RowanUniversity
COLLEGE OF SCIENCE & MATHEMATICS
SCHOOL OF HEALTH PROFESSIONS
College Office of Undergraduate Research Initiatives

New Jersey Space Grant Consortium

Alignment Quality Can Have An Effect on Phylogenetic Model Selection

Molly Miraglia[1], Stephanie J. Spielman[2]

Department of Molecular and Cellular Biosciences, Rowan University, Glassboro NJ

Abstract

An alignment is a character matrix containing DNA or amino acid sequences from several species. These sequences, known as orthologs, are genes in different species that evolved from a common ancestor. Alignments contain gaps which represent insertions and deletions in sequence evolution. Creating an alignment is the first step for comparing orthologs and building phylogenies. Producing phylogenies requires specification of a suitable model of sequence evolution. To determine these models, we often use model selection based using theoretic information criteria, such as Akaike Information Criterion (AIC). However, generating alignments is prone to error, and alignment quality is known to affect the quality of phylogenies. Here, we ask whether the alignment also affects finding the model that best fits the data. We generate a set of perturbed alignments for 200 protein and nucleotide datasets each from the *Selectome* database and analyze whether the best-fitting model is consistent for all alignment versions. We find that the alignment does have the potential to affect model selection, such that different models are identified as the best-fitting model for a given alignment version. Future work will examine how certain features of the data may further affect model selection.

