.net/projects/apbest/) was used for analysis of each BLOCK (and 18 BLOCKs) for extraction of amino acid substitution related information and conservation parameters.

III. Results And Discussion

Orthologous proteins sequences belong to different domain of lives that evolve in response to environmental change for achieving defined substrate specificity. Substitution, deletion and insertion are the available mechanisms for the purpose of evolution of which the former is comparable among different domains as it happens at locus specific positions. In these study six orthologous sets of proteins that are procured from halophiles, thermophiles and mesophiles are compared using detailed evolutionary parameters as extracted using APBEST (http://sourceforge.net/projects/apbest/) program. While mesophiles dwell under normal environmental conditions, halophiles and thermophiles function at near saturated salt solution and at high temperature respectively. Fig. 1 shows plot of different evolutionary parameters against six orthologous proteins from halophiles (H), thermophiles (T) and mesophiles (M). Each plot (plot A through D) is the comparison of an evolutionary parameter for three domains of lives for six different proteins. Each plot of the figure is presented in the following subsections.

Usage of hetero-pairs (E) show consistent pattern for extremophiles than mesophiles

Substitution of amino acids occurs at homologous positions by the utilization of a maximum of 190 hetero-pairs and 20 homo-pairs. In a given orthologous sequence (e.g. **NDK**); **E** acts as measure of usage of hetero-pairs. It varies from 0 to 1. While 0 indicates no evolution of the protein, 1 indicates maximum of it (Unpublished result of AK Bandyopadhyay). In all six protein families (with analysis of ≥30 sequences for each family), studies here show that thermophiles, have greater **E** values for all functionally different protein families (i.e. **Che C**, **DHFR**, **NDK**, **PCNA**, **RF1** and **MDH**) with highest for **Che C** and lowest for **MDH** than that of halophiles (Fig. 1 plot A). This observation indicates substitution of amino acids play dominant role in the former than the later. Interestingly, although thermophilic orthologous follow higher profile for **E**, all of its values are far less than unity. This restricted usage of hetero-pairs might be due to maintenance of sequence structure as parental ones. As far as the said sets of orthologous proteins are concerned, mesophilic pattern is far less uniform than its extremophilic counterparts. For example, in mesophiles the **E** value for **PCNA** is least (almost zero) and in **MDH**, unlike thermophiles and halophiles, shows sudden rise.

Unused hetero-pair (N) is almost opposite related with E value

As mentioned above, there are 190 possible hetero-pairs that could be utilized in the course of evolution for a given protein family. How these hetero-pairs are managed in evolution? Fig. 1 plot B shows fraction of unused hetero-pairs. The plot shows that the profile for thermophilic orthologous is less selective than that of halophiles. In halophiles major fraction of hetero-pairs (total 190) remains unused. It might highlight salt is more drastic stress that of high temperature. However, although plot pattern of **E** and **N** almost

reflecting to each other, both are not derived from hetero-pair frequency. While **E** is the sum of frequencies of used hetero-pairs in the evolution for a given BLOCK of sequence, **N** is simply the count (but not frequency) of unused hetero-pair types.

Ratio R bears protein family and domain specific evolutionary signature

R is the ratio of non-conservative to conservative type hetero-pair frequency. There are 99 types in the former and 91 in the later. Non-conservative substitutions are reported to be harmful in cellular processes in that such substitutions may cause diseases, loss of structure and mal functioning of proteins [20, 21]. The plot (Fig. 1 plot C) shows that **R** value is less than unity for all protein families and for all domains of lives indicate that the use of non-conservative substitution is less than the conservative ones. Remarkably, in case of halophilic **MDH** the **R** value is far lower than that of both of thermophilic and mesophilic ones. This would mean non-conservative substitutions play very selective role in halophilic evolution relative to its thermophilic and mesophilic counterparts. Overall, unlike **E** and **N** values, **R** value does not follow a generalised pattern in that thermophilic profile which intersects halophilic ones. This would mean non-conservative substitutions are important in relation to protein specific structural and functional constraints in their respective environments.

