[Frontiers in Bioscience 12, 3367-3376, May 1, 2007]

Mammalian specific mouse genes are evolving faster than mouse genes conserved across other eukaryotic lineages

Iti Chaturvedi[1], Mya Myintzu Hlaing[2], Lim Chu Sing[2], Kishore R. Sakharkar[3], Meena Kishore Sakharkar[1,2]

[1]Advanced design and modeling lab, Affiliated staff BioInformatics Research Centre, Nanyang Technological University, Singapore, [2] BioMedical Engineering Research Centre (BMERC), Nanyang Technological University, Singapore, [3]National University Medical Institutes, Yong Loo Lin School of Medicine, National University of Singapore, Singapore

TABLE OF CONTENTS

1. ABSTRACT

Positive selection is usually considered in the context of a higher rate of substitutions in non-synonymous as compared to synonymous sites in complete coding sequences of genes or individual positions. We show that genes conserved in eukaryota, coelomata, and bilateria, that is, proteins that arose earlier in evolution as compared to mammalia specific genes evolve slowly and are subjected to negative selection. This finding supports the notion that evolutionary rates progressively diminish with the age of a gene. The data suggests that in both intron-containing and intronless genes synonymous sites may be subject to some degree of selection that is indicative of a relative acceleration of amino-acid substitution, which could be due to a relaxation of functional constraints and/or directional selection.

2. INTRODUCTION

Multiple, complete genome sequences from taxonomically diverse species create unprecedented opportunities for new phylogenetic approaches (1). Studying the evolution of protein coding regions of genes has increased our understanding of the selective pressures that have shaped organisms fitness. As a consequence, a large amount of sequence data for these regions and many comparative studies of nucleotide sequence for these regions have been published. Nucleotide substitutions in protein coding regions are divided into two classes, ones that change amino acid (non-synonymous) and those that do not (synonymous). For protein coding sequences, the synonymous rate (K_s) is often regarded as a measure of the underlying mutation rate (2), though it may be influenced by other factors (3). By contrast, the non-synonymous rate (K_a) or the ratio K_a/K_s (which corrects for variation in K_s

among proteins) is often regarded either as a measure of the amount of purifying selection on the protein or as a measure of the amount of positive selection. For most genes, non-synonymous rates are lower than synonymous rates and are much more variable from gene to gene; this is thought to reflect differences in the extent of selective constraint and purifying selection among proteins (4, 5). By contrasting silent (synonymous) substitutions among amino acid altering (non-synonymous) substitutions, it is possible to detect the different selective forces acting on a protein. Genes with relatively low K_a/K_s ratios have been subject to negative (or purifying) selection, in contrast, genes with high ratios have been subject to positive (or adaptive) selection. Recently, Agarwal performed an analysis on human intronless genes from GSEGE database and reported that mammalia specific intronless genes evolve faster (6, 7). However, to perform an analysis on functional characteristics and annotations of genes it is essential to remove non-functional pseudogenes and get a "clean" list of protein coding genes. Pseudogenes are complete or partial copies of genes unable to code for functional polypeptides (8, 9). According to the theory of neutral evolution (4), pseudogenes are unconstrained by selection. Therefore, over time they randomly accumulate mutations (insertions, deletions, and substitutions) that often cause disruptions of the original reading frame. The identification of pseudogenes has also become a necessary component of the primary genome annotation in Metazoans, mainly because of their significant (up to 20%) mis-incorporation into gene collections (10, 11, 12). K_a/K_s ratios of pseudogenes and those of the vast majority of genes are generally different, as mutations in genes causing amino acid replacements with functional consequences are selected against, in contrast to mutations occurring in pseudogenes. Moreover, it is also essential to study the evolutionary rates in mammalia specific intron-containing genes. In this report, we propose a methodology to remove putative pseudogenes (based on a procedure by Harrison *et al*) and use full-length cDNA sequences to confirm the gene's structure, expression and function before performing analysis on evolution (13). This approach makes the list of identified genes more reliable and accurate. It also circumvents the greatest challenges in using EST databases to understand gene structure and expression.

This study using evolutionary rates in human and mouse orthologous genes examines whether the mouse genes classified in different eukaryotic specific lineages evolve at similar rates. The results show that there are significant differences in evolutionary rate among all eukaryotic lineages with mammalia specific mouse genes being the most rapidly evolving. The data also suggests that both mammalia specific intron-containing and intronless genes in mouse evolve at similar rates.

3. MATERIALS AND METHODS

3.1. Dataset Creation

GenBank format files for *Mus musculus* (mouse) genome were downloaded from NCBI (27th December, 2005), (National Centre for Biotechnology Information) to create a dataset on ''intronless'' and "intron-containing" genes based on the CDS FEATURE table annotation (14). We identified 5076 intronless and 21849 intron-containing genes that coded for proteins (Table 1).

3.2. Identification of expressed genes

The mouse full-length cDNA sequences were downloaded from the MGC (Mammalian Gene Collection) at ftp://ftp1.nci.nih.gov/pub/MGC/. The MGC contains full-length open reading frame for mammalian genomes (including mouse) (15). We compared the mouse intronless and intron-containing genes against the MGC using BLASTN at a low E-value cutoff of 10^{-50}. We identified 2150 intronless genes and 16275 intron-containing genes that had a hit in MGC. In the present study, analysis has been restricted to only those genes that have homologous protein coding sequence in human. Genes that had more than one homologous sequence in human genome were eliminated.

3.3. Human Orthologs and substitution rates

HomoloGene is a system for automated detection of homologs among the annotated genes of several completely sequenced eukaryotic genomes including mouse (16). HomoloGene data was downloaded from NCBI (Build 41). Match against the HomoloGene identified 666 intronless and 11233 intron-containing mouse genes that had human homologs with K_a and K_s value less than 1 (16). We discarded pairs with high substitution rates $K_a \geq 1$ or $K_s \geq 1$. This gene list represents the curated/validated dataset of genes and was subject to further evolutionary analyses. The K_a, K_s and K_a/K_s values were calculated and verified by statistical procedures.

