BIOINFORMATION
Discovery at the interface of physical and biological sciences

Open access

BIOMEDICAL INFORMATICS

www.bioinformation.net

Volume 13(1)

Hypothesis

Substitutional Analysis of Orthologous Protein Families Using BLOCKS

Parth Sarthi Sen Gupta[1], Shyamashree Banerjee[1], Rifat Nawaz Ul Islam[1], Vishma Pratap Sur[2] and Amal K. Bandyopadhyay[1]*

[1]Department of Biotechnology, The University of Burdwan, Golapbag, Burdwan, 713104, West Bengal, India; [2]Indian Institute of Chemical Biology, Animal House (IICB), Kolkata, West Bengal, India; Email: akbanerjee@biotech.buruniv.ac.in; Phone - +91-342-2657231(O), 9474723882(M), Fax - +91-3422657231; *Corresponding author

Received December 14, 2016; Revised December 26, 2016; Accepted December 27, 2016; Published January 19, 2017

Abstract:
Orthologous proteins, form due to divergence of parental sequence, perform similar function under different environmental and biological conditions. Amino acid changes at locus specific positions form hetero-pairs whose role in BLOCK evolution is yet to be understood. We involve eight protein BLOCKs of known divergence rate to gain insight into the role of hetero-pairs in evolution. Our procedure APBEST uses BLOCK-FASTA file to extract BLOCK specific evolutionary parameters such as dominantly used hetero-pair (***D***), usage of hetero-pairs (***E***), non-conservative to conservative substitution ratio (***R***), maximally-diverse residue (***MDR***), residue (***RD***) and class (***CD***) specific diversity. All these parameters show BLOCK specific variation. Conservative nature of ***D*** points towards restoration of function of BLOCK. While ***E*** sets the upper-limit of usage of hereto-pairs, strong correlation of ***R*** with divergence-rate indicates that the later is directly dependent on non-conservative substitutions. The observation that ***MDR***, measure of positional diversity, occupy very limited positions in BLOCK indicates accommodation of diversity is positionally restricted. Overall, the study extract observed hetero-pair related quantitative and multi-parametric details of BLOCK, which finds application in evolutionary biology.

Keywords: evolution, substitution, non-conservative, conservative, hetero-pairs, divergence rate.

doi: 10.6026/97320630013001

Background:

Homologous proteins, emerged due to speciation event, are structurally and functionally similar **[1].** Evolution accommodates changes in these sequences. Amino acid changes are mostly achieved by substitution, deletion and insertion mechanisms **[2],** of which earlier is the result of accumulation of changes at locus specific positions. In evolution, two types of substitutions namely conservative and non-conservative occur of which most of the later changes are deleterious. Thus these are eventually eliminated through purifying selection. Beneficial ones (both conservative and non-conservative) are restored in sequence population and thus contribute to species differentiation **[3].** Comparison among homologous sequences of database reveals sequences of closely related species (e.g. human and mouse) are more similar than that of distantly related species (human vs. bacteria). When homologous positions (column-wise in a BLOCK) are fixed, it would be seen that each of these positions bears characteristic details. While some are invariant, other is either conservative or non-conservative type of substitutions **[3].** Henikoff and Henikoff (1992) pioneered the concept of BLOCK of sequences. A BLOCK contains homologous sequences whose allelic positions are fixed. These types of BLOCKs of different level of sequence similarity were used to develop different series of average BLOSUM matrices **[4].**

The concept divergence rate has become an important tool in the assessment of mechanisms of diversification in sequence evolution **[5].** Table values of divergence rates of few of these protein families are available **[6, 7].** Although different homologous proteins possess different divergence rate **[7, 8],** for a given family, it is constant **[9].** For example, Fibrino-peptide, a blood-clotting factor, has the highest and histone, a DNA binding protein, has the lowest divergence rate **[7, 8].** The variability in these rates are related to structural and functional requirements of these molecules **[10].** In this aspect, great deals of studies and developments are available **[6, 7, and 9].** Understanding the mechanism of substitutions largely involve comparison of locus-specific positions **[11],** for its effect on physicochemical properties

[12] and identity **[13]** or similarity **[14].** Similarity or identity scores are used for pair-wise comparison of sequence that eventually helps their alignment, finding relatedness **[14],** obtaining functional significance and constructing phylogenetic trees **[12, 15].** Further sequence-based studies also include analyses and extraction of information from INDEL regions of alignment. It is an additive alternative to substitution-mechanism for understanding protein evolution **[16].** While these studies have widened our understanding in different aspects of molecular evolution of protein sequences, the governing principles of evolution for homologous protein families in relation to acquired substitutions (i.e. the usage of observed hetero-pairs) still remain an enigma. Fundamental question concerning the non-conservative substitutions, as to how these are managed in these functionally similar proteins when they are known to be deleterious **[3, 17],** remain to be answered.

