**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. Since mismatch repair is critical for maintaining genome stability, from this project, evolution of the MutS protein family in animals will be investigated.

**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. 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.

Table1: 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 |

**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.

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.

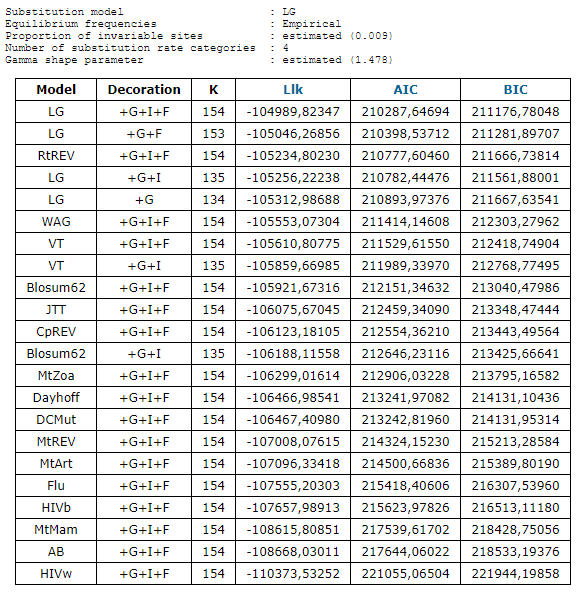
Kozlov, et al. “RAxML-NG: a Fast, Scalable and User-Friendly Tool for Maximum Likelihood Phylogenetic Inference.” *OUP Academic*, Oxford University Press, 9 May 2019, academic.oup.com/bioinformatics/article/35/21/4453/548.

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.

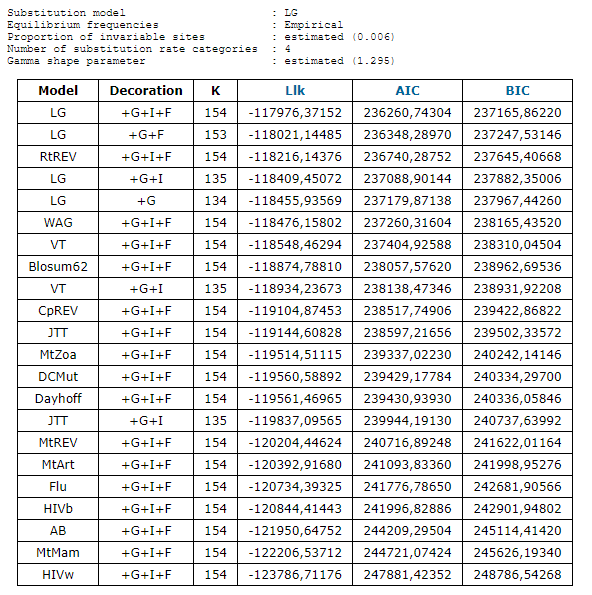
Salvador Capella-Gutierrez, Jose M. Silla-Martinez, Toni Gabaldon. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, 25: 1972-1973.

**Supplementary Materials**

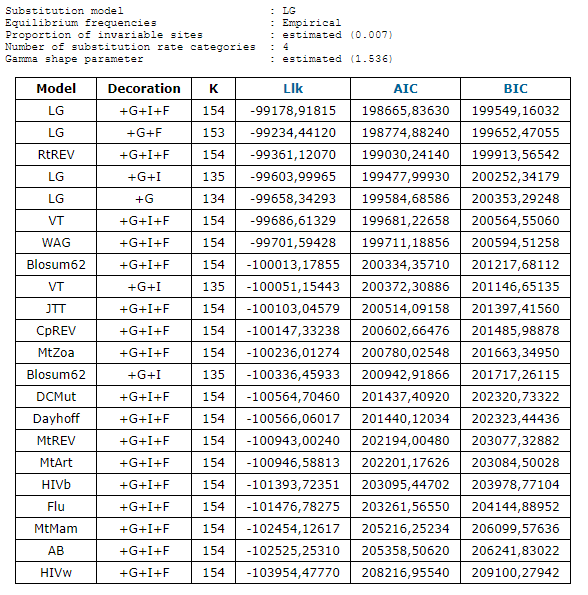
Supplementary Table1: Complete results for MSH2 model selection by SMS smart model selection tool.



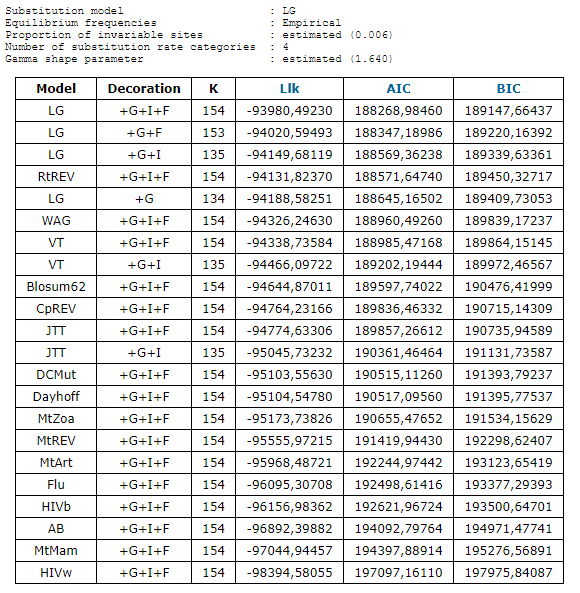
Supplementary Table2: Complete results for MSH3 model selection by SMS smart model selection tool.



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



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



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