3.4. Effective number of codons (ENC)

A more direct measure of synonymous codon bias was proposed by Wright (1990). ENC measures the effective number of codons by a method used in population genetics to determine the effective number of alleles segregating in a population. ENC measures bias in an intuitively clear manner. The larger the variety of synonymous codons used by a gene, the larger is ENC. The minimum expected value is 20 and the maximum is 61 (though actual values sometimes fluctuate outside these limits). Random synonymous codon usage should lead to ENC values close to 60. A gene with ENC value equal to 20 uses only one type of codon for each synonymous codon set and thus it shows the strongest codon usage bias whereas a gene with ENC equal to 61 indicates no synonymous codon usage preference. Effective number of codons (ENC) values was obtained for all sequences. In addition $GC3_S$, the frequency of use of G + C in synonymously variable third position of codon was calculated using codonW (http://codonw.sourceforge.net//culong.html).

3.5. Protein family classification

Subsequently, functions were assigned using PFAM database to the top20 protein families represented in mouse intronless and intron-containing genes (17). Mean K_a values and mean ENC values were calculated for these families from HomoloGene and CodonW.

Table 1. Expressed Intronless and Intron-containing genes in Mouse

Mouse	Intronless	Intron-containing
# of genes	5076	21849
# having a hit in MGC at $1e^{-50}$	2150	16275
% expressed	42.36%	74.49%

4. RESULTS AND DISCUSSION

The protein sequence evolutionary rate, which can be effectively measured as the number of nonsynonymous substitutions per nonsynonymous site (K_a), is indicative of the intensity of the selective forces acting on a protein. A large number of genes is shared by all living organisms, whereas many others are unique to some specific lineages, indicating their different times of origin. In this article, we classify the mouse genes into two subgroups the intron-containing and intronless genes. We further subdivide each subgroup into four lineages - mammalia, eukaryota, coelomata, and bilateria and estimate evolutionary rates in mouse genes with human orthologs for each of these subdivisions. The results of our analyses are presented.

4.1. Datasets for intron-containing and intronless genes

The final dataset for the subsequent analysis contains 666 intronless genes with human orthologs in total with 495, 97, 37, and 37 genes in mammalia, eukaryota, coelomata, and bilateria, respectively and 11233 intron-containing genes with human orthologs in total with 6251, 2590, 1127, and 1265 genes in mammalia, eukaryota, coelomata, and bilateria. These datasets were used to calculate the evolutionary rates.

4.2. Evolutionary rates of intron-containing genes and intronless genes

In order to characterize the evolution of a DNA sequence, one needs to know how fast it evolves, i.e., calculation of rate of nucleotide substitution. Knowing the rate of nucleotide substitutions enables us to date evolutionary events such as divergence between species or higher taxa. It is observed that the K_a value distribution of orthologous pairs classified into four phylogenetically distinct groups - mammalia, eukaryota, coelomata, and bilateria are significantly different (Figure 1A and Figure 1B). It is interesting to note that compared to mouse intron-containing and intronless genes conserved in eukaryota (K_a = 0.049, 0.036), coelomata (K_a = 0.060, 0.051), and bilateria (K_a = 0.050, 0.060), mammalia specific genes (K_a = 0.105, 0.088) exhibit almost two times higher non-synonymous rates.

The synonymous rate is also observed to be higher in case of mammalia specific mouse intron-containing and intronless genes (K_s = 0.616, 0.613) compared to (K_s = 0.602, 0.572) for eukaryotes; (K_s = 0.597, 0.566) for coelomata; and (K_s = 0.588, 0.615) however, to a lesser degree. The average evolutionary rates of intron-containing genes and intronless are shown in Table 2A and Table 2B, respectively.

Compared with intron-containing genes, intronless genes on average have lower non-synonymous rates and exhibit a smaller degree of variation in all the four sub-divisions. The synonymous rates are also lower in intronless genes but to a lesser degree. However, these differences are not significant. As expected, it is observed that the rate of non-synonymous substitution is generally lower than the rate of synonymous substitution in each category Table 2A and Table 2B both for intron-containing and intronless genes. The distribution profile for K_a and K_s for intron-containing genes and intronless genes shown in Figure 1A, 1B, respectively and Figure 2A, 2B respectively, shows variation in different groups, with skew towards higher value in case of mammalia specific human-mouse orthologs for intron-containing and intronless genes, indicating towards their recent origin. This observation is congruent with the previous report that vertebrate-specific genes evolve faster (18).

4.3. Why do mammalian specific genes evolve faster in general than other genes?

To answer this question, we examined several factors that might affect evolutionary rates. First, the higher non-synonymous substitution rates in mammalia specific genes could be due to higher mutation rates. However, this explanation is unlikely because if higher mutation rates are the main reason, we expect to see much higher synonymous substitution rates in mammalia specific genes than in genes conserved across other subgroups. The average synonymous rate in mammalia specific genes does not exhibit the same magnitude of increase as the average non-synonymous substitution rate (10% increase vs. 100% increase), suggesting that mutation rate differences are not the main cause for the non-synonymous rate difference. The differences among the rates of sequence divergence for different sub-groups are more pronounced for K_a than for K_s, which suggest that the acceleration of a gene's divergence rate may be mainly caused by more relaxed purifying selection against amino acid replacement.