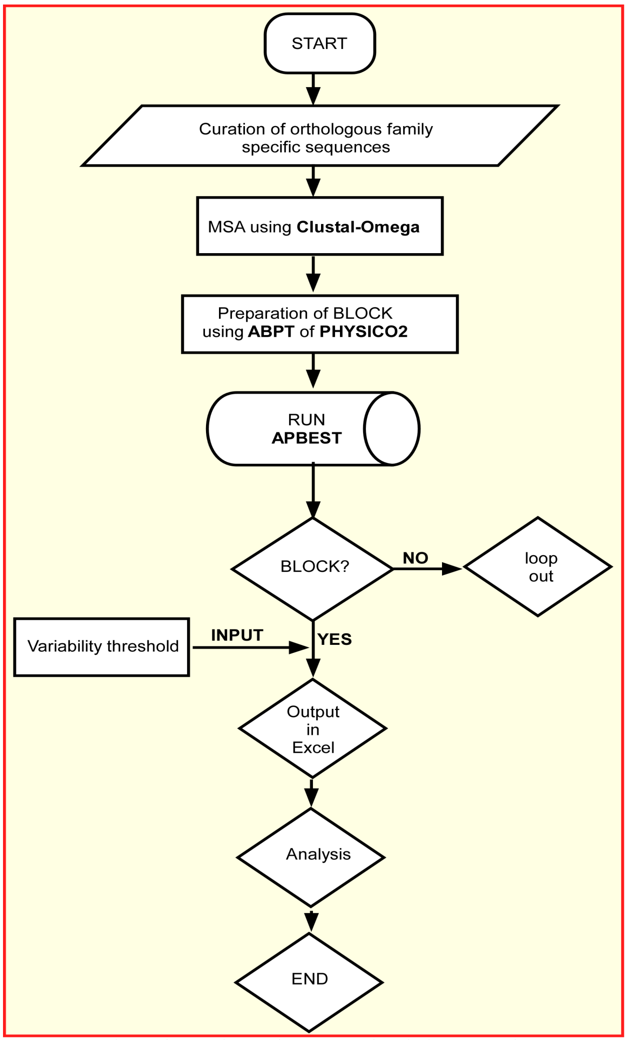

START
Curation of orthologous family specific sequences
MSA using **Clustal-Omega**
Preparation of BLOCK using **ABPT** of **PHYSICO2**
RUN **APBEST**
BLOCK?
NO
loop out
Variability threshold
INPUT
YES
Output in Excel
Analysis
END

Figure: 1 Flowchart describing methodology and operation of APBEST for extraction of analytical parameters from orthologous protein family.

In this work, we report results on SHPs (substitution-hetero-pairs) for eight protein BLOCKs of known divergence rate **[6, 7]** to work out a general model of evolution of homologous proteins. We use APBEST for efficient extraction of BLOCK parameters (***D***, ***R***, **E**, ***MDR***, ***RD*** and ***CD***). The study then shows the application of these parameters in relation to amino acid substitution of which the role of ***R*** and ***MDR*** are highlighted for the first time in this work. Overall our study extracts evolutionary parameters, the knowledge of which has potential application in understanding molecular evolution of homologous protein families.

Methodology

Collection of Data

A total of eight homologous protein families (Ubiquitin, Glyceraldehyde-3-phosphate dehydrogenase (G3PDH), Lactate dehydrogenase (LDH), Acid-protease, Hemoglobin, Ribonuclease, Somatotropin and Kappa-casein.) were taken in the present study. These families were chosen in such a way that their divergence rate give a wide coverage. For example Ubiquitin has 0.1% per 100/mYr and that for Kappa-casein is 33% per 100/mYr **[6, 7].** Family specific sequences were obtained from UNIPROT **[18],** database. Obtained sequences were then aligned using ClustalW2 **[13]**, for each of the eight protein families.

Preparation of BLOCK FASTA files

BLOCK-FASTA files were prepared using automated block preparation tool (ABPT) of PHYSICO2 **[19].** As the method involve manual step during removal of partial sequences, care was taken such that maximal sequence information is restored in the BLOCK. The BLOCK FASTA file thus produced was used as input for APBEST. An example input BLOCK FASTA file can be downloaded at (*https://sourceforge.net/projects/apbest/files/*). A flowchart starting from methodology to analysis using APBEST is shown in **Figure 1**.

Analyses of BLOCK FASTA file and extraction of evolutionary parameters

Analysis of BLOCK FASTA files was performed using in house procedure APBEST. The program is written in AWK-programming-language and runs in CYGWIN-UNIX like environment. It is efficient, error free and user-friendly. A compact itemized (Item A through F) output is redirected in excel file. It is freely available at *http://sourceforge.net/projects/APBEST/* for academic users. ***D***, ***R***, **E**, ***MDR***, ***RD*** and ***CD*** parameters were computed using relevant observed frequency of substitution-hetero-pair (SHP) (**Figure 2**). BLOCK positions undergo different types of substitutions. Different positions of BLOCK are also assessed based on residue types. If there is only one type of amino acid in a given position then it is marked as **invariant**. If substituted then qualitatively positional substitutions are assessed as different categories such as hydrophobic-hydrophobic, hydrophilic-hydrophilic and hydrophobic to hydrophilic etc.

