Investigation of the Evolution of MutS Protein Family in Animals

EEOB 563 Final Project

Mudith Ekanayake

Introduction

Throughout different stages of the cell cycle, many DNA repair pathways act in the cells allowing the cells to repair the DNA damage. Among these pathways mismatch repair (MMR), base excision repair (BER), nucleotide excision repair (NER), homologous recombination (HR) and non-homologous end joining (NHEJ) are the major pathways that are active in the cells. More importantly DNA mismatch repair (MMR) is responsible for recognizing and repairing erroneous insertion, deletion and mis-incorporation of bases during DNA replication and recombination. In MMR, there are three major steps including mismatch identification, mismatch excision and DNA re-synthesis. Throughout this whole process several proteins are involved and among those proteins, MutS is incorporated in detecting mismatches in the sequences. These proteins are well conserved in prokaryotes, eukaryotes and even in viruses. There are many homologs of MutS protein including MutS1, MutS2 in bacteria and MSH 1 to 6 in eukaryotes. But In this project, MSH2, MSH3, MSH4, MSH5, MSH6 were selected for the analysis. There are many theories about the evolution of MSH protein and still there are some confusions which protein evolved first and then how the others got descended from that. Some studies suggest that eukaryotic mitochondrial MutS has evolved due to horizontal gene transfter (HGT) event from a large DNA virus into the mitochondrial genome (Bilewitch and Degnan, 2011) while others hypothesize eukaryotic MSH proteins have originated from a HGT event between bacteria and eukaryotes. Since mismatch repair is critical for maintaining genome stability, from this project, evolution of the MutS protein family in animals will be investigated in order to find evidence for the above mentioned theories.

Methods

MSH protein Sequences were retrieved from five different groups including Archaea, Bacteria, Metagenomics, Viruses and Eukaryotes excluding animals, plants and fungi. Protein BLAST (blastp) was carried out in NCBI for human MSH proteins from 2 to 6 against above mentioned groups and only the top 10 hits were selected for the analysis. Five separate datasets were generated for five different MSH proteins and to each dataset, a reference dataset was added which included MSH1-6 from *Saccharomyces cerevisiae*, MSH1-6 from *Nematostella*, MSH2-6 from human and mtMutS from *Dendronephthya gigantea*.

All the steps in the phylogenetic analysis were performed five times for five different MSH datasets. MAFFT (Katoh *et al*., 2002) sequence alignment program was used for aligning the protein sequences and the alignments were curated and cleaned using TrimAl (Capella-Gutierrez *et al*., 2009) selecting "automated 1" method. Prior to constructing the phylogeny, model selection step was carried out for all the alignments incorporating SMS: Smart Model Selection tool in PhyML (Lefort *et al*., 2017) using the selection criterion as Akaike Information Criterion (AIC) (Akaike, 1973).

Phylogenetic reconstruction was undertaken under a maximum likelihood framework implemented in RAxML-NG (Kozlov, *et al*., 2019) using the models selected in SMS model selection, and with confidence levels estimated using bootstrap resampling with 1000 replicates. Interactive Tree Of Life (iTOL) v4 tool (Letunic and Bork, 2019) was used for the visualization of the tree.

Results

After aligning with MAFFT, TrimAl removed poorly aligned regions from multiple sequence alignments and it significantly shortened the alignments. SMS tool selected LG+G+I+F as the best model for all the 5 MSH datasets. Table 1 shows a summary of the model selection step. Complete results are attached in the supplementary materials.

Table 1: Summary of the model selection by SMS smart model selection tool (K: number of model free parameters, Llk: Log-likelihood of the data given the substitution model and the tree, AIC: Akaike Information Criterion, BIC: Bayesian Information Criterion)

MutS Homolog	Model	Decoration	K	Llk	AIC	BIC
MSH2	LG	+G+I+F	154	-104989.82347	210287.64694	211176.78048
MSH3	LG	+G+I+F	154	-117976.37152	236260.74304	237165.86220
MSH4	LG	+G+I+F	154	-99178.91815	198665.83630	199549.16032
MSH5	LG	+G+I+F	154	-93980.49230	188268.98460	189147.66437
MSH6	LG	+G+I+F	154	-110644.83080	221597.66160	222490.88671

RAxML-NG estimated maximum likelihood values and tree topologies for the best tree found according to the selected models for each and every MutS homolog (20 tree search). The log likelihood values of the best tree and AIC, AICc, BIC scores predicted by RAxML-NG are tabulated in the table 2. The bootstrap analyses with 1000 replicates returned bootstrap support for all the nodes and the values are shown near the nodes of the trees (Figure 1-5).

Table 2: Summary of the RAxML-NG phylogeny reconstruction.

MutS Homolog	Best Tree logLH	AIC score	AICc score	BIC score
MSH2	-42591.341315	85490.682629	85637.574937	86133.445689
MSH3	-59140.561955	118589.123909	118690.054142	119273.275095
MSH4	-56558.934077	113425.868153	113522.118153	114115.558638
MSH5	-37632.010775	75572.021549	75723.577105	76211.542384
MSH6	-45121.922768	90551.845537	90685.197492	91204.848017

In the constructed phylogenetic trees, many interesting relationships could be seen. All the viral sequences were clustered in one single clade in each and every tree and interestingly, *Dendronephthya gigantea* sponge mitochondrial MutS (mtMutS) was also included in all the viral clades (Figure 1-5). The bootstrap support value for the clade was 100 at every occasion. In the figure 1, all the eukaryotic MSH2 proteins were incorporated in a clade with metagenome samples as well as bacteria with 100 percent bootstrap support. More importantly, another bacteria protein was outgrouped with the whole clade. This specific relationship could be seen in the MSH6 phylogenetic tree as well (Figure 5).

