The RAS subfamily Evolution – tracing evolution for its utmost exploitation

Ismail IF Saad[1], Saurav B Saha[2]* & George Thomas[3]

[1]Department of Zoology, Faculty of Science, Omar Al Mukhtar University, Al Bayda, Libya; [2]Department of Computational Biology and Bioinformatics, JSBB, SHIATS, Allahabad - 211007, India; [3]Department of Molecular and Cellular Engineering, JSBB, SHIATS, Allahabad - 211007, India; Saurav B Saha – Email: saurav.saha@shiats.edu.in; Phone: +91- 9621675416; *Corresponding author

Abstract:
In the development of multicellularity, signaling proteins has played a very important role. Among them, RAS family is one of the most widely studied protein family. However, evolutionary analysis has been carried out mainly on super family level leaving sub family information in scanty. Thus, a subfamily evolutionary study on RAS evolutionary expansion is imperative as it will aid in better drug designing against dreadful diseases like Cancer and other developmental diseases. The present study was aimed to understand RAS evolution on both holistic as well as reductive level. All human RAS family genes and protein were subjected to BLAST tools to find orthologs and paralogs with different parameters followed by phylogenetic tree generation. Our results clearly showed that H-RAS is the most primitive RAS in higher eukaryotes and then diverged into other RAS family members due to different gene modification events. Furthermore, a site specific selection pressure analysis was carried out using SELECTON server which showed that H-RAS, M-RAS and N-RAS are evolving faster than K-RAS and R-RAS. Thus, the results ascertain a new ground to cancer biologists to exploit negatively selected K-RAS and R-RAS as potent drug targets in cancer therapeutics.

Keywords: RAS, evolution, cancer, evolutionary tree, selection pressure

Background:
Development of multicellularity requires a sophisticated but robust machinery to communicate among different cells. In this machinery, the sensing unit is usually activated by ligand-receptor interaction, generating specific effects mostly in gene transcription, morphology, motility, adhesion and endo/exocytosis [1]. Ras (Rat sarcoma) proteins family comprises of highly conserved molecular switches involved in controlling canonical process like homeostasis, cell growth, differentiation and apoptosis [2]. These switches works by altering their conformation to Guanosine triphosphate (GTP) bound active mode to Guanosine diphosphate (GDP) bound inactive mode [3]. Their significance in human life form can easily be assessed by the fact that nearly in 25 - 30% of human tumors, RAS mutations are present [2, 4]. Its role in cancer development has raised this promising protein family as one of the most lucrative and extensively studies cancer drug target for

nearly three decades with more than 40,000 research articles published on it till year 2011 [5]. However, comparatively fewer efforts have been made to decipher its evolutionary history creating a huge lacuna in its utmost exploitation. Thus, an in-depth evolutionary study will broaden the scope of understanding of this particular target and will aid in designing better drugs against different menacingly dangerous cancers and developmental diseases.

Ras family consists of key regulatory proteins mainly conserved in eukaryotes [3]. Despite of early reports on its evolutionary presence in lower eukaryotes and prokaryotes by Robins et al. in the year 1990 [6] and Dong *et al*. in 2007 [7], a comprehensive route to its evolutionary expansion in higher and complex eukaryotes was still missing. To facilitate the understanding of higher eukaryotic evolution, RAS evolutionary study was imperative. Ras evolutionary canvas can be drawn in two levels