ISSN 0973-2063 (online) 0973-8894 (print)

1. Substitution hetero-paiir frequency (***SHP***)

$$f^{B}_{XY} = \sum_{m=1}^{w} f^{m}_{XY} = \sum_{m=1}^{w}\left[{}^{n_X}C_1 * {}^{n_Y}C_1\right]_m = \sum_{m=1}^{w}\left[n_X * n_Y\right]_m \qquad \textbf{[1A]}$$

$$f^{RPF}_{XY} = \frac{f^{B}_{XY} * 100}{f^{B}_{SHP}} \qquad \textbf{[1B]}$$

Where $\boldsymbol{n_X}$ and $\boldsymbol{n_Y}$ are frequencies of amino acid **X** and **Y** respectively for a given column position (***m***) of block ***B*** of width ***w***. Block specific observed frequency ($\boldsymbol{f^B}_{\boldsymbol{XY}}$) for a given pair (**XY**) is the sum of frequencies of all column positions. $\boldsymbol{f^{RPF}}_{\boldsymbol{XY}}$ is the relative percentile frequency or probability for the hetero-pairs **XY**.

2. Substitution homo-pair frequency (***SMP***):

$$f^{B}_{ZZ} = \sum_{m=1}^{w} f^{m}_{ZZ} = \sum_{m=1}^{w}\left[{}^{n_Z}C_2\right]_m = \sum_{m=1}^{w}\left[\frac{n_Z(n_Z-1)}{2}\right]_m \qquad \textbf{[2A]}$$

$$f^{RPF}_{ZZ} = \frac{f^{B}_{ZZ} * 100}{f^{B}_{SMP}} \qquad \textbf{[2B]}$$

Where $\boldsymbol{n_Z}$ is the count of amino acid ***Z*** for a given column position (***m***).

3. Non-conservative to conservative ratio parameter (***R***)

$$R = \frac{f^{B}_{HB-HL} * 100}{(f^{B}_{HB-HB} + f^{B}_{HL-HL})} \qquad \textbf{[3]}$$

$\boldsymbol{f^B}_{\boldsymbol{HB\text{-}HL}}$ hydrophobic to hydrophilic, $\boldsymbol{f^B}_{\boldsymbol{HB\text{-}HB}}$ hydrophobic to hydrophobic and $\boldsymbol{f^B}_{\boldsymbol{HL\text{-}HL}}$ hydrophilic to hydrophilic ***SHP***s.

4. Hetero-pair usage parameter (***E***)

$$E = \frac{f^{B}_{SHP} * 100}{f^{B}_{SHP} + f^{B}_{SMP}} \qquad \textbf{[4]}$$

$\boldsymbol{f^B}_{\boldsymbol{SHP}}$ is substitution-hetero-pair frequency and $\boldsymbol{f^B}_{\boldsymbol{SMP}}$ is substitution-homo-pair frequency.

5. Residue and class-specific diversities ($\boldsymbol{RD_X}$ and $\boldsymbol{CD_{CL}}$)

$$RD_X = \sum_{k=1}^{k=19} f_{Xk} \qquad \textbf{[5A]}$$

Here ***k*** is any of 20 amino acids except **X**. ***MDR*** refer to the maximum value of $\boldsymbol{RD_X}$. ***CD***s (classes or ***CL***: acidic, basic, non-polar, hydrophobic and hydrophilic) are calculated as:

$$CD_{CL} = \sum_{j=1}^{n_{CL}}\left(\sum_{k=1}^{r} f_{j,k}\right) \qquad \textbf{[5B]}$$