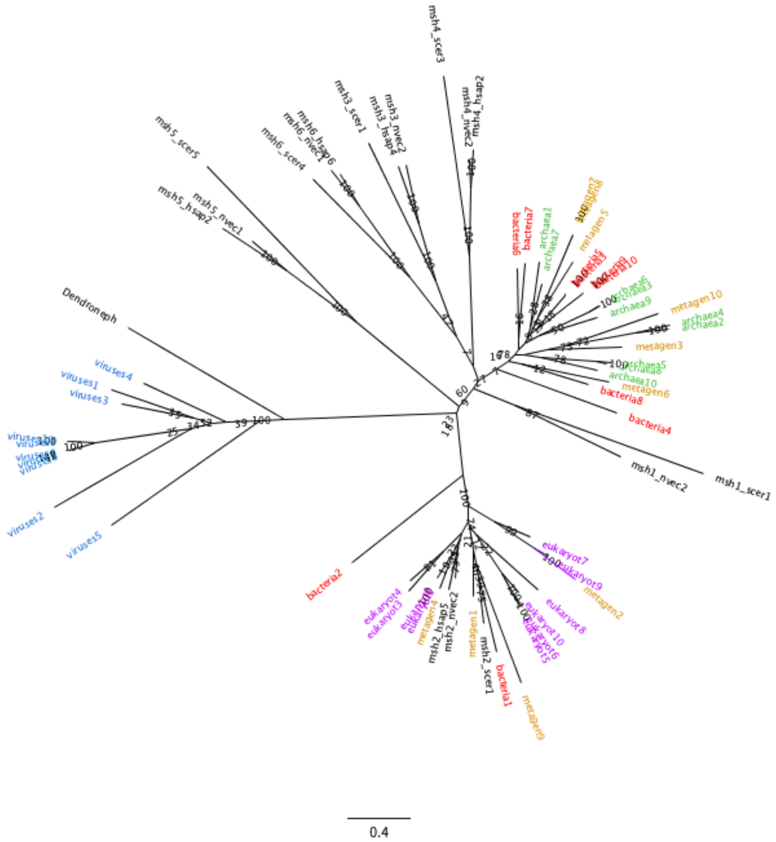


Figure 1: Phylogenetic tree of MSH2 from archaea, bacteria, eukaryotes, metagenomes, and viruses. Phylogeny was built using RAxML-NG with 1000 bootstraps.

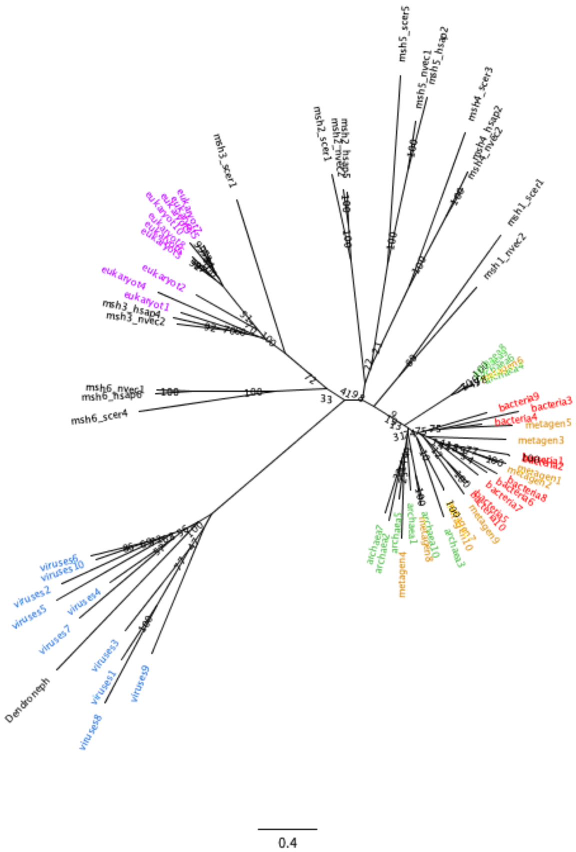


Figure 2: Phylogenetic tree of MSH3 from archaea, bacteria, eukaryotes, metagenomes, and viruses. Phylogeny was built using RAxML-NG with 1000 bootstraps.

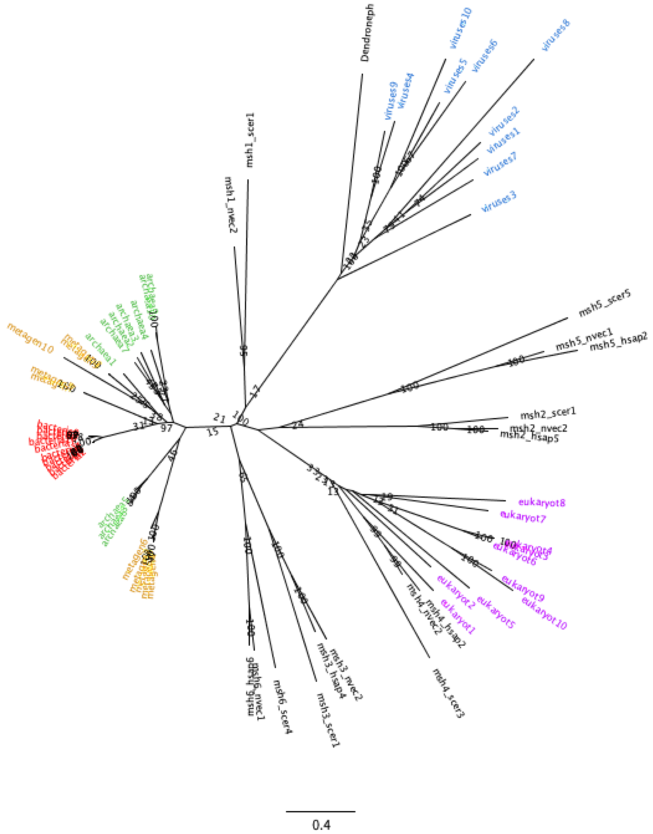


Figure 3: Phylogenetic tree of MSH4 from archaea, bacteria, eukaryotes, metagenomes, and viruses. Phylogeny was built using RAxML-NG with 1000 bootstraps.