viz., Ras Superfamily and Subfamily. The Ras superfamily consists of five major families i.e., Ras, Rho, Arf/Sar, Ran, and Rab. However, to best of our knowledge, evolutionary studies have been conducted on RAS superfamily [3, 7, 8, 9], Rho subfamily [1] and Rab subfamily [10, 11] only, leaving rest of the family members still untouched. The present study is an endeavor to understand RAS subfamily evolutionary history and sequence structure relationship to an unprecedented in-depth level. The aim of this study was to decipher the evolutionary path followed by individual RAS subfamily members viz., K-Ras, H-Ras, M-Ras, N-Ras and R-Ras [4] during the course of evolution on both sequence and structural level.
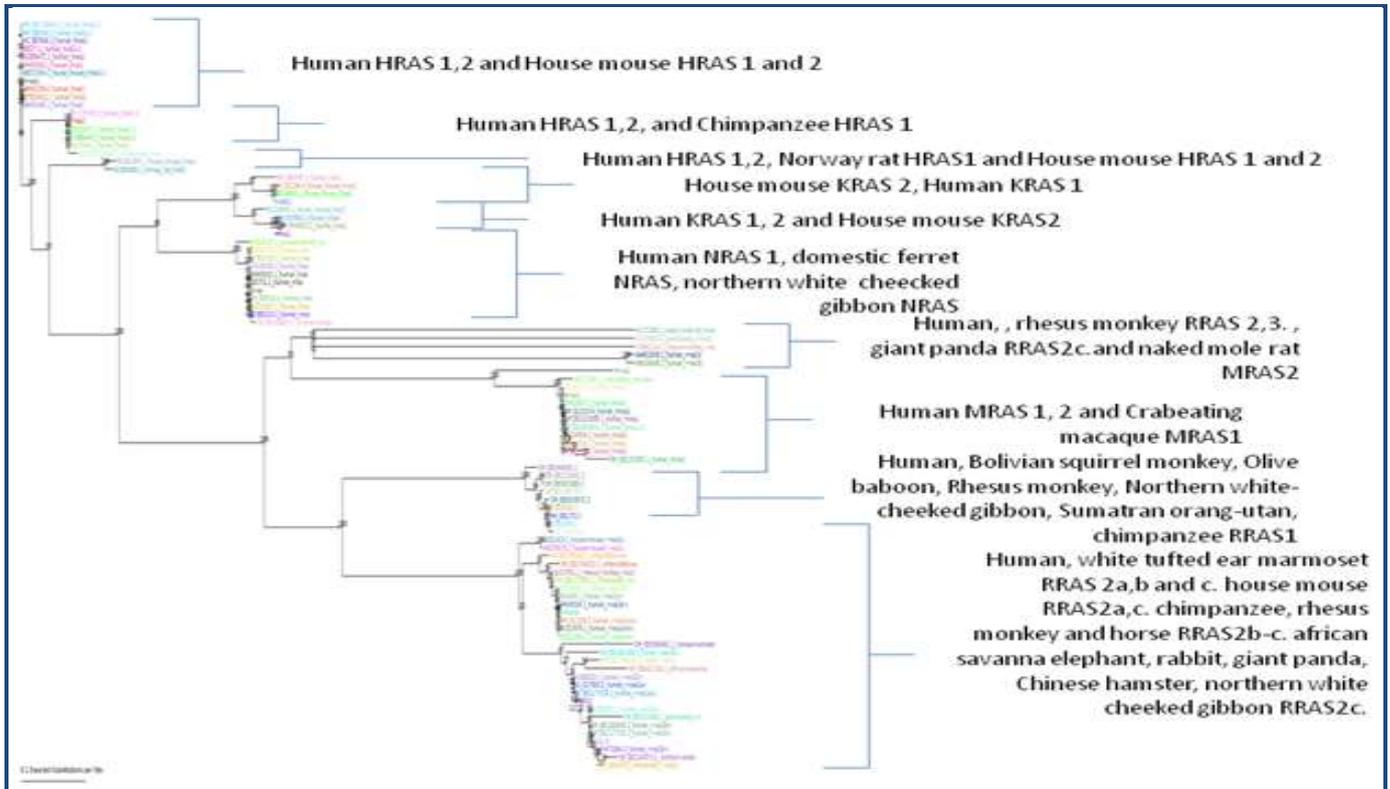