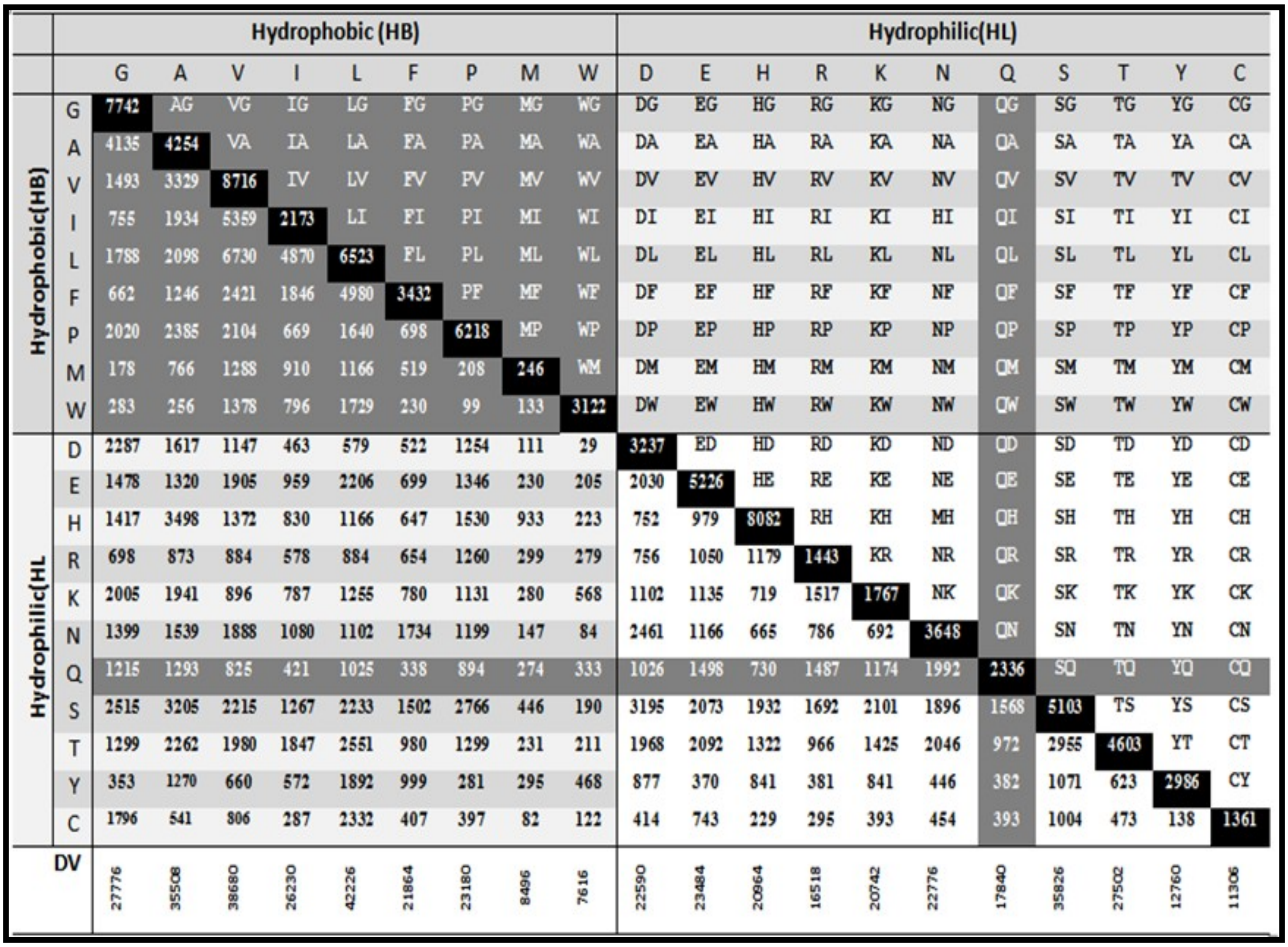
$\boldsymbol{n_{CL}}$ is the total count of class residues (for basic residues HRK, ***CL***=basic class, $\boldsymbol{n_{CL}}$=3). ***r*** is always constituted by 19 residues except the one for which ***SHP*** diversity is considered to avoid inclusion of its homo-pair.

6. Shannon Entropy (***H***)

$$H = -\sum_{i}^{M} P_i \log_2 P_i \qquad \textbf{[6]}$$

$\boldsymbol{P_i}$ is the fraction of residues of amino acid types ***i***, and M is the number of amino acid types (20 in number). Typically, positions with ***H*** >2.0 are considered variable, whereas those with ***H*** < 2 are consider conserved. Highly conserved positions are those with ***H*** <1.0.

Figure: 2 APBEST implemented equations and their clarity.