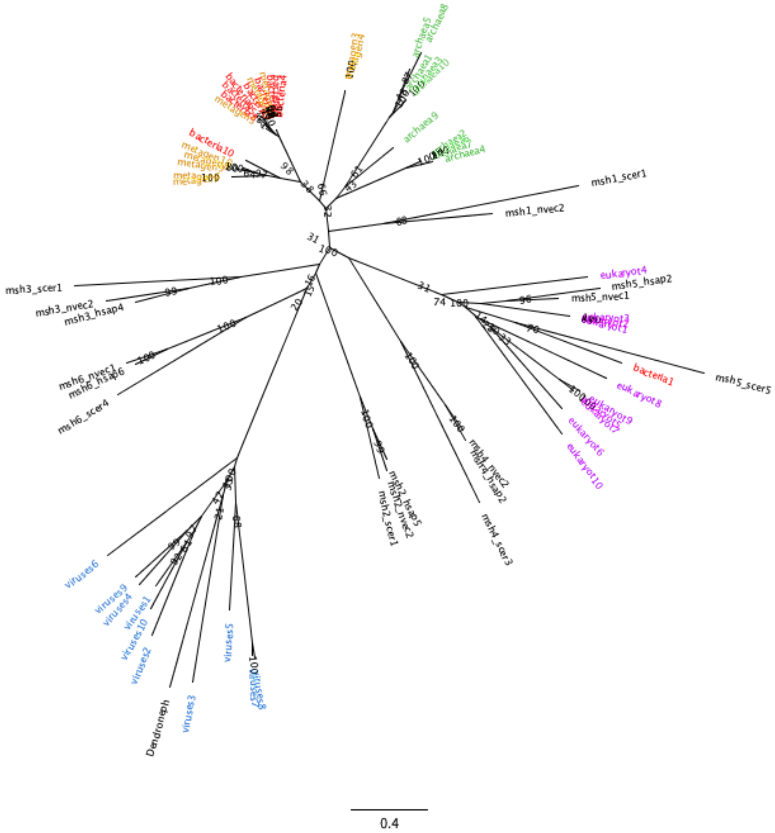


Figure 4: Phylogenetic tree of MSH5 from archaea, bacteria, eukaryotes, metagenomes, and viruses. Phylogeny was built using RAxML-NG with 1000 bootstraps.

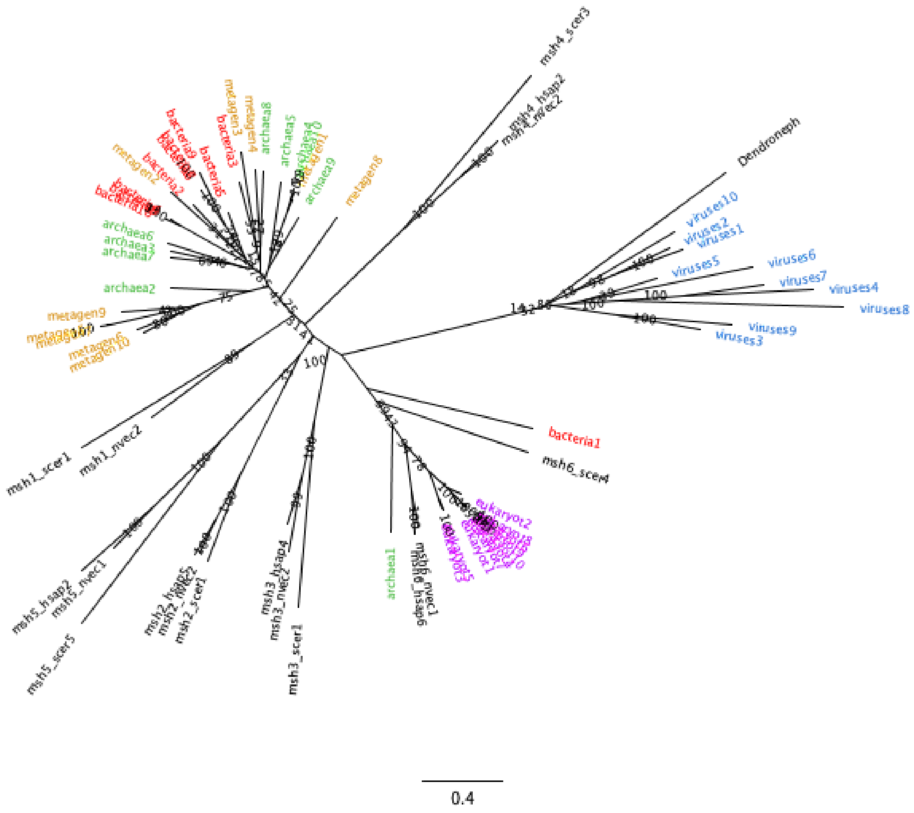


Figure 5: Phylogenetic tree of MSH6 from archaea, bacteria, eukaryotes, metagenomes, and viruses. Phylogeny was built using RAxML-NG with 1000 bootstraps.

Discussion

The origin of mtMutS has been unsolved for many years even though many hypotheses were proposed. All the constructed phylogenies cluster viral sequences and mtMutS of *Dendronephthya gigantea* sponge together indicating that mtMutS does not belong to the MSH family and also it does not have a eukaryotic origin. Both having this relationship in all the trees and indicating 100 percent bootstrap value provide more support for this phenomena. This observation provide evidence for the hypothesis that animal mtMutS has originated due to a horizontal gene transfer (HGT) event from a large DNA virus. This would be the first known case of horizontal gene transfer into an animal mitochondrial genome.

There are several studies that propose opposing theories about eukaryotic MutS family evolution. While some suggest that MSH homologs form a monophyletic clade in the evolutionary process, many others establish the theory of paraphyletic MSH evolution with some relationship to bacteria (Bilewitch and Degnan, 2011; Muthye and Lavrov, 2020). In this study, phylogenetic tree for MSH2 (Figure 1) clearly shows a clade of eukaryotic sequences including an outgroup of bacteria (bacteria 2) with 100 bootstrap support. Bacteria 2 refers to HHG10973.1 MutS family DNA mismatch repair protein [bacterium]. This bacteria sequence is from a hot springs metagenome. This result can be an evidence for horizontal gene transfer event in between bacteria and eukaryotic genomes. Figure 6 also illustrates the same relationship, but with small changes. In here, archaea 1 which is a sequence from phyllosphere metagenome is included in eukaryote cluster and the whole cluster is outgrouped with bacteria 1 (MBE7180883.1 hypothetical protein). This protein is from *Terriglobus roseus* plastic metagenome and the bacterial species is commonly found in agricultural soils. In the figure 5 also bacteria 1 includes in the eukaryotic clade providing more evidence for the relationship between eukaryotic MSH and bacteria. In addition to that specific clade, most of the trees cluster MSH sequences with bacterial sequences and it gives the idea that there can be a strong relationship between bacteria and other groups of organisms.

In conclusion, this study provides more evidence to the hypothesis that the eukaryotic MSH can be evolved due to a horizontal gene transfer event in between bacteria and eukaryotic genome. Furthermore it supports the theory that the origin of eukaryotic mtMuts has occurred due to another HGT event from a large DNA virus.