**Figure 1:** Phylogenetic tree on gene basis
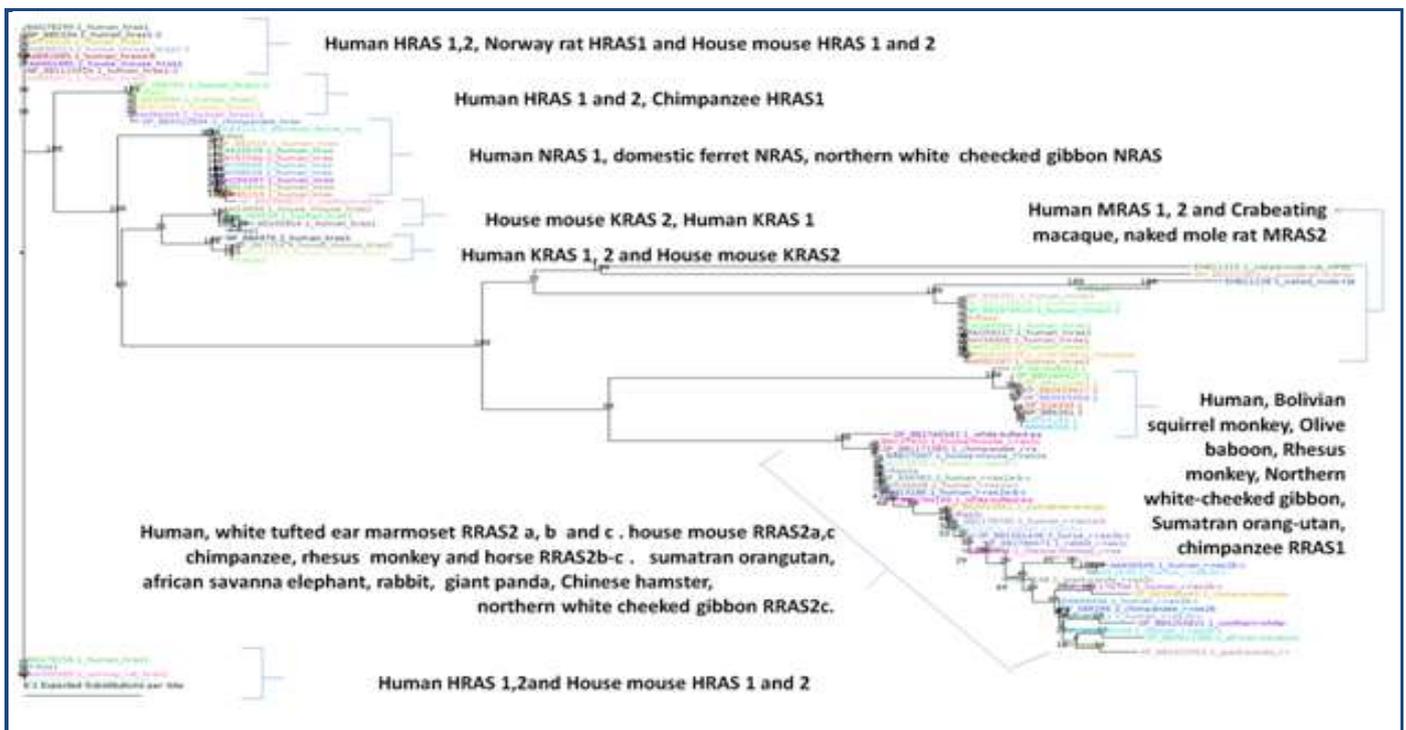
Figure 2: Phylogenetic tree on Protein basis

BIOINFORMATION

Methodology:

Tracing RAS family architecture and sequence retrieval

The gene protein architecture (forms and isoforms) of all individual RAS family members i.e., K-Ras, H-Ras, M-Ras, N-Ras and R-Ras were traced in *Homo sapiens* by generating Boolean queries against NCBI's Gene database *(http://www.ncbi.nlm.nih.gov/gene/)*. Furthermore, the Gene sequences, the Coding sequences (CDS) and the Protein sequences in fasta format were downloaded from genbank (*(http://www.ncbi.nlm.nih.gov/nuccore/)*) and genpept *(http:// www.ncbi. nlm.nih.gov/ protein/)* databases respectively.

Homolog (Ortholog and Paralog) Searching

A homolog searching was carried out for all the RAS family members. A rigid parameter was set to get only the true positive results. The searching was done against Non Redundant Database by using NCBI's BLAST tools, blastn and blastp [12]. All the subject sequences which were synthetic constructs and sequences that aligned to the queries with e-value more than 0, identities less than 100 % and coverage less than 90% were eliminated. Apart from tracing the distribution of RAS family among the different biodiversity, our endeavor was also to understand the extent of evolution in individual genomes. Therefore, a paralog sequence search was conducted against all the homologous genomes already found out through previous analysis. However, the analysis was done with flexible parameters i.e., all the sequences with their identity more than 90% to the query and spanned more than 90% of the query were considered as paralogous sequences. The flexibility in the parameters was introduced to extract all the false negative sequence from the entire raw data.

Multiple Sequence Alignment, Structure Prediction and Validation

Multiple sequence alignment was carried out through guide tree approach using windows ClustalX program [13] on both gene and protein level. To leverage the potentiality of protein structures on tracing evolutionary level, 3D structure of all the longest isoform of each individual Human RAS family members were predicted using Modeller software [14], except for RRAS1 which was modeled by I-TASSER server using ab-initio methodology [15] due to its high degree of divergence from other RAS subfamily members. The structural models were further evaluated using Structural Analysis and Verification Server's PROCHECK tool which checks stereochemical quality of a protein structure utilizing Ramachandran plot [16]. The best models from each individual isoforms were taken under consideration for further analysis.

