Simulating Nucleotide Evolution: Summary Report

# Abstract

This project attempted to evaluate several common programs for genetic simulation. I did this by simulating the evolution of a nucleotide sequence along the phylogenetic tree for ten species of bat, using four different programs: DAWG, INDELible, EVOLVER, and one of my own design.

I then analysed the simulated sequences of these four programs using Shannon entropy as my main metric, so as to gauge the complexity of the simulated sequences, and then compared the results of this analysis to the real sequences of each species.

I did not complete the project, but have suggested several areas which could be extended, as discussed in Section 4.

# Introduction

## Models of Evolution

### Substitution

There are two main elements of any theory of nucleotide evolution – substitution and indels. A substitution event is an event where a single base in the sequence is substituted for another base, through various chemical processes, primarily governed by the type of base in question (purine-pyrimidine).

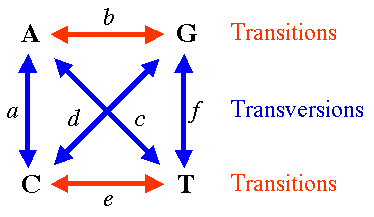


Figure 1: A depiction of the K2P model of evolution.

There are several theories which attempt to explain nucleotide substitution:

* Jukes and Cantor 1969 (**JC69**) – the most basic model. This assumes all transitions between nucleotide bases have equal probability.
* Kimura 2 Parameter 1980 (**K2P**) – assumes 2 rates of evolution, corresponding to transition and transversion.
* General Time Reversible (**GTR**) – assigns ten parameters corresponding to the rates of transition between each base, and the equilibrium base frequencies. This is the most complex model, and is the model I used when writing my own simulator.

In addition, the rate of evolution at each site is not homogeneous. The best way to model the distribution of rates is to use a Gamma distribution (Yang, 1994). A Gamma distribution has two parameters: a shape parameter α, which controls the peaked-ness of the distribution, and a scaling parameter β, which was set to 1/α to ensure that the mean of the distribution was equal to one.

The Gamma model used is not continuous – it assigns equal numbers of sites to a pre-defined number of categories of rates. Three rate categories, simulating the rates of each of the three sites in a codon, were used in each simulation.

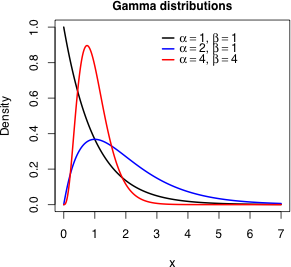


Figure 2: The variation of the Gamma distribution with parameter values.

### Indel Events

The second main element to simulate is insertion-deletion (indel) events. An indel is an event where a sequence of nucleotides is removed from or added to the main sequence, at random. This type of event can lead to frameshift mutations, where every codon sequence is changed, if the length of the inserted or deleted sequence is not a multiple of three.

Some areas of genetic sequences are more likely to undergo indels, however this fact is not taken into account by any of the programs I used to model the evolution. To model the variation in length of the indels, several distributions can be used: the negative binomial, the inverse power law, or the Lavalette distribution. All these distributions have a roughly similar shape – a large value close to zero, decreasing rapidly closer to one, and flattening out as they approach infinity. In practice, this means that indels of long length are much less likely than those of short length. (Yang, INDELible, 2010)

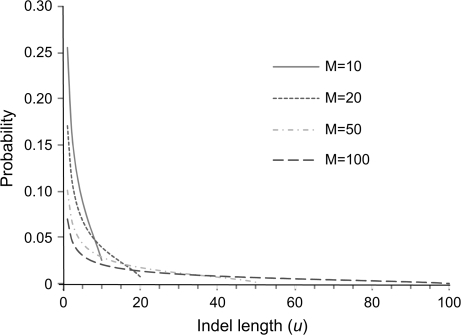


Figure 1: The Lavalette distribution. The value M represents the predefined maximum allowed length of the indels.

## Aims

The aims of this project were as follows:

* Become familiar with the theories of nucleotide evolution.
* Learn how to simulate non-deterministic scenarios.
* Design a simple simulator and contrast its results with more complex programs.
* Compare the results of each program, including my own, to real sequences.

### Comparing Sequences

To compare the results of the programs, the complexity of the sequences was measured. As no two simulated sequences from the same program are likely to be the same, as the programs are probabilistic in nature, it was necessary to use a measure of the information stored in a sequence instead. To this end, the metric chosen to compare the complexity was the Shannon entropy of the sequences.