References

Akaike H. (1973). Information theory and an extension of the maximum likelihood principle In: Petrov BN, Csaki F, editors. Second international symposium on information theory. Budapest (Hungary:): Akademiai Kiado; p. 267–281.

Bilewitch, J. P., Degnan, S. M. (2011). A unique horizontal gene transfer event has provided the octocoral mitochondrial genome with an active mismatch repair gene that has potential for an unusual self-contained function. *BMC Evol Biol* 11, 228. https://doi.org/10.1186/1471-2148-11-228.

Ivica Letunic, Peer Bork, Interactive Tree Of Life (iTOL) v4: recent updates and new developments, *Nucleic Acids Research*, Volume 47, Issue W1, 02 July 2019, Pages W256–W259, https://doi.org/10.1093/nar/gkz239.

Katoh, et al. “MAFFT: a Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform.” *OUP Academic*, Oxford University Press, 15 July 2002, academic.oup.com/nar/article/30/14/3059/2904316.

Kozlov, et al. “RAxML-NG: a Fast, Scalable and User-Friendly Tool for Maximum Likelihood Phylogenetic Inference.” *OUP Academic*, Oxford University Press, 9 May 2019, academic.oup.com/bioinformatics/article/35/21/4453/548.

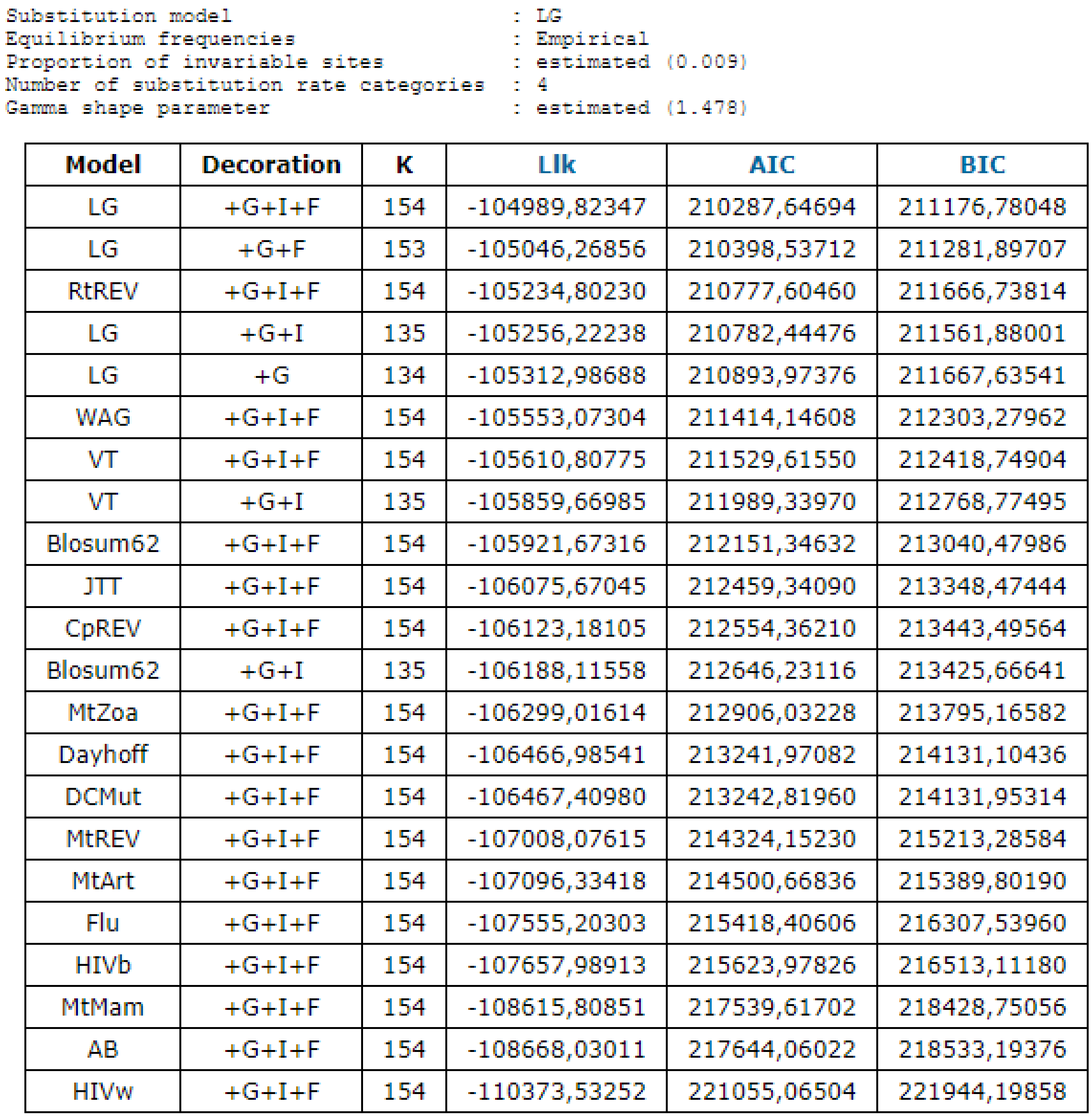
Lefort, V., Longueville, J. E., & Gascuel, O. (2017). SMS: Smart Model Selection in PhyML. *Molecular biology and evolution, 34*(9), 2422–2424. https://doi.org/10.1093/molbev/msx149.

Salvador Capella-Gutierrez, Jose M. Silla-Martinez, Toni Gabaldon. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25: 1972-1973.

Viraj Muthye, Dennis V. Lavrov. (2020). Dynamic evolution of the MutS family in animals: multiple losses of MSH paralogues and gain of a viral MutS homologue in octocorals. *bioRxiv* 2020.12.22.424024; doi: https://doi.org/10.1101/2020.12.22.424024