Second, non-mammalia specific genes might be under stronger selective constraints than mammalia-specific genes. To compare selective constraints on these two types of genes, we calculated the K_a/K_s of these genes. As K_a/K_s has commonly been used as an indicator of selective constraint. K_a/K_s is expected to increase as the level of negative selection decreases and as the level of positive selection increases. The ratio K_a/K_s (*i.e.*, the rate of nonsynonymous substitutions corrected for neutral rates) showed a trend similar to K_a, namely, the values of K_a/K_s for genes of mammalian sub group are higher than those for genes of other sub groups. The average K_a/K_s for non-mammalia specific genes is <0.097 for both intron-containing and intronless genes, whereas it increases more than 1.5 fold for mammalia-specific genes (mean K_a/K_s = 0.149 for intron-containing and K_a/K_s =0.140 for intronless genes).

The Mann-Whitney U test shows that the average K_a/K_s is statistically higher for the mammalia specific genes (Table 3) indicating that, on average, mammalia genes are under weaker selective constraints, thus supporting the hypothesis that both mammalia specific intronless and intron-containing genes are evolving faster as compared to

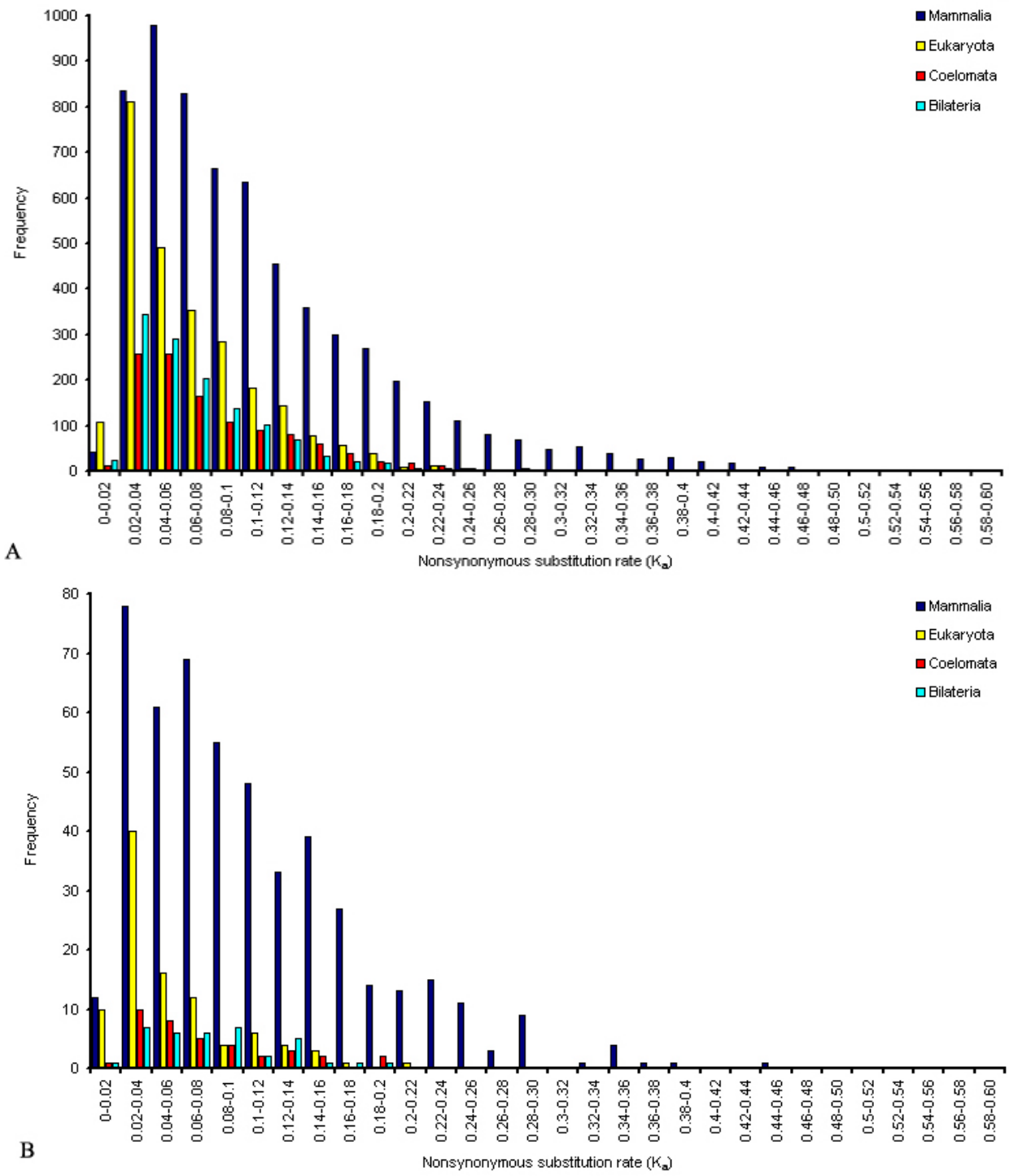

Figure 1. A: Distribution of K_a in Intron-containing Mouse genes, B: Distribution of K_a in Intronless Mouse genes.

other lineage specific genes. Selective constraint differences could arise from differences in gene function. Therefore, to get further insight into the connection between protein evolutionary rate and function of mammalia specific intronless and intron-containing genes, the K_a values of proteins associated with different PFAM annotation were compared.

4.4. Functions of representative genes in mammalian sub group

The rates at which substitutions accumulate in a protein sequence are known to vary widely in different protein families. Therefore, it is of interest to explore the function of the mammalia specific intron-containing and intronless genes along with their functions. Hence, we