Phylogenetic tree generation and Selection Pressure analysis

Topali software [17] was used to generate trees and clustering. Phylogenetic trees were generated in both gene level and protein level using neighbor joining method with a bootstrapping iteration value of 100. To analyze positive or negative selection of specific amino acid sites within the full-length sequences of all the RAS individuals, the protein sequence and the modeled structures were subjected to SELECTON server which calculated substitution rate ratios of nonsynonymous (Ka) versus synonymous (Ks) mutations. A Ka/Ks ratio greater than 1 suggests positive selection and a ratio less than 1 suggests purifying selection [18].

Discussion:

The existence and role of RAS family in lower eukaryotic and prokaryotic development are reported early by Dong et al, and Robbins et al. [6, 7, 19]. Moreover, Dı́ez et al., showed expansion of RAS family in complex higher order of eukaryotes like humans [3]. However, the route followed by this canonical family during the course of evolution was still eluded. This study was carried out to understand the biodiversity of RAS family and how this family might have been derived during evolution. All the form and isoforms of RAS subfamily in human were traced from NCBI gene database using Boolean queries like AND and OR **Table 1 (see supplementary material).** An ortholog search for each forms and isoforms were then carried out against Non Redundant Database with strict parameters **Table 2 (see supplementary material)**. The expansion of RAS family in complex higher organisms was clearly visible conforming Diego Dı́ez et al.'s work [3]. However, the rigid parameters also removed false negative data from results i.e., due to long evolutionary distance and time, the RAS eukaryotic forms substantially diverged from lower organisms showing no significant similarity on gene and protein level for entire stretch of sequence. Therefore, all the prokaryotic and lower eukaryotic RAS family members were subjected to BLAST all-to-all method. Surprisingly even in lower eukaryotes and prokaryotes, no significant similarity for entire stretch of sequences was found indicating a long evolutionary distance and many parallel evolutionary events among lower organisms. Thus to decipher the evolutionary path from prokaryotes to eukaryotes, gene and protein sequences were utilized to draw phylogenetic tree using neighbor joining approach followed by clustering **(Figure 1 & Figure 2).**

From both the gene and protein tree, it was evident that it was H-RAS which came from prokaryotes to eukaryotes which eventually due to evolutionary gene modification events like gene duplication, bifurcated among different RAS forms present today in just higher eukaryotes. The study clearly showed that HRAS subfamily has further evolved, diverged into NRAS on one cluster and KRAS, MRAS and RRAS in the other cluster/sub clusters. To assess the different selection pressures on all the sites of all RAS forms, SELECTON server was used. All the protein structures of longest RAS forms were modeled through homology modeling except R-RAS1 which was modeled by I-TASSER server using ab-initio methodology **Table 3** (**see supplementary material**) and then were subjected to site specific selection pressure analysis using SELECTON server **Table 4** (**see supplementary material**) [18]. While considering the distribution of calculated Ka/Ks ratio (ω), the evaluation of selection pressures indicated that maximum number of the codons in RRAS and KRAS were under purifying selection and HRAS, NRAS and MRAS showed neutral selection. It is also interesting to note the correlation on location of RAS members on the human chromosome arm i.e., the classical RAS members, HRAS (11p15.5), KRAS (12p21.1), NRAS (1p13.2) are located on the short arm (p) of human chromosome whereas MRAS (3q22.3) and RRAS (19q13.3-qter) are located on the long arm (q) of the human chromosome. The phylogenetic tree and selection pressure results showed that HRAS diverged into KRAS and NRAS on the short arm and then a gene duplication event was observed where either of the then present form duplicated to long arm of the human

chromosome with sudden evolutionary expansion giving rise to more evolutionary dynamic RRAS and MRAS. It is worth mentioning that RRAS itself comprises of 53.5% of the entire RAS subfamily bio-distribution. The maximum percentage bio-distribution density was observed in RRAS2c (24.56) followed by RRAS2b (14.91), HRAS1 (13.16), NRAS (8.78), HRAS2 (7.89), MRAS1 (7.89) and RRAS1 (7.89), RRAS2a (6.14), MRAS2 (3.52) and KRAS1 (2.63) and KRAS2 (2.63).

