A Computational Investigation of the Mouse Genome

Heather Noonan, Dr. Kevin Liu
Department of Computer Science and Engineering

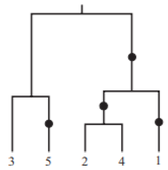
Introduction

Table 3.2 A data set with five sequences and four segregating sites with relative positions[a]

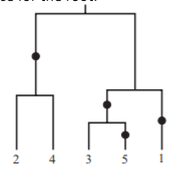
	0.012	0.41	0.47	0.97
1	0	0	0	0
2	0	0	1	1
3	1	0	1	0
4	0	0	1	1
5	1	1	1	0

[a] The sites partition the sequences into the following sets: {1, 2, 4}, {3, 5}; {1, 2, 3, 4}, {5}; {1}, {2, 3, 4, 5}; and {1, 3, 5}, {2, 4}.

Representation as Ancestral Sequence

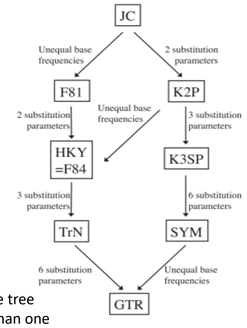


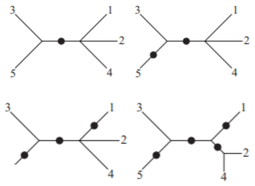
An ancestral point is chosen to root the tree from. Most of the time there is more than one choice for the root.



Tree Making Process

Models of Evolution





Unrooted gene trees are constructed site by site. The filled circles denote mutations.

The branches without mutations are removed and replaced with unfilled circles.

All diagrams in this section were obtained from pages 73-74 of *Sequence variation, genealogies and evolution* [5]

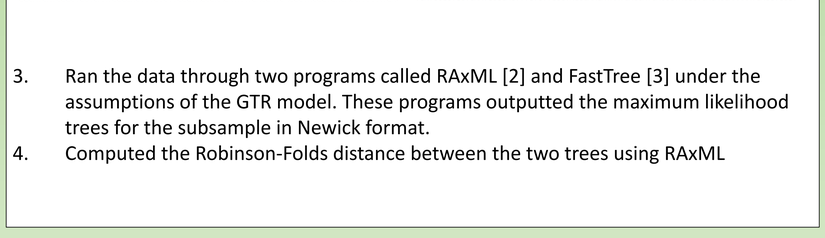
Experiment

In this project we found the best phylogenetic tree for a sample of various *Mus musculus* subspecies and *Mus spretus* mice. Two trees were computed using two separate algorithms, FastTree and RAxML, both under the assumption of the Generalized Time Reversible (GTR) model of evolution.

Computational Methods

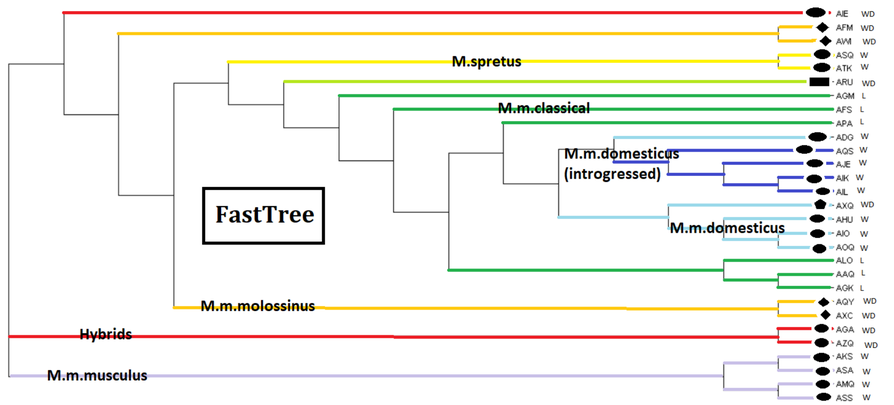
1. Using the published data set from the Didion et. al 2012 study, chose a random sample comprised of the various *Mus musculus* subspecies, *Mus spretus* individuals, and laboratory hybrids [1].
2. Wrote a program in C++ that concatenated the genetic sequences from each chromosome for each chosen individual and put them in phylip and fasta formatted documents.