Table 2. Mean(SD) of K_a, K_S, K_a/K_S and ENC in mouse

	# seq	%	K_a mean (SD)	K_S mean (SD)	K_a/K_S mean (SD)	ENC mean (SD)
A. Intron-containing genes						
Mammalia	6251	55.65	0.105(0.083)	0.616(0.162)	0.149(0.118)	52.41(1.79)
Eukaryota	2590	23.06	0.049(0.047)	0.602(0.157)	0.077(0.069)	52.71(1.66)
Coelomata	1127	10.03	0.060(0.052)	0.597(0.159)	0.097(0.076)	52.52(1.63)
Bilateria	1265	11.26	0.050(0.044)	0.588(0.156)	0.083(0.067)	52.30(1.57)
Total	11233					
B. Intronless genes						
Mammalia	495	74.32	0.088(0.073)	0.613(0.170)	0.140(0.112)	47.15(7.17)
Eukaryota	97	14.56	0.036(0.041)	0.572(0.229)	0.056(0.063)	42.26(10.89)
Coelomata	37	5.56	0.051(0.047)	0.566(0.168)	0.088(0.076)	50.16(6.06)
Bilateria	37	5.56	0.060(0.044)	0.615(0.146)	0.096(0.070)	48.39(5.88)
Total	666					

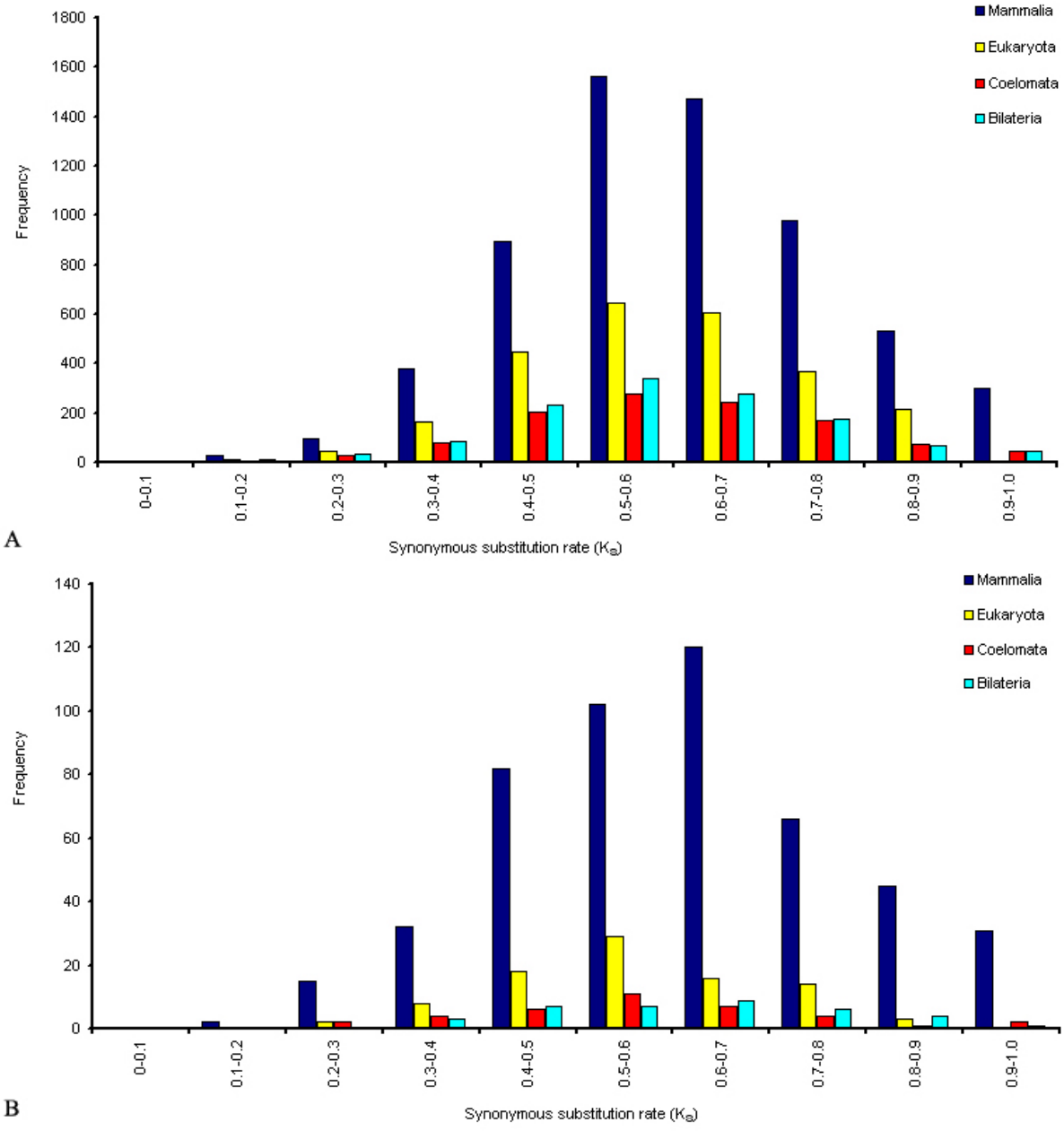

Figure 2. A: Distribution of K_s Intron-containing Mouse genes, B: Distribution of K_s in Intronless Mouse genes.

Table 3. Mann-Whitney U test of mammalian genes against different subset. NS = Non-significant, HS=Highly significant, S= Significant , Vs = Versus

Mann-Whitney U test	**Intronless**			**Intron-containing**		
Mammalia	K_a Vs	K_s Vs	K_a / K_s Vs	K_a Vs	K_s Vs	K_a / K_s Vs
Eukaryota	HS	NS	HS	HS	HS	HS
Coelomata	S	NS	S	HS	HS	HS
Bilateria	NS	NS	MS	HS	HS	HS