Conclusion:
Ras family proteins play pivotal role in key regulatory process in eukaryotes. The study on evolution of this family has immense significance in both understanding of multicellularity evolution and cancer biology. From our finding, we speculate and hypothesize that RAS family in higher eukaryotes started with H-RAS and then parallely evolved in two distinct domains viz. K-RAS with N-RAS and M-RAS with R-RAS where N-RAS and M-RAS with their higher divergence rate are involved in ever expanding RAS forms. Moreover evidence of K-RAS and R-RAS being negatively selected opens opportunity to exploit these proteins as preferred targets over other family members.

Conflict of interests:
None to declare

References:
[1] Boureux A *et al. Mol Biol Evol.* 2007 **24**: 203 [PMID: 17035353]
[2] Castellano E & Downward J, *Genes Cancer.* 2011 **2**: 261 [PMID: 21779497]
[3] Dı́ez D *et al. Nucleic Acids Res.* 2011 **39**: 5526 [PMID: 21447561]
[4] Adjei AA, *J Natl Cancer Inst.* 2001 **93**: 1062 [PMID: 11459867]
[5] Fernandez-Medarde A & Santos E, *Genes Cancer* 2011 **2** :344 [PMID: 21779504]
[6] Robbins SM *et al. Nucleic Acids Res.* 1990 **18**: 5265 [PMID: 2205839]
[7] Dong JH *et al. Gene* 2007 **396**: 116 [PMID: 17449198]
[8] Colicelli J, *Sci STKE.* 2004 **2004:** RE13 [PMID: 15367757]
[9] Rojas AM *et al. J Cell Biol.* 2012 **196**: 189 [PMID: 22270915]
[10] Stenmark H, *BMC, Biol.* 2012 **10** : 68 [PMID: 22873178]
[11] Klöpper TH, *BMC Biol.* 2012 **10**: 71 [PMID: 22873208]
[12] Altschul SF *et al. J Mol Biol.* 1990 **215**: 403 [PMID: 2231712]
[13] Thompson JD *et al. Nucleic Acids Res.*1997 **25**: 4876 [PMID: 9396791]
[14] Sali A *et al. Proteins* 1995 **23**: 318 [PMID: 8710825]
[15] Zhang Y, *BMC Bioinformatics* 2008 **9:** 40 [PMID: 18215316]
[16] Laskowski RA *et al. J Appl Crystallogr.* 1993 **26**: 283
[17] Milne I *et al. Bioinformatics* 2008 **25**: 126 [PMID: 18984599]
[18] Stern,A *et al. Nucleic Acids Res.* 2007 **35:** W506 [PMID: 17586822]
[19] Weeks G & Spiegelman GB, *Cellular Signal.* 2003 **15**: 901 [PMID: 12873703]

BIOINFORMATION

Supplementary material:

Table 1: The gene-protein architecture (forms and isoforms) of all RAS subfamily members i.e., H-Ras, K-Ras, M-Ras, N-Ras and R-Ras in *Homo sapiens* by generating Boolean queries against NCBI's Gene database

No	Ras	Isoforms	Gene	Length(bp)	Protein	Length(aa)
1	H-Ras	H-Ras1	NM_001130442.1	1169	NP_001123914.1	189
			NM_005343.2	1061	NP_005334.1	189
		H-Ras2	NM_176795.3	1251	NP_789765.1	170
2	K-Ras	K-Ras1	NM_033360.2	5436	NP_203524.1	189
		K-Ras2	NM_004985.3	5312	NP_004976.2	188
3	N-Ras	N-Ras	NM_002524.4	4454	NP_002515.1	189
4	M-Ras		NM_001085049.2	4033	NP_001078518.1	208
		M-Ras1	NM_001252090.1	4183	NP_001239019.1	208
			NM_012219.4	4609	NP_036351.3	208
			NM_001252091.1	4398	NP_001239020.1	132
		M-Ras2	NM_001252092.1	3903	NP_001239021.1	132
			NM_001252093.1	3822	NP_001239022.1	132
5	R-Ras	R-Ras1	NM_006270.3	1013	NP_006261.1	218
			NM_012250.5	2360	NP_036382.2	204
		R-Ras2	NM_001102669.2	2209	NP_001096139.1	127
			NM_001177315.1	2186	NP_001170786.1	127
			NM_001177314.1	2054	NP_001170785.1	169

Table 2: The Bio-distribution of RAS forms in terms of both orthologs and paralogs generated through BLAST tools.