Dominant hetero-pair type and frequency contribute to overall protein properties

Do all hetero-pairs are equally important in the evolution of a given BLOCK of sequence? The answer to this question is simply no. A specific hetero-pair play dominant role for a given BLOCK. To check its variation in halophiles, thermophiles and mesophiles, plot of dominant hetero-pairs for all orthologous protein families are presented in the Fig. 1 plot D. Several points are noteworthy form the plot. Firstly, for all domains dominant hetero-pairs are conservative type. Conservative nature of dominant hetero-pair helps to retain sequence properties as ancestral one which would otherwise be harmful in case of non-conservative ones. Secondly, in halophiles dominant hetero-pair mostly **ED** types indicates most of the sequence positions are occupied with both **E** and **D** residues. At neutral pH these residues are negatively charged that help screening of deleterious effect of salt [1, 2]. Thus contend in **ED** types has direct relevance with halophilic stability of proteins. This observation apparently seems not unique for all halophilic proteins in that for **Che C** and **PCNA** the dominant hetero-pairs are **IV** and **SA** respectively. Although halophilic **PCNA** possess hydrophobic hetero-pair its normalized frequency is much lower than that of thermophilic one (**LI** in **PCNA**). Further, in these cases although **ED** is not dominant, it occurs at second highest rank. In contrast, thermophilic protein families show dominant hetero-pairs are mostly hydrophobic type with exception in case of **PCNA** and **MDH** which have **ED** types. Notably although both halophilic and thermophilic **MDH** have **ED** as dominant hetero-pair, the normalized frequency of the later is far lower than the former. Again, the same is true for **PCNA**. Thirdly, in mesophilic cases all dominant hetero-pairs are **IV** types. Comparison of observation for a given homologous protein (e.g. **RF1**) provides insight into the evolution for a given environment. In the present example dominant hetero-pair of **RF1** for halophilic is **ED** and that for thermophilic and mesophilic ones are **IV** types. This might highlight importance of salt and temperature in the evolution of **RF1** (an orthologous protein). Appearance of **ED** as dominant hetero-pair imparts stability of the protein in high salt which at high temperature destabilized the protein. In turn existence of **IV** as dominant hetero-pair may contribute to the overall stability by hydrophobic interactions. Finally, the dominant hetero-pairs as seen in different homologous proteins in these three domains of lives contribute to protein structural and functional stability in their respective environment (such as high salt, high temperature).

In some cases, candidates of dominant pair in homologous proteins have no tendency to participate as maximally diverse residues (see below; Table 1). In other words neither of the two candidate amino acids of the pair participates as maximally diverse residue. Such inert pairs are **IV** in **Che C** (MDR is **E**), **ED** in **NDK** (MDR is **A**) in halophiles; **IV** in **DHFR** (MDR is **K**), **ED** in **MDH** (MDR is **A**), **IV** in **RF1** (MDR is **L**), in thermophiles; **IV** in **DHFR** (MDR is **L**), **IV** in **MDH** (**A** is MDR), **IV** in **NDK** (**A** is MDR), **IV** in **PCNA** (**S** is MDR) in mesophiles. These dominant hetero-pairs thus have more tendencies to remain conserve than to be diverse which seems to have crucial structural role.

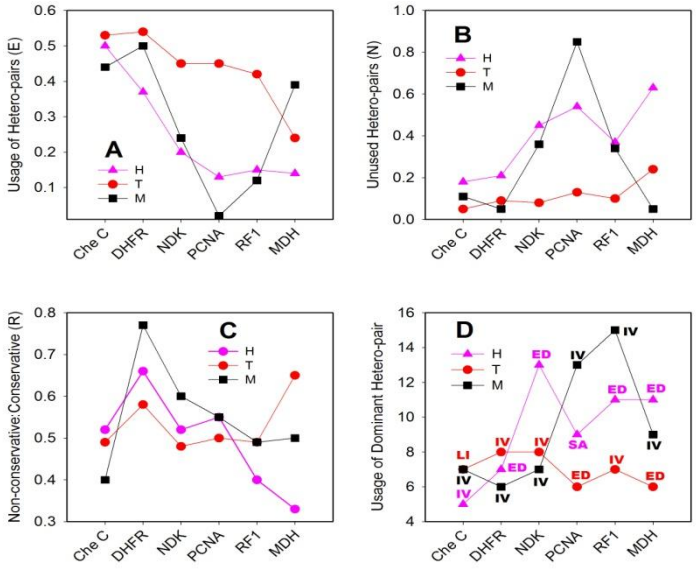

Figure 1 Plot of E (plot A), N (plot B), R (plot C) and dominant hetero-pair (plot D) values against six orthologous proteins procured from halophiles (H; violet), thermophiles (T; red) and mesophiles (M; black). In plot D hetero-pair types are also presented using appropriate color codes (violet halophilic; red thermophilic; black mesophilic) against each normalized frequency.