Table 4. Functional class evolutionary rates in mammals

A. Specific intron-containing genes					
S. No	**PFAM family**	**PFAM accession**	**Number of sequences**	**Mean K_A(SD)**	**Avg ENC**
1	Immunoglobulin domain	PF00047.10	132	0.242(0.155)	51.69
2	Fibronectin type III domain	PF00041.7	52	0.227(0.153)	52.21
3	EGF-like domain	PF00008.10	40	0.148(0.086)	52.74
4	Leucine Rich Repeat	PF00560.15	66	0.119(0.099)	51.97
5	M protein repeat	PF02370.6	35	0.119(0.079)	51.61
6	TPR Domain	PF00515.11	39	0.115(0.087)	51.96
7	bZIP transcription factor	PF00170.9	33	0.099(0.074)	53.00
8	PH domain	PF00169.12	35	0.095(0.066)	51.33
9	Zinc finger, C2H2 type	PF00096.11	184	0.091(0.075)	52.86
10	Ankyrin repeat	PF00023.12	74	0.086(0.055)	52.30
11	7 transmembrane receptor (rhodopsin family)	PF00001.7	50	0.081(0.043)	52.29
12	Ion transport protein	PF00520.13	34	0.081(0.067)	52.48
13	WD domain, G-beta repeat	PF00400.14	48	0.074(0.083)	52.35
14	Homeobox domain	PF00046.13	44	0.074(0.094)	53.71
15	PDZ domain (Also known as DHR or GLGF)	PF00595.8	36	0.073(0.064)	52.59
16	EF hand	PF00036.14	52	0.072(0.070)	52.09
17	Zinc finger, C3HC4 type (RING finger)	PF00097.9	39	0.062(0.052)	52.27
18	Protein kinase domain	PF00069.11	151	0.061(0.053)	52.64
19	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)	PF00076.7	50	0.051(0.055)	52.19
20	Ras family	PF00071.9	46	0.026(0.026)	52.51
B. Specific intronless genes					
1	Pancreatic ribonuclease	PF00074.8	5	0.209(0.170)	52.51
2	Zinc knuckle	PF00098.8	5	0.208(0.149)	50.76
3	Slow voltage-gated potassium channel	PF02060.5	5	0.189(0.162)	43.18
4	Protein of unknown function (DUF634)	PF04826.3	7	0.145(0.092)	52.42
5	Nucleosome assembly protein (NAP)	PF00956.7	4	0.143(0.099)	49.30
6	EGF-like domain	PF00008.10	4	0.142(0.114)	52.26
7	Cadherin domain	PF00028.7	13	0.128(0.059)	52.23
8	PMP-22/EMP/MP20/Claudin family	PF00822.8	8	0.117(0.104)	44.64
9	Zinc finger, C2H2 type	PF00096.11	23	0.094(0.078)	48.63
10	BTB/POZ domain	PF00651.15	5	0.085(0.066)	54.05
11	7 transmembrane receptor (rhodopsin family)	PF00001.7	35	0.081(0.051)	47.53
12	P21-Rho-binding domain	PF00786.12	4	0.079(0.056)	50.14
13	Fork head domain	PF00250.7	4	0.078(0.073)	39.90
14	Zinc finger, C3HC4 type (RING finger)	PF00097.9	6	0.072(0.076)	46.49
15	Connexin	PF00029.7	9	0.062(0.052)	46.87
16	Leucine Rich Repeat	PF00560.15	22	0.052(0.036)	48.79
17	Helix-loop-helix DNA-binding domain	PF00010.12	15	0.044(0.033)	41.75
18	Ras family	PF00071.9	6	0.029(0.024)	42.99
19	HMG (high mobility group) box	PF00505.7	6	0.012(0.012)	35.26
20	Core histone H2A/H2B/H3/H4	PF00125.7	12	0.007(0.012)	29.72

tabulated the K_a values of proteins associated with the top-20 protein families derived from PFAM annotation (Table 4A, Table 4B). It is noted that the rate of non-synonymous substitution (K_a) is extremely variable among top-20 protein families in both intron-containing and intronless genes (ranging from mean value of 0.026-0.242 for mammalia specific intron-containing genes and ranging from mean value of 0.007 – 0.209 for mammalia specific intronless genes). For example, the Immunoglobulin domain protein family with 132 sequences was found to have the highest mean K_a value, followed by the Fibronectin type III domain, EGF-like domain, Leucine Rich Repeat and M protein repeat among the intron-containing genes. On the other hand Pancreatic ribonuclease, Zinc knuckle, Slow voltage-gated potassium channel, Protein of unknown function (DUF634) and Nucleosome assembly protein (NAP) have highest mean K_a values for intronless genes. These results reinforce the previous findings that proteins which have a clear role in direct binding of pathogen antigens as part of the specific acquired immune response e.g. major histocompatibility complex (MHC) Class I (19), MHC Class II (20), and immunoglobulin heavy chain (21), Fibronectin type III domain (The extracellular region of CD45 comprises a variable domain, a linker region, and three type III fibronectin domains) are under strong selective pressures created by the need to respond to rapidly evolving pathogens with short generation times. On the other hand, housekeeping proteins like histones, HMG have been reported to evolve slower than tissue-specific ones (22).