		Hydrophobic (HB)									Hydrophilic(HL)										
		G	A	V	I	L	F	P	M	W	D	E	H	R	K	N	Q	S	T	Y	C
Hydrophobic(HB)	G	7742	AG	VG	IG	LG	FG	PG	MG	WG	DG	EG	HG	RG	KG	NG	QG	SG	TG	YG	CG
	A	4135	4254	VA	IA	LA	FA	PA	MA	WA	DA	EA	HA	RA	KA	NA	QA	SA	TA	YA	CA
	V	1493	3329	8716	IV	LV	FV	PV	MV	WV	DV	EV	HV	RV	KV	NV	QV	SV	TV	TV	CV
	I	755	1934	5359	2173	LI	FI	PI	MI	WI	DI	EI	HI	RI	KI	HI	QI	SI	TI	YI	CI
	L	1788	2098	6730	4870	6523	FL	PL	ML	WL	DL	EL	HL	RL	KL	NL	QL	SL	TL	YL	CL
	F	662	1246	2421	1846	4980	3432	PF	MF	WF	DF	EF	HF	RF	KF	NF	QF	SF	TF	YF	CF
	P	2020	2385	2104	669	1640	698	6218	MP	WP	DP	EP	HP	RP	KP	NP	QP	SP	TP	YP	CP
	M	178	766	1288	910	1166	519	208	246	WM	DM	EM	HM	RM	KM	NM	QM	SM	TM	YM	CM
	W	283	256	1378	796	1729	230	99	133	3122	DW	EW	HW	RW	KW	NW	QW	SW	TW	YW	CW
Hydrophilic(HL	D	2287	1617	1147	463	579	522	1254	111	29	3237	ED	HD	RD	KD	ND	QD	SD	TD	YD	CD
	E	1478	1320	1905	959	2206	699	1346	230	205	2030	5226	HE	RE	KE	NE	QE	SE	TE	YE	CE
	H	1417	3498	1372	830	1166	647	1530	933	223	752	979	8082	RH	KH	MH	QH	SH	TH	YH	CH
	R	698	873	884	578	884	654	1260	299	279	756	1050	1179	1443	KR	NR	QR	SR	TR	YR	CR
	K	2005	1941	896	787	1255	780	1131	280	568	1102	1135	719	1517	1767	NK	QK	SK	TK	YK	CK
	N	1399	1539	1888	1080	1102	1734	1199	147	84	2461	1166	665	786	692	3648	QN	SN	TN	YN	CN
	Q	1215	1293	825	421	1025	338	894	274	333	1026	1498	730	1487	1174	1992	2336	SQ	TQ	YQ	CQ
	S	2515	3205	2215	1267	2233	1502	2766	446	190	3195	2073	1932	1692	2101	1896	1568	5103	TS	YS	CS
	T	1299	2262	1980	1847	2551	980	1299	231	211	1968	2092	1322	966	1425	2046	972	2955	4603	YT	CT
	Y	353	1270	660	572	1892	999	281	295	468	877	370	841	381	841	446	382	1071	623	2986	CY
	C	1796	541	806	287	2332	407	397	82	122	414	743	229	295	393	454	393	1004	473	138	1361
DV		27776	35508	38680	26230	42226	21864	23180	8496	7616	22590	23484	20964	16518	20742	22776	17840	35826	27502	12760	11306

Figure 3: 190 SHP types (upper-half of diagonal) and observed frequencies (lower-half of diagonal) are shown. Substitution-homo-pair frequencies (i.e. 20) are at the diagonal position. Both these types and their frequencies divided into three categories: a] **HB-HB** category: total 36 (upper dark shade), b] **HL-HL** category: total 55 (lower white shade) and **HB-HL** category: 99 in number (middle gray shade region). Residue Q is shown by gray-strip for explanation of the calculation of diversity of a given hetero-pair.

Result and discussion

To explore evolutionary and functional significance of substitution-hetero-pairs (*SHP*s) for any given homologous protein family, we have analyzed eight homologous protein BLOCKs of known divergence rate **[6, 7]**, (**Table 1**: second column) using APBEST. A representative output is available at *https://sourceforge.net/projects/apbest/files/*. It provides details of six different items (Item A through F). Items A to E compute quantitative results on substitutions. Item F provides qualitative and quantitative insight into the positional mutations and variability respectively. The study is a first time attempt to gain insight into the mechanism of substitution based on observed hetero-pairs and its diversity. It is worth noting here that, BLOSUM series of fundamental matrices made use of observed hetero-pair for the computation of odd-score **[4].** However, their use in relation to the above is rare.

In the course of evolution, observed *SHP*s, the source of diversity in BLOCK, emerge in expense of homo-pairs in the ancestral protein. A total of 20 homo-pairs (diagonal) and 190 hetero-pairs (off-diagonal) participate in this process. BLOCK specific frequency parameters such as ***R***, ***E*** and ***N***, and diversities parameters such as ***RD***, ***CD*** and ***MDR*** are presented in Table 1. Homo-pair and hetero-pair frequencies and types for a typical BLOCK are presented in **Figure 3**. Several points are noteworthy from Table 1 and Figure 3. First, type specific hetero-pair frequencies are seen to be non-identical for BLOCKs (**Figure 3**) and usage of hetero-pair (*E*) for different BLOCKs are seen to be different (**Table 1: column 5**). Second, dominantly used hetero-pair (*D*) is seen to be conservative in nature (**Table 1: column 8**). Third, residue (**Table 1: column 6-7**) and class-specific (**Table 1: column 9-13**) diversities (***RD*** and ***CD*** respectively) also show BLOCK specific variation. Interestingly, type of ***MDR*** (**Table 1:** column 6; Frequency: 18 to 26) is more versatile than that of minimally diverse residue (**Table 1:** column 7; Frequency: 0 to 2). Finally, ratio parameters (***R***, *E* and *N*) also show BLOCK specific variation.