Amino acid diversity

Divergence or diversity is the measure of evolution which is computed by the sum of frequency of all hetero-pairs produced from a given amino acid residue. Although all amino acid residues contribute to overall diversity of a given BLOCK, their contribution varies over a wide range of frequencies. Thus there exists a maximally diverse residue which could be considered for understanding BLOCK evolution in their respective environment. Thus, it is the residue which incorporates maximum diversity in a given orthologous BLOCK. Residue maximum diversity is presented in the Table 1. Maximally diverse residue (MDR) for halophilic proteins is either **E** or **A** (Table 1, column 5) which indicate that these residues are maximally substituted by other residues in their self-dominating BLOCK positions. As the sum of frequency in maximally diverse residue also includes the dominant hetero-pair, maximally diverse residue is expected to be one of the two candidate amino acid present in the dominating pair. This is the case for **DHFR** (DP is **ED** and MDR is **E**), **MDH** (DP is **ED** and MDR is **E**), **PCNA** (DP is **SA** and MDR is **A**) and **RF1** (DP is **ED** and MDR is **E**) in halophiles. As far as substitution mechanism is concerned, in these proteins residue **E** but not **D** plays critical role in their evolution in high salt. In this respect halophilic **Che C** (DP is **IV** and MDR is **E**) and **NDK** (DP is **ED** and MDR is **A**) are exceptions. Interestingly in **Che C**, although **IV** is dominant hetero-pair, in its evolution residue **E** (highest diversity) plays critical role. In **NDK** (in respect to **Che C**) reverse situation is entertained wherein **ED** is dominant hetero-pair, residue **A** (with highest diversity) play critical role in its evolution. Overall, in high salt evolution of halophilic proteins, amino acid **E** and **A** play crucial role for maintenance of functional structure of halophilic proteins. Notably both these amino acids are seems suitable at hyper saline brine situation. **E** is negatively charged at cellular pH that has direct role in charge screening and **A** is boarder line hydrophobic residue that impart reasonable hydrophobic stability under low water activity situation (i.e. at high salt condition) [1, 2, 4]. In thermophiles, it is seen that most of the homologous proteins possess bulky hydrophobic residues as maximally diverse residues (Table 1, column 6). At high temperature hydrophobic interactions are less affected than the electrostatic ones. Thus existence of bulky hydrophobic residue in thermophilic evolution seems suitable.

As mentioned above that some dominant pairs are inert type, have more tendency to maintain the pair conservation than to be substituted by other residues, maximum diversity is thus achieved by different residues (except the candidate amino acids in dominant pair). These diversity accommodating residues seems to play crucial role in positional substitutions in evolution. In halophiles and thermophiles, such residues are less populated than mesophiles. For example, **E** in **Che C** (DP is **IV**), **A** in **NDK** (DP is **ED**) in halophiles; **K** in **DHFR** (DP is **IV**), **A** in **MDH** (DP is **ED**), **L** in **RF1** (DP is **IV**) in thermophiles and **L** in **DHFR** (DP is **IV**), **A** in **MDH** (DP is **IV**), **A** in **NDK** (DP is **IV**), **S** in **PCNA** (DP is **IV**) in mesophiles (Table 1) are diversity accommodating residues but not participate as candidate in dominant hetero-pair.

Table 1 Type and normalized frequency of dominant hetero-pair and maximally diverse residue for halophilic (H), thermophilic (T) and mesophilic (M) homologues. Type of amino acid is presented with their corresponding frequency in the parenthesis.

	Dominant hetero-pair			Maximally diverse residue		
	H	T	M	H	T	M
Che C	IV (05)	LI (07)	IV(07)	E(17)	L(24)	I (20)
DHFR	ED (07)	IV (08)	IV(06)	E(25)	K(20)	L(20)
MDH	ED (11)	ED (06)	IV(09)	E(26)	A(21)	A(21)
NDK	ED (13)	IV (08)	IV(07)	A(28)	V(20)	A(26)
PCNA	SA (09)	ED (06)	IV(13)	A(25)	E(22)	S (39)
RF1	ED (11)	IV (06)	IV(15)	E(30)	L(20)	V(21)

H halophilic; T thermophilic; M mesophilic; amino acids are expressed as single letter codes.