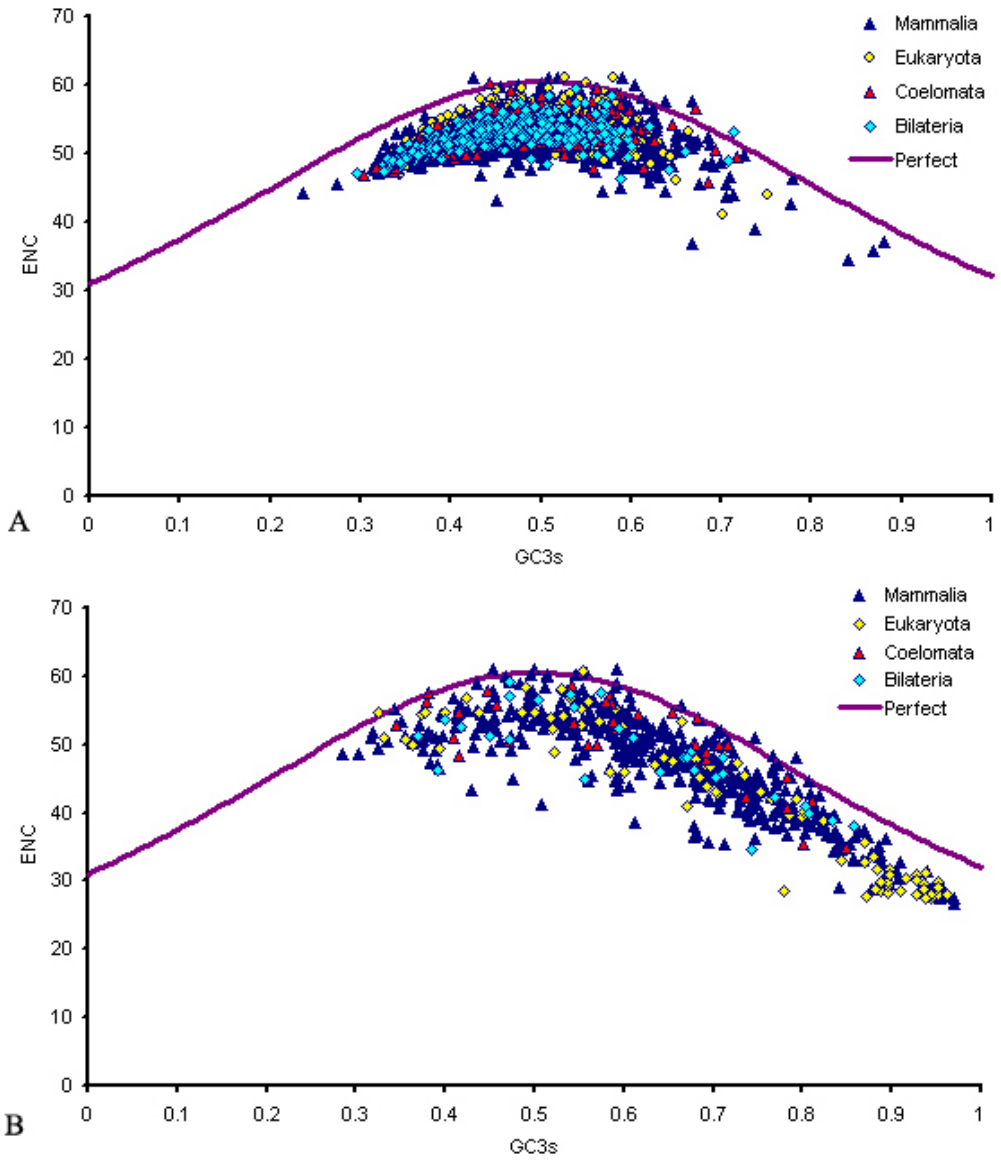

Figure 3. A: Distribution of ENC (Effective number of codons) or codon usage against GC3s (GC of silent 3rd codon position) content in Intron-containing genes. ENC when only biased by G+C content and hence under the assumption of no selection is given by solid curve line B: Similar distribution as A for Intronless genes.

The data indicates (as previously reported) that certain protein families are evolving much more rapidly as compared to others (Table 4A, Table 4B).

4.5. ENC Usage

The effective number of codons index (ENC), is a very useful preliminary tool for codon usage analysis (23). It gives the number of equally used codons that would generate the same codon usage bias as observed, lower values indicating stronger bias i.e. the smaller the ENC, the stronger the codon usage bias. The larger the variety of synonymous codons used by a gene, the larger is ENC. It is known that highly biased genes evolve far more slowly, on average presumably because selection for codon use limits the number of acceptable synonymous nucleotide substitutions in highly biased genes (24). To examine whether differences in selective constraints have any effect on codon usage, we calculated the ENC values for all genes studied. The average ENCs are 47.15, 42.26, 50.16 and 48.39 for Mammalia, Eukaryota, Coelomata and Bilateria intronless genes and 52.41, 52.71, 52.52,and 52.30 for Mammalia, Eukaryota, Coelomata and Bilateria intron-containing genes, respectively (Table 2A and Table 2B) (Figure 3A and Figure 3B). In all the categories, the ENC values are observed to be dispersed over a wide range.

If a particular gene is subject to G+C compositional constraints, it will lie on or just below the GC3s curve. It is interesting to note that although there were a small number of genes lying on the continuous plot