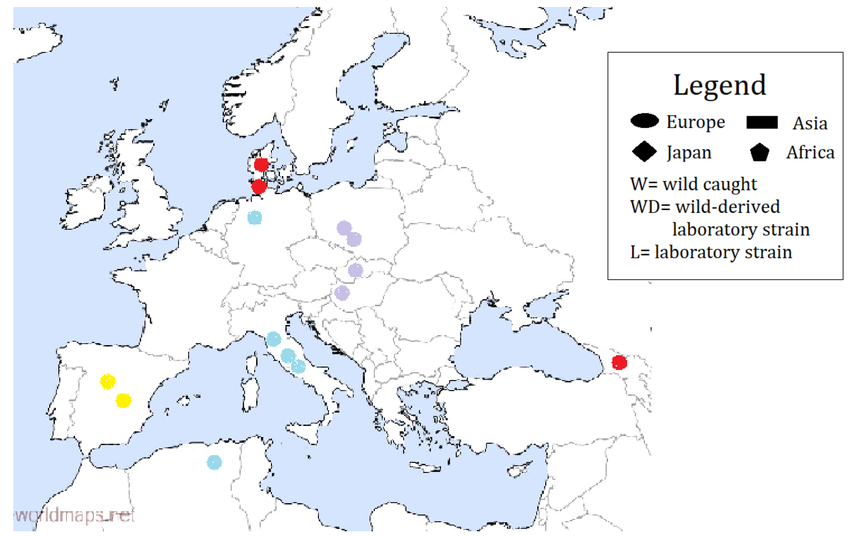
PHYLIP FASTA



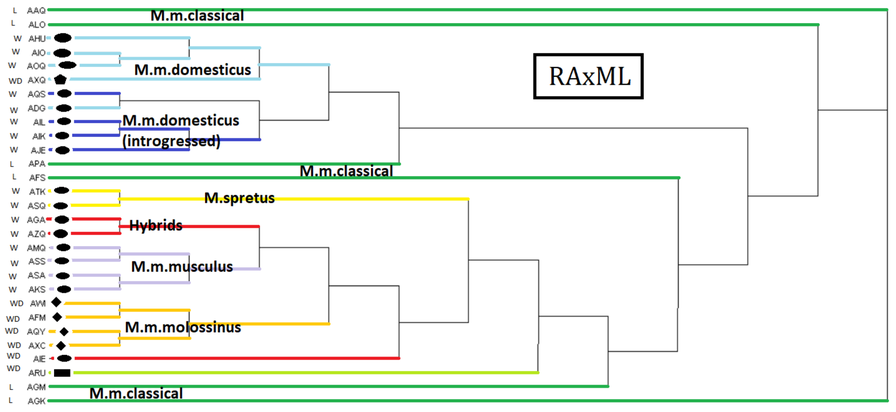
3. Ran the data through two programs called RAxML [2] and FastTree [3] under the assumptions of the GTR model. These programs outputted the maximum likelihood trees for the subsample in Newick format.
4. Computed the Robinson-Folds distance between the two trees using RAxML

Results





Map image obtained from http://americafuntrip.com



Discussion

The two trees constructed by RAxML and FastTree illustrate the ability of two algorithms based on the same model to construct two different phylogenies for the same sample of sequences. Despite obvious differences in the two trees, they still exhibit multiple similarities. Upon, closer inspection, much of the dissimilarity appears to result from a different choice for a root.

Some of the more interesting disparities include the placement of the subspecies, *Mus musculus classical.* These mice are the standard strain used in laboratory experiments and have been inbred in isolation for generations, so much so that their classification as an outgroup on the RAxML tree is not surprising. The phylogeny from FastTree places them within larger subtree, but still splits the subspecies, which doesn't really occur with the other subspecies in the sample. Another notable difference between the two comes when looking at the placement of the *M.m.musculus* subspecies. On one tree they are placed within a subtree, whereas on the other, they are considered an outgroup. However, the subtree they are a part of appear to be comprised of the outgroups of the other tree, again suggesting a different choice of roots between the two algorithms.

One characteristic that both trees had in common that was particularly interesting involved the *M. spretus* species. This species is not actually a *Mus musculus* subspecies so one would expect to see them identified as an outgroup in both phylogenies, however this was not the case when it came to either algorithm. This could be in part due to introgression, although there is not much support for this as the mice who were identifies to carry an introgressed *M.spretus* allele, were not deemed that closely related to the *M.spretus* mice.

It's important to note that variations between the two phylogenetic trees are not limited to the visuals. Both programs were ran under the same conditions with regard to memory and CPU allocation, yet FastTree ran significantly faster than RAxML.

Future Research

Moving forward, there are a couple direction/questions we would like to investigate:

- Analysis of the full sample of *Mus musculus* subspecies
- Investigation into the optimization of maximum likelihood inferencing
- Exploration into the genetic effects of inbreeding and hybridization in terms of the event of speciation
- Closer look at introgression in a larger data sample

References

[1] J. P. Didion, H. Yang, K. Sheppard, C.-P. Fu, L. Mcmillan, F. D. Villena, and G. A. Churchill, "Discovery of novel variants in genotyping arrays improves genotype retention and reduces ascertainment bias," *BMC Genomics*, vol. 13, no. 1, p. 34, Jan. 2012
[2] Alexandros Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, *Bioinformatics*, Volume 30, Issue 9, 1 May 2014, Pages 1312–1313, https://doi.org/10.1093/bioinformatics/btu033
[3] Price, M.N., Dehal, P.S., and Arkin, A.P. (2009) FastTree: Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix. Molecular Biology and Evolution 26:1641-1650, doi:10.1093/molbev/msp077
[4] Price, M.N., Dehal, P.S., and Arkin, A.P. (2010) FastTree 2 -- Approximately Maximum-Likelihood Trees for Large Alignments. PLoS ONE, 5(3):e9490. doi:10.1371/journal.pone.0009490.
[5] J. Hein, M. Schierup, and C. Wiuf, *Sequence variation, genealogies and evolution*. Oxford: Oxford University Press, 2004.
[6] T. Warnow, *Computational phylogenetics: an introduction to designing methods for phylogeny estimation*. Cambridge: Cambridge University Press, 2018.