ISSN 0973-2063 (online) 0973-8894 (print)

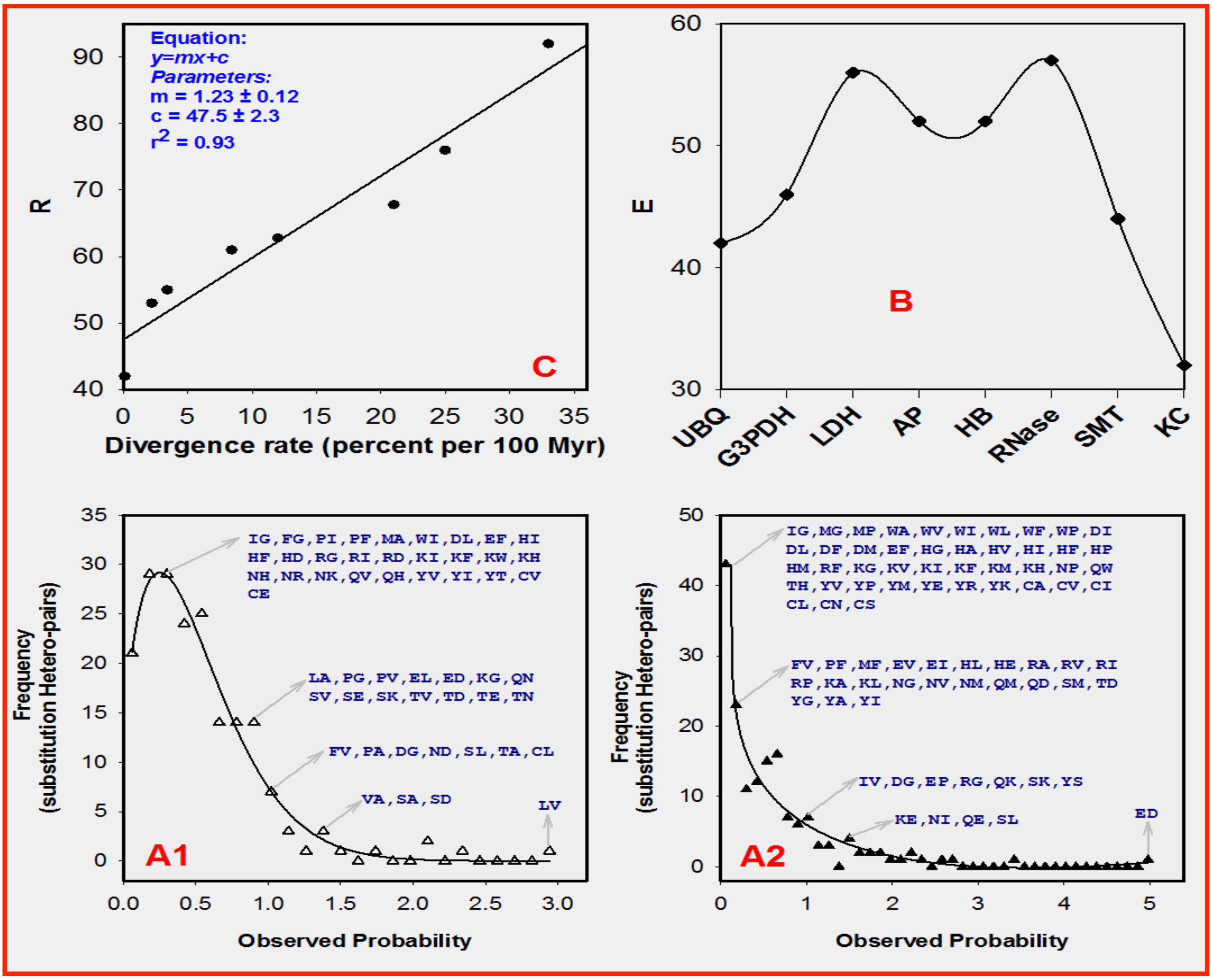


Figure: 4 Plot of derived ratios (***R*** and ***E***) vs. divergence rate and observed hetero-pair frequency vs. probability range. Two typical frequency distributions are shown (Graph A1 and Graph A2) where the observed-data are fitted with Weibull distribution function. Used (***E***) fraction of hetero-pair is plotted against divergence rate (Graph B). Graph C, shows the correlation between ***R*** with divergence rate **[6, 7].**

The fact that for a given BLOCK, individual SHP frequency varies from one another (**Figure 3**) and among BLOCKs, ***E*** also shows variation (**Table 1:** column 4), we have presented hetero-pair frequency against observed probability in **Figure 4** (plot A1 and A2). It is seen in the figure that overall distribution pattern and region specific details of observed hetero-pair types vary greatly for BLOCKs. At low probability range, observed hetero-pair frequency is very high and non-selective. As we move towards higher probability range, the frequency and type of hetero-pair become narrower and selective. For example, at highest probability range, the sole and lone observed hetero-pairs are LV and ED for plot A1 and A2 respectively (**Figure 4**). It is worth noting here that both of these are conservative types with the former is hydrophobic and the later is hydrophilic.