Supplementary Materials

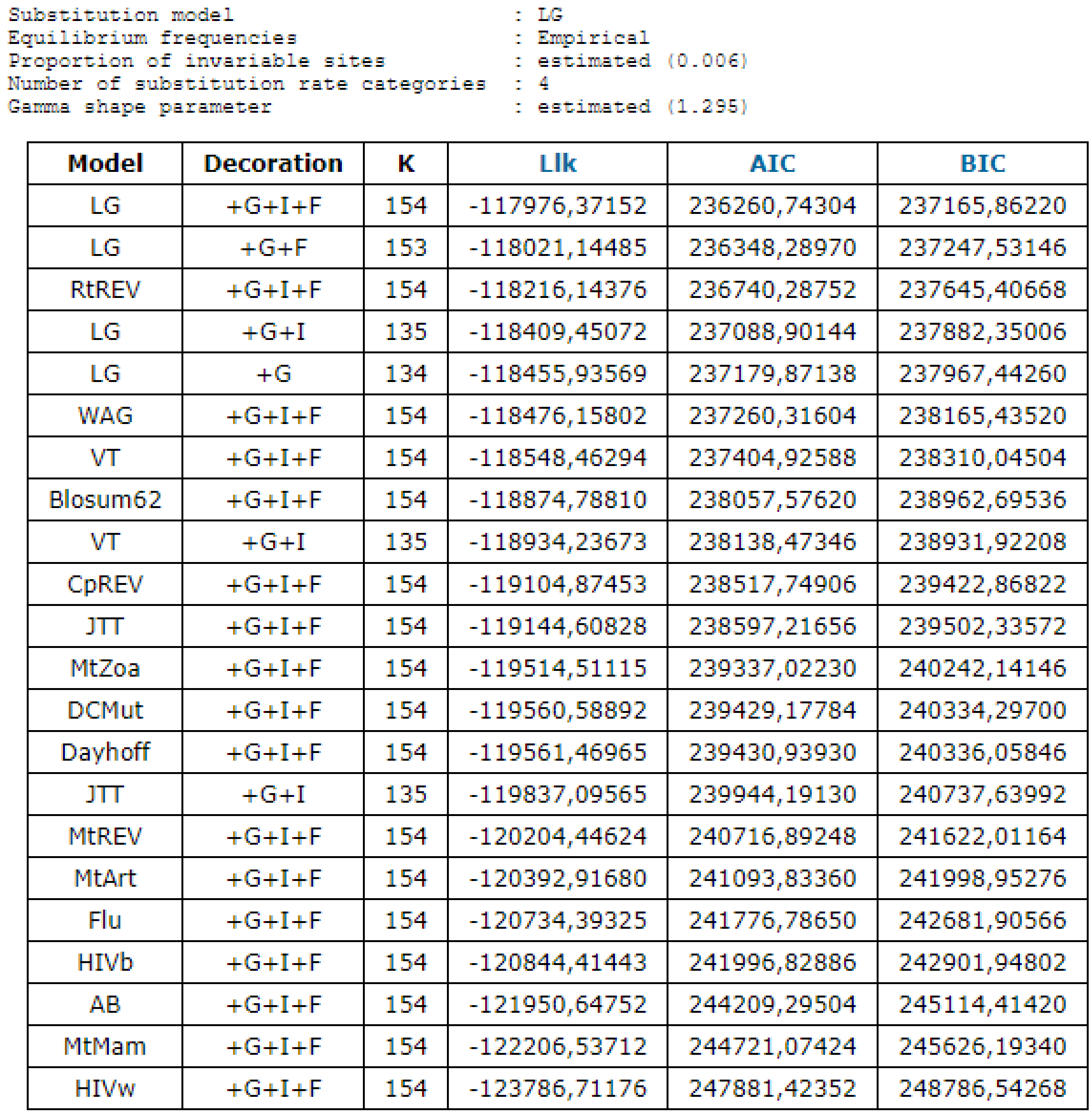
Supplementary Table1: Complete results for MSH2 model selection by SMS smart model selection tool.



```
Substitution model                    : LG
Equilibrium frequencies               : Empirical
Proportion of invariable sites        : estimated (0.009)
Number of substitution rate categories  : 4
Gamma shape parameter                 : estimated (1.478)
```

Model	Decoration	K	Llk	AIC	BIC
LG	+G+I+F	154	-104989,82347	210287,64694	211176,78048
LG	+G+F	153	-105046,26856	210398,53712	211281,89707
RtREV	+G+I+F	154	-105234,80230	210777,60460	211666,73814
LG	+G+I	135	-105256,22238	210782,44476	211561,88001
LG	+G	134	-105312,98688	210893,97376	211667,63541
WAG	+G+I+F	154	-105553,07304	211414,14608	212303,27962
VT	+G+I+F	154	-105610,80775	211529,61550	212418,74904
VT	+G+I	135	-105859,66985	211989,33970	212768,77495
Blosum62	+G+I+F	154	-105921,67316	212151,34632	213040,47986
JTT	+G+I+F	154	-106075,67045	212459,34090	213348,47444
CpREV	+G+I+F	154	-106123,18105	212554,36210	213443,49564
Blosum62	+G+I	135	-106188,11558	212646,23116	213425,66641
MtZoa	+G+I+F	154	-106299,01614	212906,03228	213795,16582
Dayhoff	+G+I+F	154	-106466,98541	213241,97082	214131,10436
DCMut	+G+I+F	154	-106467,40980	213242,81960	214131,95314
MtREV	+G+I+F	154	-107008,07615	214324,15230	215213,28584
MtArt	+G+I+F	154	-107096,33418	214500,66836	215389,80190
Flu	+G+I+F	154	-107555,20303	215418,40606	216307,53960
HIVb	+G+I+F	154	-107657,98913	215623,97826	216513,11180
MtMam	+G+I+F	154	-108615,80851	217539,61702	218428,75056
AB	+G+I+F	154	-108668,03011	217644,06022	218533,19376
HIVw	+G+I+F	154	-110373,53252	221055,06504	221944,19858