Class specific diversity

Residue specific diversity was used to constitute different class specific diversity (such as acidic, basic, hydrophobic and hydrophilic) and plotted in Fig. 2. Acidic vs. basic diversity is plotted at the left half and hydrophobic vs. hydrophilic are at the right half of the figure. Acidic and basic diversity for all functionally distinct proteins in halophiles are well separated (Fig. 2, left-half, H) with the former is much higher than the later.

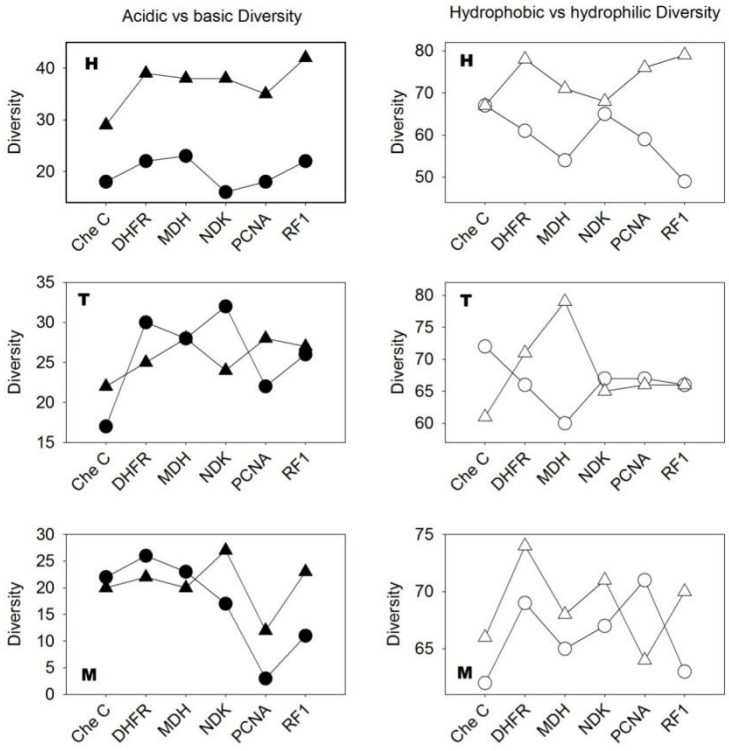

Figure 2 Acidic (Solid triangle) vs. basic (Solid circle) and hydrophobic (empty circle) vs. hydrophilic (empty triangle) diversity for halophilic (H), thermophilic (T) and mesophilic (M) proteins.

Such differentiation is not seen in case of thermophiles and mesophiles. In these cases acidic and basic diversities are more linked to protein specific functions. However, the observation in halophiles, acidic diversity play dominant role than that of basic ones in their evolution in high salt. Again the inference is seems to be unique for all proteins. Similarly, in halophiles hydrophilic diversity (Fig. 2, right half, H) dominates over hydrophobic ones. At high salt evolution of bulky hydrophobic residues is difficult due to the fact of low water activity situation [22] .In turn evolution of hydrophilic residues is utilized in halophiles. In case of thermophiles and mesophiles hydrophobic and hydrophilic diversity are not uniform for all proteins rather it is more related to protein function.

IV. Conclusion

In this study important evolutionary parameters are compared among three different domains of lives namely from halophiles, thermophiles and mesophiles using six sets of orthologous proteins. Usage of hetero-pairs in thermophiles exceed than that of halophiles for all proteins. In other word unlike halophiles, thermophiles make use of maximum hetero-pair types. Although non-conservative substitutions participate in the evolution of these proteins, conservative dominate over it (R<1). While **E** and **N** values of candidate proteins for domains of lives follow a pattern, **R** value of candidate proteins show domain specific variation. In halophiles where proteins function at high salt, **ED** acts as major dominant hetero-pair. In thermophiles and mesophiles dominant hetero-pairs are hydrophobic types. Dominant hetero-pairs are always conservative types for all domains. It can be of two kinds: first kind is non-dispensable or inert type where neither of the two candidate residues participates as maximally diverse residue and the second kind is dispensable type in that one of the two participates as maximally diverse residue. While former contributes to conservation, the later is important in further evolution of the dominant hetero-pair. Maximally diverse residue affects group diversity such as acidic, basic, hydrophobic and hydrophilic properties. Unlike thermophiles and mesophiles, halophilic shows clear separation of acidic vs. basic and hydrophobic vs. hydrophilic diversity. Overall, the study extracts and compares evolutionary parameters for six sets of proteins adapted in high salt, high temperature and normal environment.