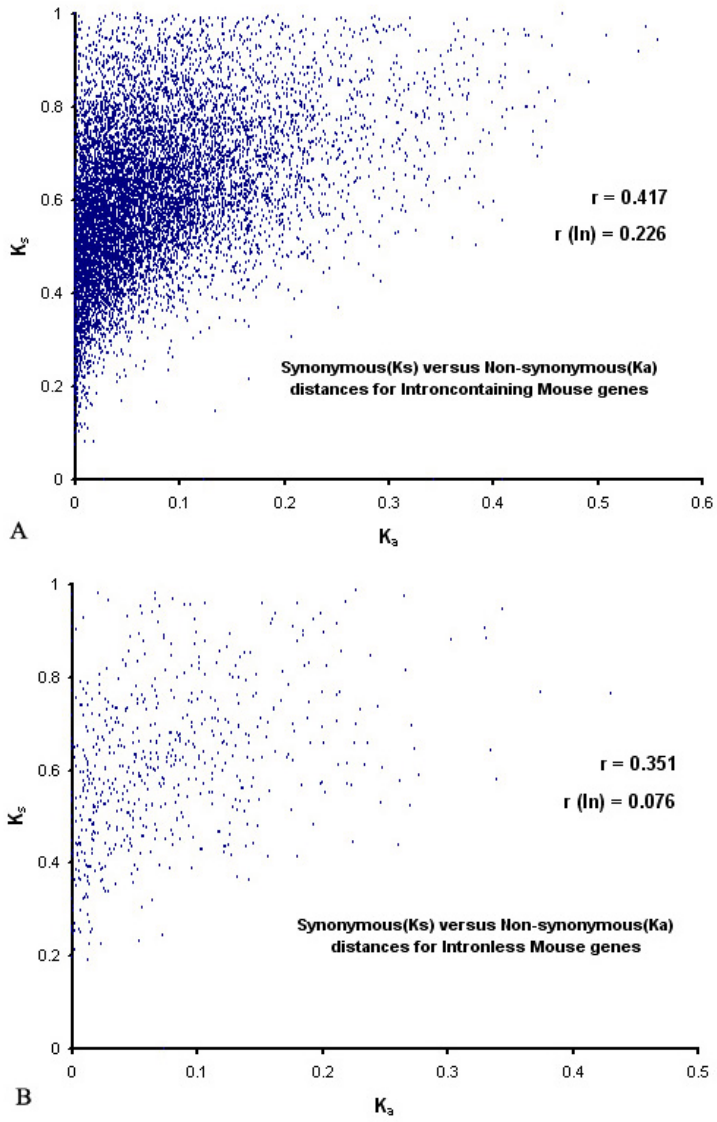

Figure 4. A: Correlation between K_a and K_s values for intron-containing genes based on homologene data, B: Correlation between K_a and K_s values for intronless genes based on homologene data.

curve, a majority of the points with low ENC values are lying well below the expected curve in both intron-containing and intronless genes, suggesting that apart from the compositional constraints other factors might have influences in dictating codon usage variation among genes (25, 26). Because almost all the ENC values of intron-containing genes (S.D ranging from 1.57 – 1.79) are much higher (ENC > 40), codon usage bias in these genes is slight. However, there is a marked variation in codon usage pattern among some intronless genes (S.D. ranging from 5.88 – 10.89). Intronless genes in general are smaller than intron-containing genes (27). Comeron, *et al* reported that long coding regions have both a lower codon bias and higher synonymous substitution rates, suggesting that they are affected less efficiently by selection (28). Thus, the bias in codon usage may be due to the smaller coding region in intronless genes.

Conversely, it has also been suggested that even some synonymous mutations are subject to constraint, often because they affect splicing and/or mRNA stability, this may be the reason for some intron-containing genes having lower ENC values (29). However, it is difficult to ascertain from this plot alone whether the variation among genes reflects different extents of selection for particular codons.

4.6. Saturation of Synonymous sites and Positive selection

We plotted K_a versus K_s and fitted linear and logarithmic regression models to the data. Saturation of synonymous sites might lead to a better fit to the logarithmic model compared with the linear one. When the linear model was applied, the correlation coefficient was r = 0.41 for intron-containing and r = 0.35 for intronless genes (Figure 4A, Figure 4B). The log model gave a

correlation coefficient of r=0.226 for intron-containing and r=0.027 for intronless genes. Thus, in both the cases the logarithmic model did not improve the linear regression model, which means that the linear model is a good explanation of the correlation of synonymous distances with respect to nonsynonymous ones. These results hint that there is no saturation of synonymous sites.

A number of cases of positive selection at the molecular level have been supported by comparison of the rates of synonymous (K_s) and non-synonymous (K_a) substitution. $K_a > K_s$ is considered clear evidence of positive selection. With this in mind, comparisons of K_s and K_a were carried out for the orthologous protein sets and the data did not reveal any cases with $K_a > K_s$. This is probably due to the fact that, if directional selection led to the acceleration of the amino-acid substitution rate in these proteins, it acted in an episodic manner (30, 31) and on a minority of residues against a constant background of purifying selection (32, 33). These results thus suggest that adaptive changes are difficult to find using the strict $K_a > K_s$ criterion and that even documented cases of positive selection can be missed using this technique alone (30, 34, 35). This correlation could reflect a mechanistic bias in mutation (36, 3) or synonymous sites may be subject to some degree of selection (37, 38, 39) (or both), that is indicative of a relative acceleration of amino-acid substitution, which could be due to a relaxation of functional constraints and/or directional selection.

5. ACKNOWLEDGEMENTS

We acknowledge funding support from A*STAR-BMRC, Singapore (Grant # 03/1/22/19/242).

6. REFERENCES

1. Huynen M. A & P. Bork. Measuring genome evolution. *Proc Natl Acad Sci USA* 95(11), 5849-5856 (1998)

2. Miyata T, T. Yasunaga & T. Nishida. Nucleotide sequence divergence and functional constraint in mRNA evolution. *Proc Natl Acad Sci USA* 77(12), 7328-7332 (1980)

3. Williams E. J & L. D. Hurst. Is the synonymous substitution rate in mammals gene-specific?. *Mol Biol Evol* 19(8), 1395-1398. (2002)

4. Kimura M. The neutral theory of molecular evolution. Sci Am. 241(5), 98-100, 102, 108 (1979)

5. Graur D & W. H. Li. Fundamentals of molecular evolution. Sinauer Assoc; ISBN: 0878932666; 2nd edition (January 2000)

6. Agarwal S. M. Evolutionary rate variation in eukaryotic lineage specific human intronless proteins. *Biochem Biophys Res Commun* 337,1192-1997 (2005)

7. Sakharkar M. K & P. Kangueane. Genome SEGE: a database for 'intronless' genes in eukaryotic genomes. *BMC Bioinformatics* 2(5), 67 (2004)

8. Vanin E. F. Processed pseudogenes: Characteristics and evolution. *Annu. Rev. Gene.* 19, 253-272 (1985)

9. Mighell A. J, N. R. Smith, P. A. Robinson & A. F. Markham. Vertebrate pseudogenes. FEBS Lett 468, 109-114 (2000)

10. Lander E. S, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* 409, 860-921 (2001)

11. Mounsey A, P. Bauer & I. A. Hope. Evidence suggesting that a fifth of annotated Caenorhabditis elegans genes may be pseudogenes. *Genome Res* 12, 770-775 (2002)

12. Mouse Genome Sequencing Consortium; R. H. Waterston, K. Lindblad-Toh, E. Birney, J. Rogers, J. F. Abril, P. Agarwal, R. Agarwala, R. Ainscough, M. Alexandersson, P. An, S. E. Antonarakis Initial sequencing and comparative analysis of the mouse genome. *Nature* 420, 520-562 (2002)