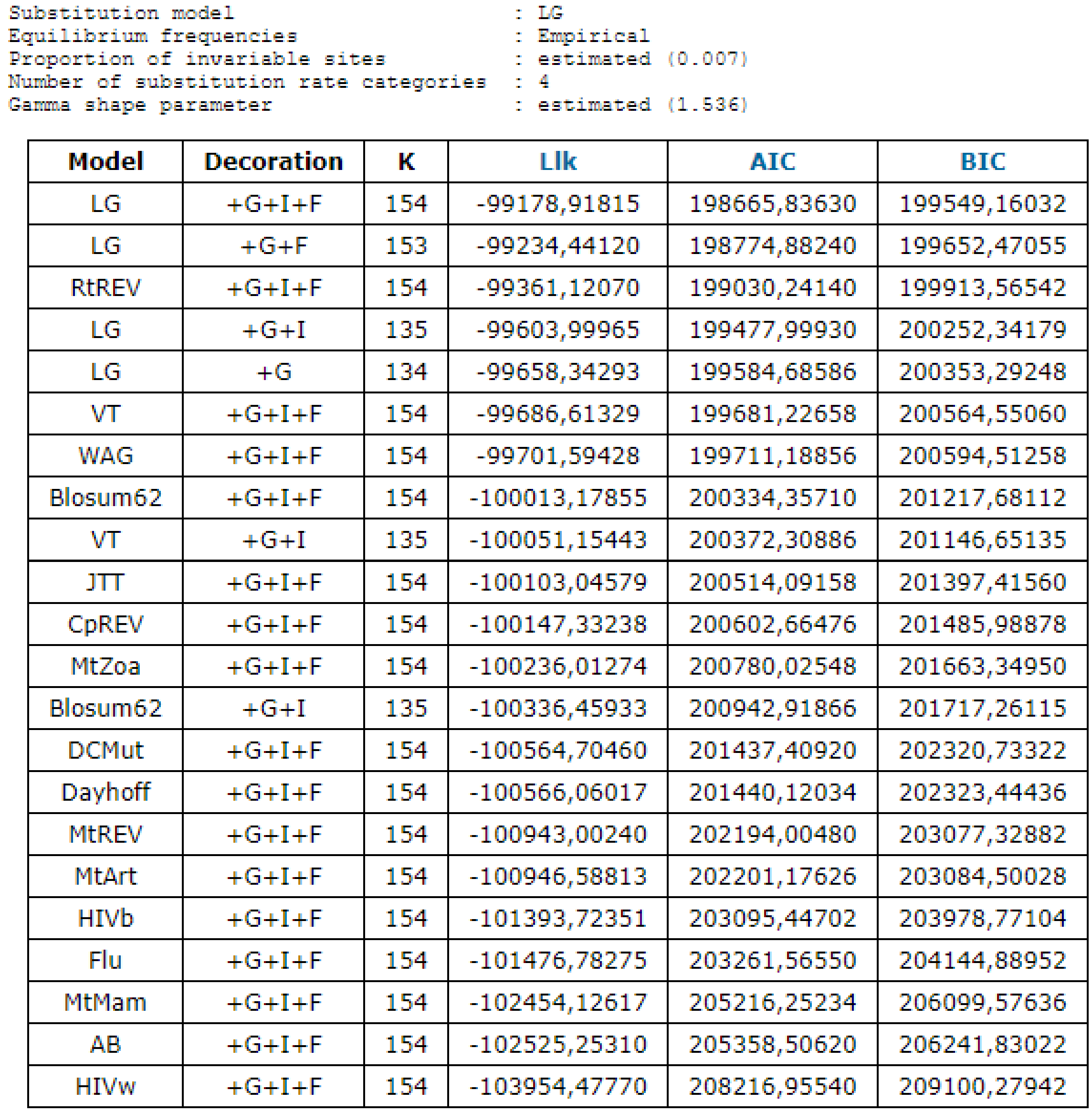
Supplementary Table2: Complete results for MSH3 model selection by SMS smart model selection tool.



```
Substitution model                      : LG
Equilibrium frequencies                 : Empirical
Proportion of invariable sites          : estimated (0.006)
Number of substitution rate categories  : 4
Gamma shape parameter                   : estimated (1.295)
```

Model	Decoration	K	Llk	AIC	BIC
LG	+G+I+F	154	-117976,37152	236260,74304	237165,86220
LG	+G+F	153	-118021,14485	236348,28970	237247,53146
RtREV	+G+I+F	154	-118216,14376	236740,28752	237645,40668
LG	+G+I	135	-118409,45072	237088,90144	237882,35006
LG	+G	134	-118455,93569	237179,87138	237967,44260
WAG	+G+I+F	154	-118476,15802	237260,31604	238165,43520
VT	+G+I+F	154	-118548,46294	237404,92588	238310,04504
Blosum62	+G+I+F	154	-118874,78810	238057,57620	238962,69536
VT	+G+I	135	-118934,23673	238138,47346	238931,92208
CpREV	+G+I+F	154	-119104,87453	238517,74906	239422,86822
JTT	+G+I+F	154	-119144,60828	238597,21656	239502,33572
MtZoa	+G+I+F	154	-119514,51115	239337,02230	240242,14146
DCMut	+G+I+F	154	-119560,58892	239429,17784	240334,29700
Dayhoff	+G+I+F	154	-119561,46965	239430,93930	240336,05846
JTT	+G+I	135	-119837,09565	239944,19130	240737,63992
MtREV	+G+I+F	154	-120204,44624	240716,89248	241622,01164
MtArt	+G+I+F	154	-120392,91680	241093,83360	241998,95276
Flu	+G+I+F	154	-120734,39325	241776,78650	242681,90566
HIVb	+G+I+F	154	-120844,41443	241996,82886	242901,94802
AB	+G+I+F	154	-121950,64752	244209,29504	245114,41420
MtMam	+G+I+F	154	-122206,53712	244721,07424	245626,19340
HIVw	+G+I+F	154	-123786,71176	247881,42352	248786,54268

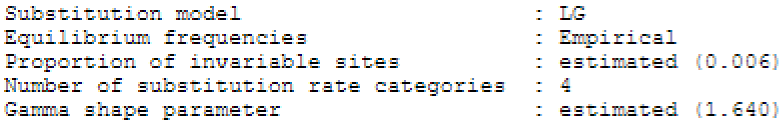
Supplementary Table3: Complete results for MSH4 model selection by SMS smart model selection tool.