Protein Model Selection

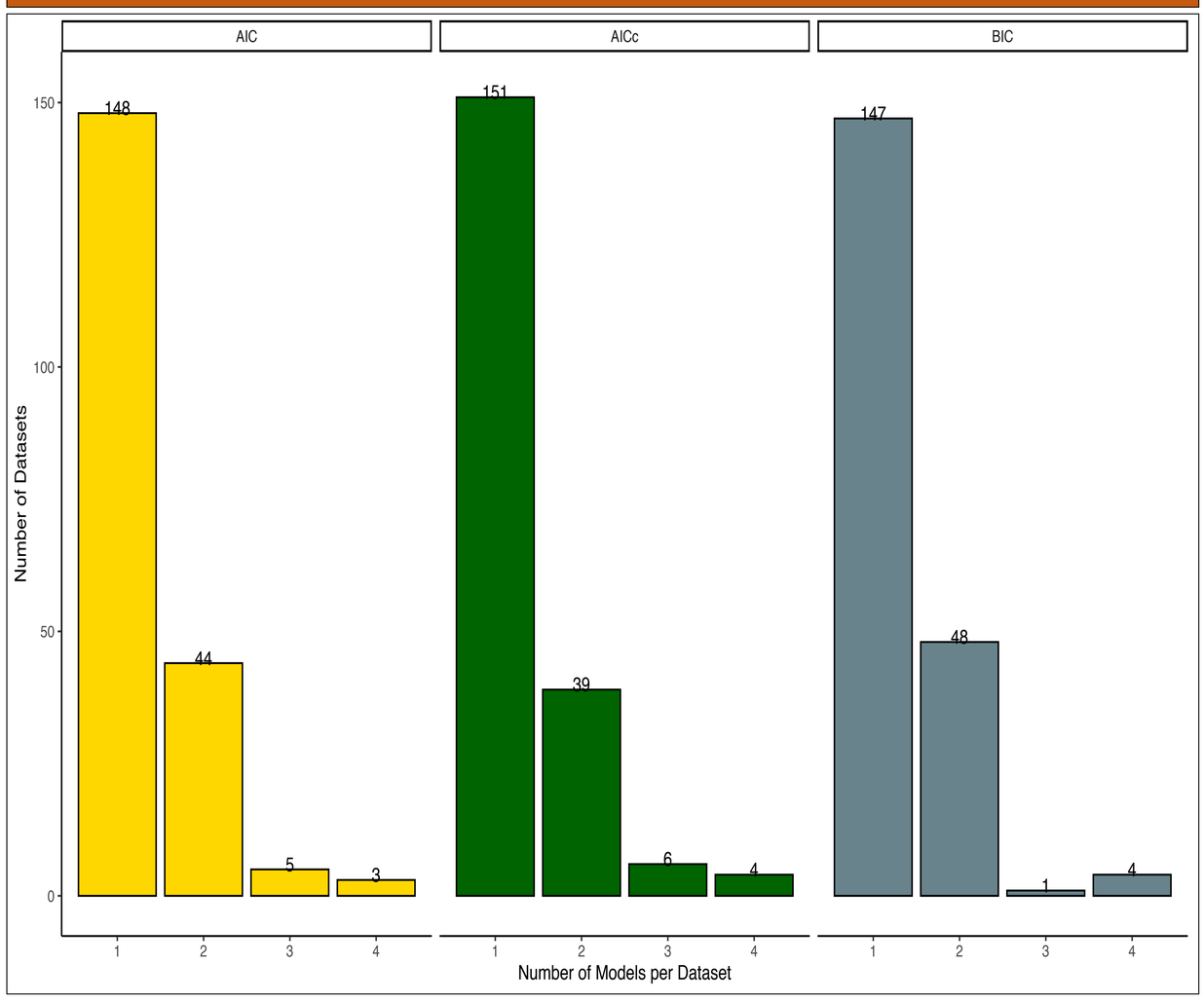


Figure 1. 200 Amino Acid Datasets with 50 alignments per dataset

Nucleotide Model Selection

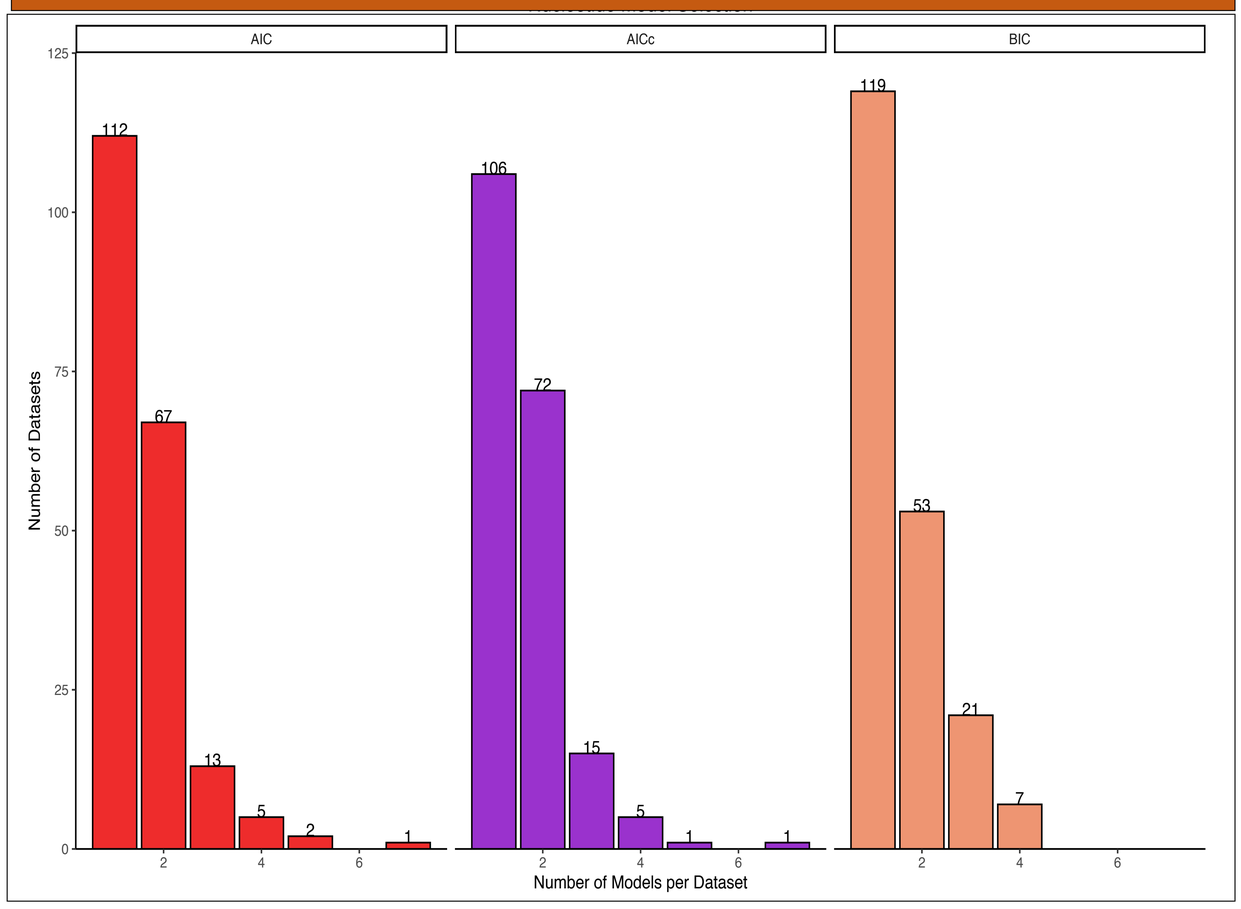
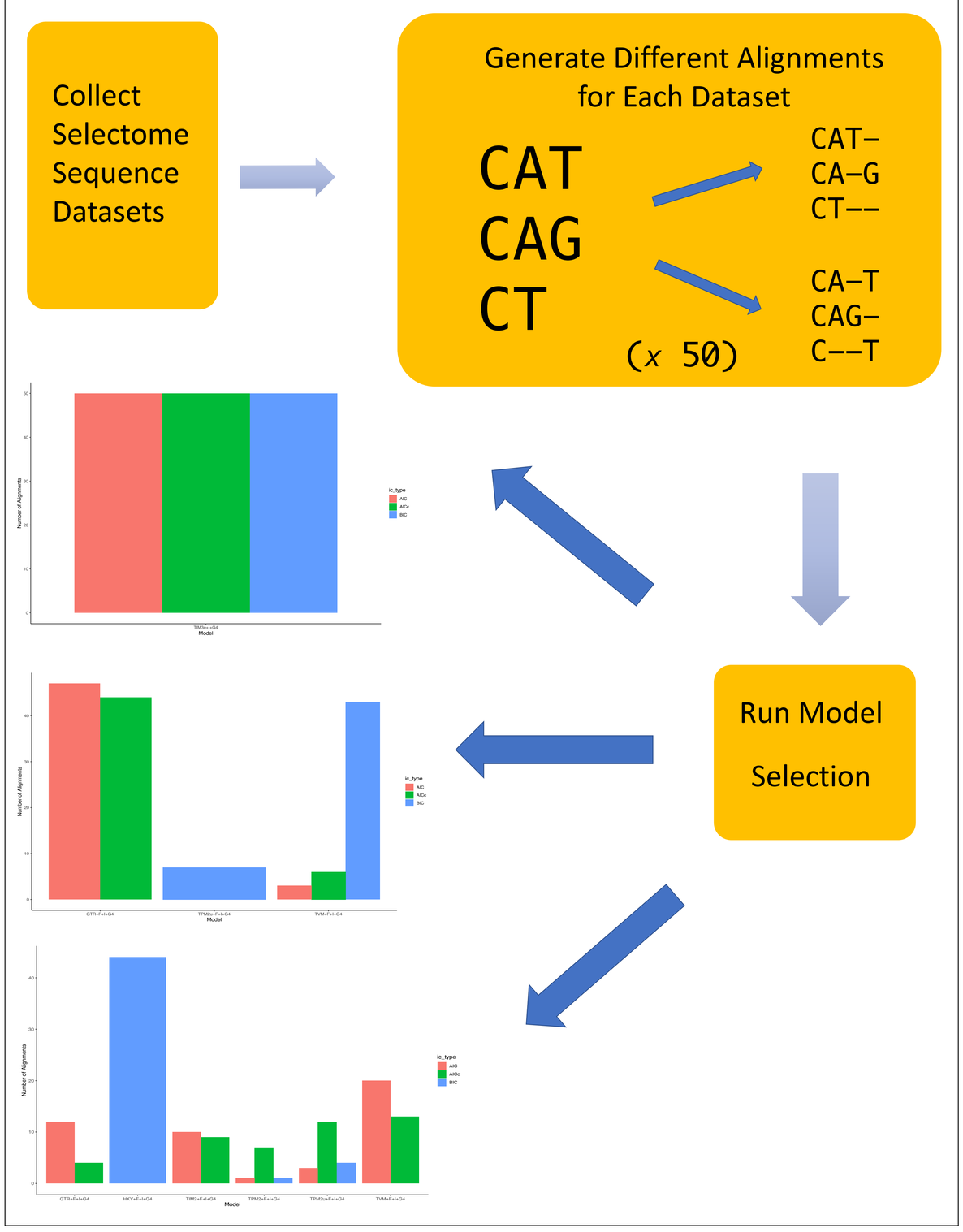
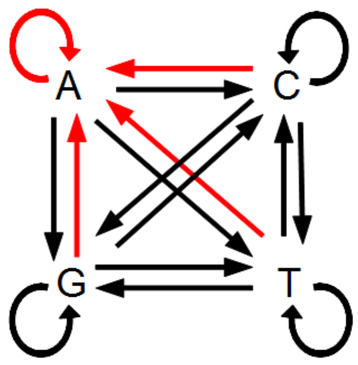


Figure 2. 200 Nucleotide Datasets with 50 alignments per dataset

Methods



What is a Model?



What is Information Criterion?:

- AIC, AICc, BIC likelihood
- Lowest score is best fitting
- Different Information Criterion based on different statistical formulas

Conclusions

- Multiple models can be determined as best fitting
- Highest number of multiple best-fitting models found is 7
- Different Information Criterion favors different models

References

1. Proux E, Studer RA, Moretti S, Robinson-Rechavi M. Selectome: a database of positive selection. *Nucleic Acids Res*. 2009;37(Database issue):D404–D407. doi:10.1093/nar/gkn768 2. L.-T. Nguyen, H.A. Schmidt, A. von Haeseler, B.Q. Minh (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum likelihood phylogenies.. *Mol. Biol. Evol.*, 32:268-274. https://doi.org/10.1093/molbev/msu300 3. Sela I, Ashkenazy H, Katoh K, Pupko T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. *Nucleic Acids Res*. 2015;43(W1):W7–W14. doi:10.1093/nar/gkv318 4. Spielman SJ, Dawson ET, Wilke CO. Limited utility of residue masking for positive-selection inference. *Mol Biol Evol*. 2014;31(9):2496–2500. doi:10.1093/molbev/msu183