In evolution, functionally similar sequences (BLOCK of homologous/Orthologous sequences) are the result of substitution in the parental one. While conservation of specific sequence positions as parental one (such as active site, binding site, protein core forming region etc) is the prerequisite for functionality, evolution demands substitutions (i.e. formation of *SHPs*) at homologous positions for environmental adaptation. At the same time, lethal substitutions may lead to the malfunctioning of proteins **[3, 17].** At this point, it is worth raising the question as to what are the lower and upper limits of usage of *SHP*s. To check this, we have plotted ***E*** for BLOCKs (**Figure 4:** Plot B). In principle, ***E*** varies between 0 and 1 (**Figure2;** Equation 4). The former and the later indicate non-use and full-use of SHP respectively. However, we see the observed lower and upper limit of ***E*** are 0.3 and 0.7 respectively. Interestingly kappa-casein, that possesses highest divergence rate (**Table 1:** column 1) shows lower ***E*** value (0.32). Similar is the case for Somatotropin. Thus, the parameter ***E*** is largely uncorrelated to the divergence rate.

Is there a BLOCK specific parameter that would correlates divergence rate? In **Figure 4** (C) ***R*** is plotted and fitted against the divergence rates **[6, 17].** Notably, it is the ratio of non-conservative to conservative substitution (**Figure 2**; Equation 3). The plot shows that the parameter is positively and linearly correlated with divergence rate (correlation coefficient of 0.93). Such strong correlation of ***R*** and divergence rate indicates the former could be useful in the analysis of substitutions of orthologous protein families.

Table 1: BLOCK specific quantitative parameters for SHPs as obtained by APBEST analysis.

		Computed Ratio parameters			**Residue diversity *RD***			**Class specific diversity *CD***				
Name of Protein BLOCK	**Divergence Rate***	**R (%)**	**E (%)**	**N (%)**	***MDR or RD_MAX***	**RD_MIN**	**€ Dominant pair (*D*)**	**Acidic**	**Basic**	**Non-polar**	**Hydrophobic**	**Hydrophilic**
Ubiquitin	0.1	42	42	18	K (18.8)	C (0.5)	IV (5.1)	30.5	34.8	49.8	50.7	78.6
G3PDH	2.2	53	46	1	V (21.2)	W (1.4)	IV (9.2)	23.3	24.0	36.2	69.8	64.9
LDH	3.4	55	56	0	V (20.9)	W (1.8)	IV (7.6)	18.8	21.9	44.0	69.4	66.1
Acid-protease	9.0	62	52	2	S (23.5)	C (1.6)	IV (4.7)	16.5	13.5	57.3	68.6	69.6
Hemoglobin	12	63	52	0	L (26.7)	P (2.2)	FL (6.7)	5.61	16.2	47.2	81.5	56.9
Ribonucleases	21	67	57	0	S (23.4)	W (3.3)	TS (4.1)	15.9	38.0	56.4	58.2	81.8
Somatotropin	25	76	44	13	S (20.0)	W (0.1)	ED (5.0)	21.6	22.9	55.2	62.8	80.4
Kappa-casein	33	92	32	22	V(26.4)	C (0.6)	VA(9.6)	12.7	12.9	48.7	80.6	48.8

*percent/100 MYr; Divergence rates (second column of the table) for protein BLOCKs (first column) are taken from (Marks, 1988; Dayhoff and Schwartz, 1978). **LDH**: Lactate dehydrogenase; **G3PDH**: Glyceraldehyde 3-phosphate dehydrogenase; €Dominant pair indicates the hetero-pair type whose observed frequency is maximum for Block.

Table 2: Positional analysis of BLOCKs for invariant line (only one type of residue), conserved position (Shannon entropy ≤1.0) and type of amino acid classes (such as ***HB, Ac, Bs, PC, ST, HB+HL*** *and* ***PU+PC*)**. Normalized values (in %) are presented for comparison among BLOCKs

Blocks	**Dv Rate**	**INV**	**CONV**	**HB**	**Ac**	**Bs**	**PC**	**ST**	**HB+HL**	**PU+PC**
Ubiquitin	0.1	7.0	43.7	11.3	-	2.8	-	-	62.0	16.9
G3PDH	2.2	15.2	39.0	19.5	0.6	0.3	2.8	1.5	56.0	4.0
LDH	3.4	3.6	21.5	8.8	0.4	-	0.4	0.4	84.3	2.2
Acid-protease	9.0	1.9	24.8	5.7	-	-	1.0	-	88.6	2.9
Hemoglobin	12	-	29.0	6.5	-	-	-	-	93.5	-
Ribonuclease	21	-	10.5	5.3	-	-	2.6	-	89.5	2.6
Somatotropin	25	14.4	39.4	7.5	2.5	3.1	0.6	0.6	60.0	11.3
Kappa-casein	33	1.0	64.6	15.2	2.0	-	-	-	76.8	5.1

INV Invariant position; **CONV** Conserve position; **HB** position contains only hydrophobic amino acids; Similarly **Ac** acidic, **Bs** basic, **PC** Polar charge, **ST** serine plus threonine; **HB+HL** position contains hydrophobic and hydrophilic amino acids; similarly **PU+PC** polar uncharged and polar charged; **-** absent.