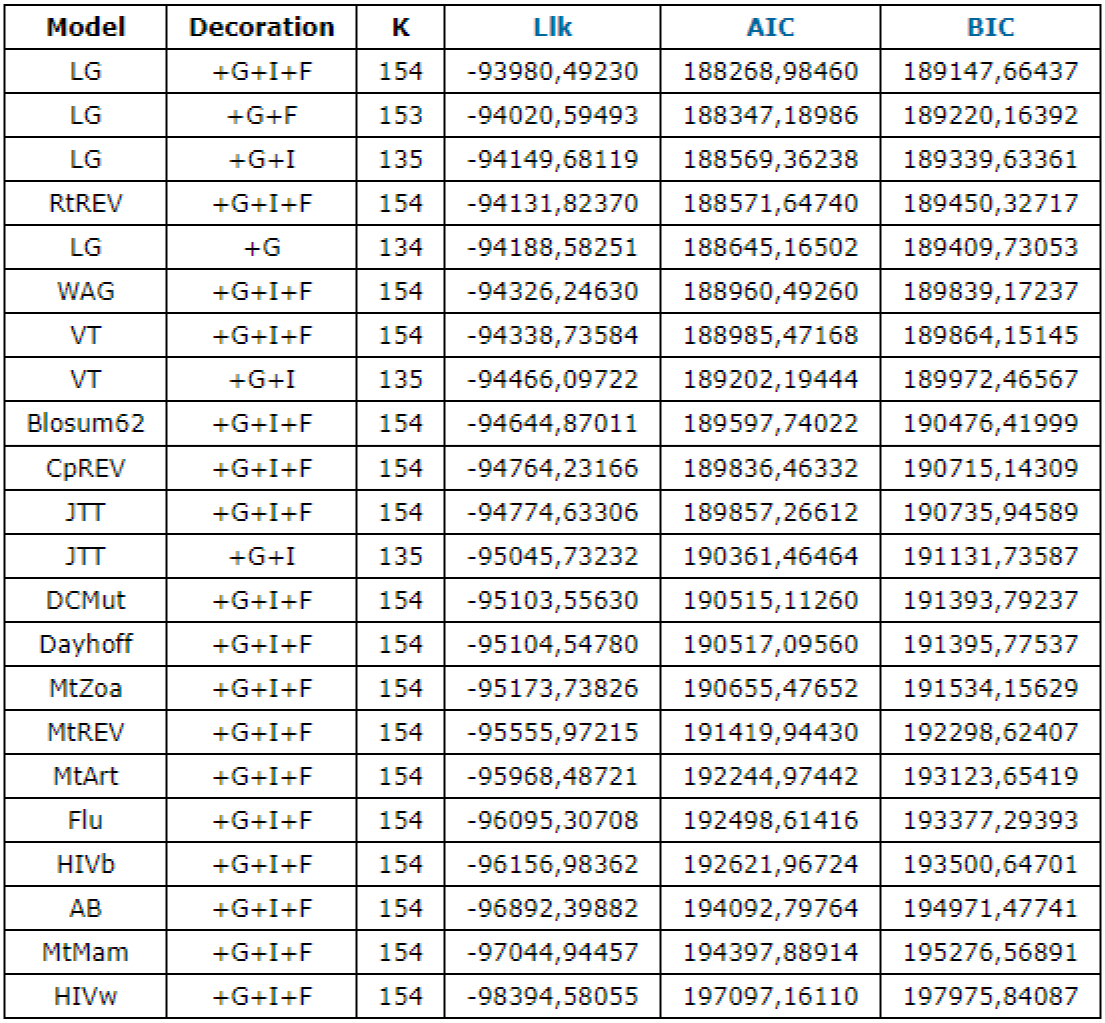
```
Substitution model                        : LG
Equilibrium frequencies                   : Empirical
Proportion of invariable sites            : estimated (0.007)
Number of substitution rate categories    : 4
Gamma shape parameter                     : estimated (1.536)
```

Model	Decoration	K	Llk	AIC	BIC
LG	+G+I+F	154	-99178,91815	198665,83630	199549,16032
LG	+G+F	153	-99234,44120	198774,88240	199652,47055
RtREV	+G+I+F	154	-99361,12070	199030,24140	199913,56542
LG	+G+I	135	-99603,99965	199477,99930	200252,34179
LG	+G	134	-99658,34293	199584,68586	200353,29248
VT	+G+I+F	154	-99686,61329	199681,22658	200564,55060
WAG	+G+I+F	154	-99701,59428	199711,18856	200594,51258
Blosum62	+G+I+F	154	-100013,17855	200334,35710	201217,68112
VT	+G+I	135	-100051,15443	200372,30886	201146,65135
JTT	+G+I+F	154	-100103,04579	200514,09158	201397,41560
CpREV	+G+I+F	154	-100147,33238	200602,66476	201485,98878
MtZoa	+G+I+F	154	-100236,01274	200780,02548	201663,34950
Blosum62	+G+I	135	-100336,45933	200942,91866	201717,26115
DCMut	+G+I+F	154	-100564,70460	201437,40920	202320,73322
Dayhoff	+G+I+F	154	-100566,06017	201440,12034	202323,44436
MtREV	+G+I+F	154	-100943,00240	202194,00480	203077,32882
MtArt	+G+I+F	154	-100946,58813	202201,17626	203084,50028
HIVb	+G+I+F	154	-101393,72351	203095,44702	203978,77104
Flu	+G+I+F	154	-101476,78275	203261,56550	204144,88952
MtMam	+G+I+F	154	-102454,12617	205216,25234	206099,57636
AB	+G+I+F	154	-102525,25310	205358,50620	206241,83022
HIVw	+G+I+F	154	-103954,47770	208216,95540	209100,27942

Supplementary Table4: Complete results for MSH5 model selection by SMS smart model selection tool.