Acknowledgements

I thankfully acknowledge the computational facility Laboratory of the Department of Biotechnology, The University of Burdwan. I also thank Dr. AK Bandyopadhyay for his help.

References

[1] A.K. Bandyopadhyay and H.M. Sonawat, Salt Dependent Stability and Unfolding of [Fe2-S2] Ferredoxin of Halobacterium salinarum: Spectroscopic Investigations, Biophysical Journal, 79(1), 2000, 501-510.
[2] A.K. Bandyopadhyay, G Krishnamoorthy and H.M. Sonawat, Structural stabilization of [2Fe-2S] Ferredoxin from Halobacterium salinarum, Biochemistry, 40(5), 2001, 1284-1292.
[3] S DasSarma and P DasSarma, Halophiles (In: eLS. John Wiley & Sons, Ltd: Chichester.pub3. 2012).
[4] J.K. Lanyi, Salt-dependent properties of proteins from extremely halophilic bacteria, Bacteriol Rev., 38(3), 1974, 272-290.
[5] M Mevarech, F Frolow and L.M. Gloss, Halophilic enzymes: proteins with a grain of salt, Biophys Chem, 86(2-3)**,** 2000, 155-164.
[6] M.T. Madigan and J.M. Martino, Brock Biology of Microorganisms (11th ed., Pearson. p. 136, 2006).
[7] A Wrba, A Schweiger, V Schultes, R Jaenicke and P Zavodszky, Extremely thermostable D-glyceraldehyde-3-phosphate dehydrogenase from the eubacterium, Thermotoga maritima, Biochemistry, 29(33), 1990, 7584-7592.
[8] J.L. Thorne, Models of protein sequence evolution and their applications, Curr Opin Genet Dev., 10(6), 2000, 602-605.
[9] M.O. Dayhoff, R.V. Eck and C.M. Park, A model of evolutionary change in proteins, In: Atlas of Protein Sequence and Structure, M.O. Dayhoff (Ed.), (National Biomedical Research Foundation, Washington, DC, 1972) 89-99.
[10] M.O. Dayhoff, R.M. Schwartz and B.C. Orcutt, A model of evolutionary change in proteins In: Atlas of Protein Sequence and Structure, M.O. Dayhoff, (Ed.), (National Biomedical Research Foundation, Washington, DC, 1978) 345–352.
[11] S French and B Robson, What is a conservative substitution? Journal of Molecular Evolution, 19(2), 1983, 171-175.
[12] A.A. Pakula and R.T. Sauer, Genetic analysis of protein stability and function, Annu Rev Genet., 23, 1989, 289-310.
[13] N.J. Tourasse and W.H. Li, Selective Constraints, Amino Acid Composition, and the Rate of Protein Evolution, Mol Biol. Evol, 17(4), 2000, 656-664.
[14] M.Y. Wolf, Y.I. Wolf and E.V. Koonin, Comparable contributions of structural-functional constraints and expression level to the rate of protein sequence evolution, Biology Direct 7(3), 2008, 40.
[15] A.T. Petroczy and D.S. Tawfik, Slow protein evolutionary rates are dictated by surface–core association, Proc Natl. Acad. Sci. U S A, 108(27), 2011, 11151–11156.
[16] M Kimura, The neutral theory of molecular evolution (Cambridge University Press, Cambridge, England, 1983).
[17] M Nei, Molecular evolutionary genetics (Columbia University Press, New York, 1987).
[18] W.H. Li, Molecular Evolution (Sinauer Associates, Sunderland, Massachusetts, 1997).
[19] D Graur, Amino acid composition and the evolutionary rates of protein-coding genes, J. Mol. Evol., 22(1), 1985, 53–62.
[20] P.C. Ng, S Henikoff, Predicting the effects of amino acid substitutions on protein function, Annu Rev Genomics Hum Genet., 7, 2006, 61-80.
[21] M.R. Shen , I.M. Jones and H Mohrenweiser, Non conservative amino acid substitution variants exist at polymorphic frequency in DNA repair genes in healthy humans, Cancer Res., 58(4), 1998, 604-608.
[22] R Karan, M.D. Capes and S DasSarma, Function and biotechnology of extremophilic enzymes in low water activity, Review Aquatic Biosystems, 8(1)**,** 2012,4.