13. Harrison M. P, H. Hegyi, S. Balasubramanian, N. M. Luscombe, P. Bertone, N. Echols, T. Johnson & M. Gerstein. Molecular Fossils in the Human Genome: Identification and Analysis of the pseudogenes in Chromosomes 21 and 22. *Genome Res* 12, 272-280 (2002)

14. Sakharkar M. K, F. Passetti, J. E. de Souza, M. Long, S. J. de Souza. ExInt: an Exon Intron Database. *Nucleic Acids Res* 30(1), 191-194 (2002)

15. Gerhard D. S, L. Wagner, E. A. Feingold, C. M. Shenmen, L. H. Grouse, G. Schuler, S. L. Klein, S. Old, R. Rasooly, P. Good, M. Guyer, A. M. Peck, J. G. Derge, D. Lipman, F. S. Collins, W. Jang, S. Sherry, M. Feolo, L. Misquitta, E. Lee, K. Rotmistrovsky, S. F. Greenhut, C. F. Schaefer, K. Buetow, T. I. Bonner, D. Haussler,. The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). *Genome Res* 14(10B), 2121-2127 (2004)

16. Wheeler D. L, T. Barrett , D. A. Benson , S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, V. Miller, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner & E. Yaschenko. Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res* 33, 39–45 (2005)

17. Finn R. D, J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. Sonnhammer & A. Bateman. Pfam: clans, web tools and services. *Nucleic Acids Res* 34, 247-251 (2006)

18. Subramanian S, S. Kumar. Gene expression intensity shapes evolutionary rates of the proteins encoded by the vertebrate genome. *Genetics* 168, 373–381 (2004)
19. Hughes A. L & M. Nei. Pattern of nucleotide substitution at MHC class I loci reveals overdominant selection. *Nature* 335, 167-170 (1988)

20. Hughes A. L & M. Yeager. Coordinated amino acid changes in the evolution of mammalian defensins. *J. Mol. Evol* 44, 675-682 (1997)

21. Tanaka T & M. Nei. Positive Darwinian selection observed at the variable-region genes of immunoglobulin. *Mol. Biol. Evol* 6, 447-459 (1989)

22. Duret L & D. Mouchiroud. Determinants of substitution rates in mammalian genes: Expression pattern affects selection intensity but not mutation rate. *Mol Biol Evol* 17, 68-74 (2000)

23. Wright F. The "effective number of codons" used in a gene. *Gene* 87, 23–29 (1990)

24. Sharp P. M & W. H. Li. The codon Adaptation Index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res* 15, 1281–1295 (1987)

25. Hou Z. C. & N. Yang. Factors affecting codon usage in Yersinia pestis. *Acta Biochimica et Biophysica Sinica* 35, 580–586 (2003)

26. Gupta S. K, T. K. Bhattacharyya & T. C. Ghosh. Synonymous codon usage in Lactococcus lactis: mutational bias versus translational selection. *J. Biomol. Struct. Dyn* 21, 1–9 (2004)

27. Fedorov A, S. Saxonov, L. Fedorova & I. Daizadeh. Comparison of intron-containing and intron-lacking human genes elucidates putative exonic splicing enhancers. *Nucleic Acids Res* 29, 1464-1469 (2001)

28. Comeron J. M, M. Kreitman & M. Aguade. Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. *Genetics* 151, 239–249 (1999)

29. Chamary J.V, J. L. Parmley & L. D. Hurst. Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nat Rev Genet* 7, 98-108 (2006)

30. Messier W & C. B. Stewart. Episodic adaptive evolution of primate lysozymes. *Nature* 385, 151-154 (1997)

31. Zhang J, H. F. Rosenberg & M. Nei. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci USA* 95, 3708-3713 (1998)

32. Goodman M, G. W. Moore & G. Matsuda. Darwinian evolution in the genealogy of haemoglobin. *Nature* 253, 603-608 (1975)

33. Gonzalez D. S. & I. K. Jordan. The alpha-mannosidases: phylogeny and adaptive diversification. *Mol Biol Evol* 17, 292-300 (2000)

34. Wolfe K. H & P. M. Sharp. Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. *J Mol Evol* 37, 441-456 (1993)

35. Endo T, K. Ikeo & T. Gojobori. Large-scale search for genes on which positive selection may operate. *Mol Biol Evol* 13, 685-690 (1996)

36. Averof M, A. Rokas, K. H. Wolfe & P. M. Sharp. Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* 287, 1283-1286 (2000)

37. Miyata T & H. Hayashida. Extraordinarily high evolutionary rate of pseudogenes: evidence for the presence of selective pressure against changes between synonymous codons. *Proc Natl Acad Sci USA*, 78, 5739-5743 (1981)

38. Lipman D. J & W. J. Wilbur. Interaction of silent and replacement changes in eukaryotic coding sequences. *J Mol Evol* 21, 161-167 (1985)

39. Akashi H. Synonymous codon usage in *Drosophila melanogaster* : natural selection and translational accuracy. *Genetics* 136, 927-935 (1994)

Abbreviations:ENC: Effective Number of Codons, HS: Highly significant, MGC: Mammalian Gene Collection, NS: Non-significant, PFAM: Protein Family, S: Significant, S.D.: Standard Deviation, Vs: Versus

Key Words: Evolution, Lineage, Selection, ENC, Evolutionary Rate, Mouse, Human

Send correspondence to: Meena K. Sakharkar (Ph.D.), Advanced Design and Modeling Lab , MAE, N3-2C-113B, 50 Nanyang Avenue, Nanyang Technological University, Singapore,Tel: 65-67905836 , Fax: +65 67924062, E-mail: mmeena@ntu.edu.sg

http://www.bioscience.org/current/vol12.htm