No	RAS	Organism	Paralogs	Homologs	Bio-distribution (%)
1	H-RAS1	*Homo sapiens* (human)	11	15	13.16
		Pan troglodytes (chimpanzee)	1		
		Mus musculus (house mouse)	2		
		Rattus norvegicus (Norway rat)	1		
2	H-RAS2	*Homo sapiens* (human)	8	9	7.89
		Mus musculus (house mouse)	1		
3	K RAS1	*Homo sapiens* (human)	3	6	5.27
4	KRAS2	*Mus musculus* (house mouse)	3		
5	MRAS1	*Homo sapiens* (human)	8	9	7.89
		Heterocephalus glaber (naked mole-rat)	1		
6	MRAS2	*Homo sapiens* (human)	2	4	3.51
		Macaca fascicularis (crab-eating macaque)	1		
		Heterocephalus glaber (naked mole-rat)	1		
7	NRAS	*Homo sapiens* (human)	8	10	8.78
		Nomascus leucogenys (Northern white-cheeked gibbon)	1		
		Mustela putorius furo (domestic ferret)	1		
8	RRAS1	*Homo sapiens* (human)	3	9	7.89
		Pan troglodytes (chimpanzee)	1		
		Pongo abelii (Sumatran orangutan)	1		
		Nomascus leucogenys (Northern white-cheeked gibbon)	1		
		Macaca mulatta (Rhesus monkey)	1		
		Papio anubis (Olive baboon)	1		
		Saimiri boliviensis boliviensis (Bolivian squirrel monkey)	1		
9	RRAS2a	*Homo sapiens* (human)	5	7	6.14
		Callithrix jacchus (white-tufted-ear marmoset)	1		
		Mus musculus (house mouse)	1		
10	RRAS2b	*Homo sapiens* (human)	11	17	14.91
		Callithrix jacchus (white-tufted-ear marmoset)	2		
		Pan troglodytes (chimpanzee)	2		
		Equus caballus (horse)	1		
		Macaca mulatta (Rhesus monkey)	1		
11	RRAS2c	*Homo sapiens* (human)	13	28	24.56
		Callithrix jacchus (white-tufted-ear marmoset)	2		
		Equus caballus (horse)	1		
		Ailuropoda melanoleuca (giant panda)	2		
		Pan troglodytes (chimpanzee)	2		

Loxodonta africana (African savanna elephant)	1	
Nomascus leucogenys (Northern white-cheeked gibbon	1	
Oryctolagus cuniculus (rabbit)	1	
Macaca mulatta (Rhesus monkey)	2	
Mus musculus (house mouse)	1	
Pongo abelii (Sumatran orangutan)	1	
Cricetulus griseus (Chinese hamster)	1	
Total	**114**	**114**	**100**

Table 3: Modeled 3-D structures of Ras subfamily members

RAS	3D structure	Favoured regions %	Additional allowed regions %	Generously allowed regions %	Disallowed regions %
H-Ras1		81.5	14.9	2.4	1.2
K-Ras1		82.5	13.5	2.3	1.8
N-Ras		82.7	14.3	2.4	0.6
M-Ras1		82.4	14.7	1.8	1.2
R-Ras1		86.3	10.3	2.3	1.1
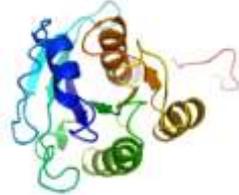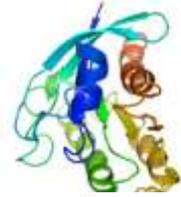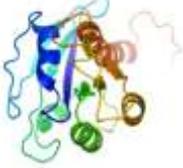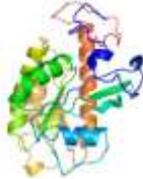
Table 4: Selection pressure on different sites of different Ras subfamily forms

No	Ras	Ras isoforms	Sequence length	0<dN/dS<1	dN/dS=1	dN/dS>1	0<dN/dS<1(%)	dN/dS=1(%)	dN/dS>1(%)
1	H-Ras	H-Ras 1	190	0	190	0	0	100	0
2	K-Ras	K-Ras1	189	180	1	8	95.23	0.534	4.23
3	N-Ras	N-Ras	190	0	190	0	0	100	0
4	M-Ras	M-Ras1	209	0	209	0	0	100	0
5	R-Ras	R-Ras1	218	218	0	0	100	0	0