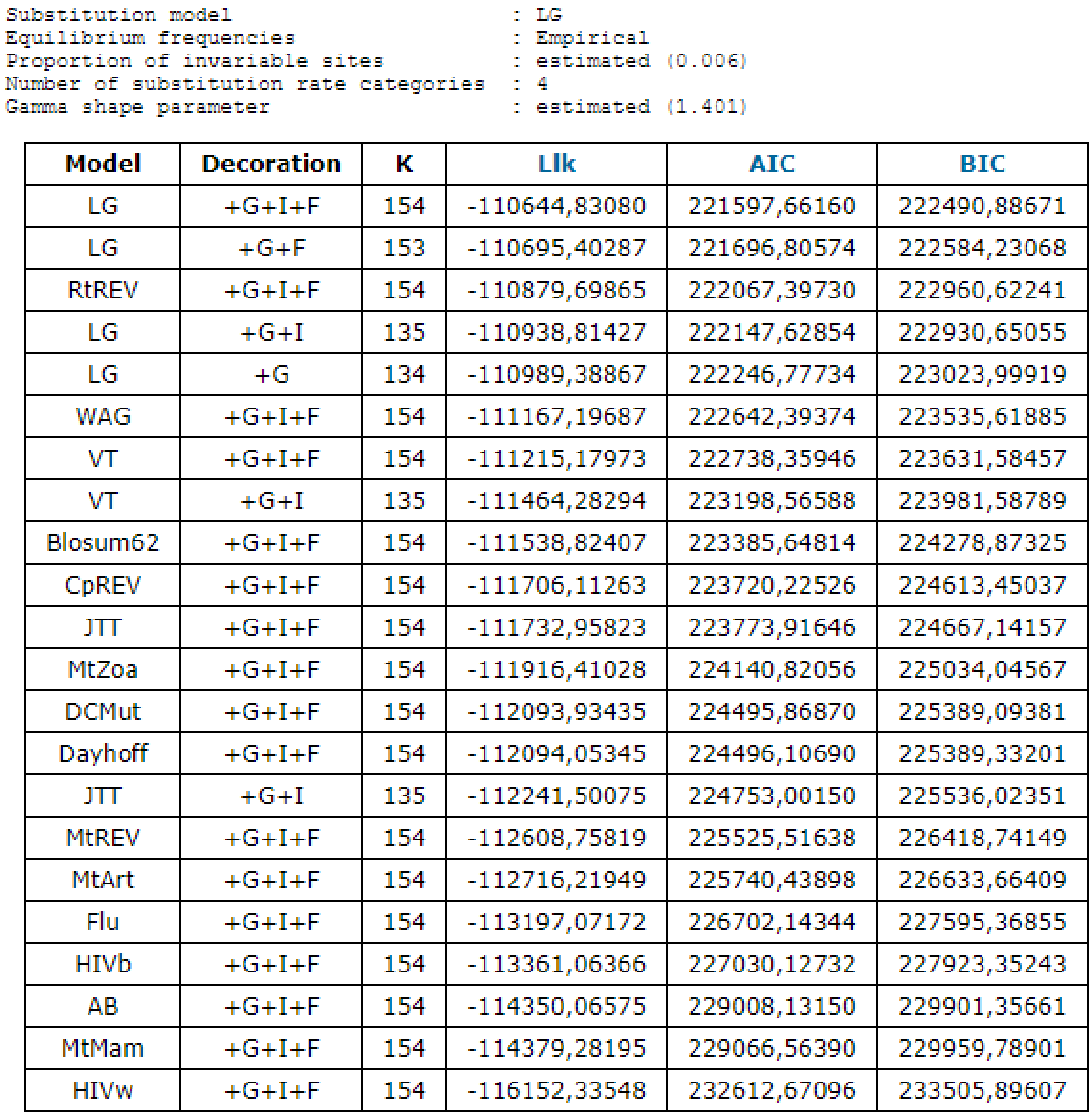


```
Substitution model                       : LG
Equilibrium frequencies                  : Empirical
Proportion of invariable sites           : estimated (0.006)
Number of substitution rate categories   : 4
Gamma shape parameter                    : estimated (1.640)
```



Model	Decoration	K	Llk	AIC	BIC
LG	+G+I+F	154	-93980,49230	188268,98460	189147,66437
LG	+G+F	153	-94020,59493	188347,18986	189220,16392
LG	+G+I	135	-94149,68119	188569,36238	189339,63361
RtREV	+G+I+F	154	-94131,82370	188571,64740	189450,32717
LG	+G	134	-94188,58251	188645,16502	189409,73053
WAG	+G+I+F	154	-94326,24630	188960,49260	189839,17237
VT	+G+I+F	154	-94338,73584	188985,47168	189864,15145
VT	+G+I	135	-94466,09722	189202,19444	189972,46567
Blosum62	+G+I+F	154	-94644,87011	189597,74022	190476,41999
CpREV	+G+I+F	154	-94764,23166	189836,46332	190715,14309
JTT	+G+I+F	154	-94774,63306	189857,26612	190735,94589
JTT	+G+I	135	-95045,73232	190361,46464	191131,73587
DCMut	+G+I+F	154	-95103,55630	190515,11260	191393,79237
Dayhoff	+G+I+F	154	-95104,54780	190517,09560	191395,77537
MtZoa	+G+I+F	154	-95173,73826	190655,47652	191534,15629
MtREV	+G+I+F	154	-95555,97215	191419,94430	192298,62407
MtArt	+G+I+F	154	-95968,48721	192244,97442	193123,65419
Flu	+G+I+F	154	-96095,30708	192498,61416	193377,29393
HIVb	+G+I+F	154	-96156,98362	192621,96724	193500,64701
AB	+G+I+F	154	-96892,39882	194092,79764	194971,47741
MtMam	+G+I+F	154	-97044,94457	194397,88914	195276,56891
HIVw	+G+I+F	154	-98394,58055	197097,16110	197975,84087

Supplementary Table5: Complete results for MSH6 model selection by SMS smart model selection tool.



```
Substitution model                       : LG
Equilibrium frequencies                  : Empirical
Proportion of invariable sites           : estimated (0.006)
Number of substitution rate categories   : 4
Gamma shape parameter                    : estimated (1.401)
```

Model	Decoration	K	Llk	AIC	BIC
LG	+G+I+F	154	-110644,83080	221597,66160	222490,88671
LG	+G+F	153	-110695,40287	221696,80574	222584,23068
RtREV	+G+I+F	154	-110879,69865	222067,39730	222960,62241
LG	+G+I	135	-110938,81427	222147,62854	222930,65055
LG	+G	134	-110989,38867	222246,77734	223023,99919
WAG	+G+I+F	154	-111167,19687	222642,39374	223535,61885
VT	+G+I+F	154	-111215,17973	222738,35946	223631,58457
VT	+G+I	135	-111464,28294	223198,56588	223981,58789
Blosum62	+G+I+F	154	-111538,82407	223385,64814	224278,87325
CpREV	+G+I+F	154	-111706,11263	223720,22526	224613,45037
JTT	+G+I+F	154	-111732,95823	223773,91646	224667,14157
MtZoa	+G+I+F	154	-111916,41028	224140,82056	225034,04567
DCMut	+G+I+F	154	-112093,93435	224495,86870	225389,09381
Dayhoff	+G+I+F	154	-112094,05345	224496,10690	225389,33201
JTT	+G+I	135	-112241,50075	224753,00150	225536,02351
MtREV	+G+I+F	154	-112608,75819	225525,51638	226418,74149
MtArt	+G+I+F	154	-112716,21949	225740,43898	226633,66409
Flu	+G+I+F	154	-113197,07172	226702,14344	227595,36855
HIVb	+G+I+F	154	-113361,06366	227030,12732	227923,35243
AB	+G+I+F	154	-114350,06575	229008,13150	229901,35661
MtMam	+G+I+F	154	-114379,28195	229066,56390	229959,78901
HIVw	+G+I+F	154	-116152,33548	232612,67096	233505,89607