Many factors might affect BLOCK's-positional divergence or diversity. Some of these factors are positional entropy (Shannon) [20], position specific physicochemical characteristics of BLOCKs. APBEST also computes some details of which few are listed in **Table 2**. Several points are noteworthy from the Table A] Majority (≥60%) of sequence positions in BLOCKs contains mixed type (HB+HL) amino acid. Thus, HB+HL-type dominates over others such as HB, PU+PC etc. b] All but hemoglobin and ribonuclease contains invariant-lines with highest for G3PDH-BLOCK. Invariant-line does not evolve over time and are largely involves in the preservation of function of BLOCK as parental one. c] Shannon entropy is the measure of positional conservation [20]. A value ≤1.0 indicate highly conserved positions. Details of conserved positions are shown in **Table 2** (column 4). Highest and lowest conservation is seen in the case of kappa-casein (65%) and ribonuclease (11%) respectively. At this point, it is worth mentioning that kappa-casein with highest divergence rate and highest ***R-value*** shows high positional conservation (64%; Shannon entropy≤1.0). This apparent contrast of high divergence rate and high conservation of kappa casein BLOCK could be resolved by the observation that non-conservative substitutions (determinant of divergence rate) occurs only at limited and unique BLOCK positions. Such limit might allow protein to use rest of the BLOCK positions for conservation to retain function.

Conclusion

Analyses of 8 protein BLOCKs of known divergence rate shows BLOCK specific variation in the distribution pattern, hetero-pair frequency and parameters such as ***D***, ***E*** and ***R, MDR, RD*** and ***CD***. ***E*** is suitable for understanding usage limit of hetero-pairs and ***R*** is directly related with the divergence rate. Non-conservative substitution acts as determinant for the divergence rate. ***MDR*** not only contributes to class-specific-variability (***CD***-parameter) but also contributes to divergence rate. It populates only at limited BLOCK positions indicates the divergence utilizes limited portion of the total width of BLOCK. In other words, BLOCK with high conservation can still have high divergence. Such a novel strategy of limited yet unique use of positions for divergence is postulated for the purpose of incorporation of other important mechanisms of substitutions such as conservation. Taken together the procedure seems to have novel applications in substitution analysis of orthologous protein families.

Acknowledgements

We are thankful for computational facility laboratory funded by DBT, Govt. of India, in the Department of Biotechnology, The University of Burdwan.

Conflict of Interest

Authors would like to declare no conflict of interest.

References:

[1] Betts MJ & Russell RB. Bioinformatics for Genet 2007 **2**:311-42
[2] Iengar P. Nucleic Acids Res 2012 **40**:14 [PMID: 22492711]
[3] Ng PC & Henikoff S. Annu Rev Genomics Hum Genet 2006 **7**:61 [PMID: 16824020]
[4] Henikoff S & Henikoff JG. Proc Natl Acad Sci 1992 **89**:22 [PMID: 1438297]
[5] Hendry AP & Kinnison MT. Genetica 2001 **112**:1 [PMID: 11838760]
[6] Marks J. Columbia University Press New York 1988
[7] Dayhoff MO & Schwartz RM. In Atlas of protein sequence and structure 1978 345-52.
[8] Dickerson R E. J Mol Evol 1971 **1**:1 [PMID: 4377446]
[9] dos Reis M *et al.* Nat Rev Genet 2016 **17**:2 [PMID: 26688196]
[10] Tourasse NJ & Li WH. Mol Biol Evol 2000 **17**:4 [PMID: 10742056]
[11] Marini NJ *et al.* PLoS Genet 2010 **6**:5 [PMID: 20523748]
[12] Baxevanis AD & Ouellette BF. John Wiley & Sons, New Jersey 2004
[13] Larkin MA *et al.* Bioinformatics 2007 **23**:21 [PMID: 17846036]
[14] Altschul SF *et al.* J Mol Biol 1990 **215**:3 [PMID: 2231712]
[15] Gabaldón T & Koonin EV. Nat Rev Genet 2013 **14**:5 [PMID: 23552219]
[16] Ajawatanawong P & Baldauf SL. BMC Evol Biol 2013 **13**:140 [PMID: 23826714]
[17] Chun S & Fay JC. Genome Res 2009 **19**:9 [PMID: 19602639]
[18] UniProt Consortium. Nucleic Acids Res 2008 **36**: D190-D195 [PMID: 18045787]
[19] Banerjee S *et al.* Bioinformation 2015 **11**:7 [PMID: 26339154]
[20] Shannon CE. ACM SIGMOBILE Comput Com Rev 2001 **5:** 1.

Edited by P Kangueane
Citation: **Sen Gupta,** Bioinformation 13(1): 1-7 (2017)