# Method

## Programs Simulated

There were three programs which were used to simulate sequence evolution:

* DAWG: A command-line program which took a text file containing the model of evolution, the tree to be evolved along, the alpha value and category number for the gamma distribution, the length of the sequence, and the indel model. (Cartwright, 2009)
* INDELible: A command-line program which took a text file containing identical parameters to the DAWG program. (Yang, INDELible, 2010)
* EVOLVER: A GUI-program which took as input the length of the sequence, the tree to be evolved, the model of evolution, the alpha value for the gamma distribution, the number of categories for the gamma distribution, and the base frequencies. It had no mechanism to simulate indels. (Yang, PAML, 2015)
* My own program: Written in Java, run from my development environment. It can only simulate substitutions, and not indels. It uses the GTR model of substitutions and the Gamma model for site rate variance (Link to a Google Drive folder containing source code in the references). (Watson, 2015)

For each program, 10 taxa were simulated, with a start length of 1000 nucleotides. The tree to be used was first estimated using a phylogenetic inference program, and then used in each simulation program. The input files for both INDELible and DAWG are available in the shared Drive folder linked above.

## Parameter Optimisation

The parameters for the models were chosen by trial and error. Realistic starting values were chosen, and incremented individually for each parameter over a small range. For example, for a starting value of 4, the simulation was rerun in the range 2-6, in increments of 0.1. The entropy of the sequence generated by each of these repeats was calculated, and the value of the parameter which gave the smallest entropy was chosen.

|  |  |
| --- | --- |
| **Parameter** | **Value** |
| A-C | 1 |
| A-G | 3.9 |
| A-T | 0.35 |
| C-G | 1.7 |
| C-T | 3.7 |
| G-T | 1 |
| Alpha | 1.6 |
| A-Frequency | 0.17 |
| C-Frequency | 0.28 |
| G-Frequency | 0.17 |
| T-Frequency | 0.38 |
| Indel Rate | 0.0125 |
| Maximum Indel Size | 50 |

For most parameters, this was not a problem. However, in particular the indel rate and model were found to have a greater effect on the entropy of the simulated sequence. Given my inexperience and lack of knowledge surrounding indels, the results would be far more meaningful if this area was analysed in depth.

Additionally, the estimation of the base composition was very difficult, as each of the four parameters had to add to one. Therefore, it was impossible to vary a single parameter, but had to instead vary all four simultaneously, making estimation of each single parameter very difficult.

*Figure 2: The optimized parameters*

## Analyzing Sequences

The analysis of the real and simulated sequences was carried out using the Phylogenomics Dataset Browser. (Parker, 2015) This program takes a Fasta or Nexus sequence file as input, and displays the sequences in table form, as well as calculating the entropy of the sequence, the number of invariant sites, the length of each sequence, the number of taxa, the number of codons represented, and so on.

# Discussion

Although I did not complete my analysis, several conclusions could be drawn from my experience:

* Although EVOLVER could not simulate indels, it could still produce sequences with entropy values very close to the real, approximately 0.04. This shows that another method of evaluating the complexity of the simulated sequences is required, as the real sequences had multiple indel events. I feel that another metric, in addition to entropy, is required.
* Both INDELible and DAWG could produce sequences with entropies in the same order of the real sequence, though usually higher. I also had no way of checking the accuracy of the indel model chosen. Additionally, the number of invariant sites in the simulated sequences of both programs were consistently lower than in the real sequences.

Some areas of possible extension include:

* Extending my own program to include a Newick tree input mechanism, as well as an indel model.
* Varying the length of the sequences simulated, and measure the entropy of the sequences produced, so as to gauge the effect of length upon the simulation’s accuracy.
* Comparing the different models of indel occurrence (power law, negative binomial, Lavalette, etc) and use another metric as well as entropy to gauge their complexity.

# References

Alex J. Drummond, R. R. (2014). *Bayesian Evolutionary Analysis.*

Cartwright, A. (2009). *DAWG*. Retrieved from Reed A. Cartwright: http://scit.us/projects/dawg

Felsenstein, J. (2004). *Inferring Phylogenies.* Sunderland: Sinauer Associates, Inc.

Parker, J. (2015, 08 21). *Phylogenomics Dataset Browser*. Retrieved from Github: https://github.com/lonelyjoeparker/qmul-genome-convergence-pipeline/blob/master/trunk/bin/PhylogenomicDatasetBrowser.jar

Watson, J. (2015, August). *Own Program.* Retrieved from https://drive.google.com/folderview?id=0By\_ODAJCH67UfkI4V2VQcm9keG92YXRhSUd2TXdQMlh2VFJZY1pDbjZveUVFSFEtemthT2M&usp=sharing

Yang, Z. (1994). Maximum Likelihood Phylogenetic Estimation from DNA Sequences with Variable Rates over Sites: Approximate Methods. *Journal of Molecular Evolution*.

Yang, Z. (2006). *Computational Molecular Evolution.* Oxford: Oxford University Press.

Yang, Z. (2010, May 05). *INDELible*. Retrieved from UCL: http://abacus.gene.ucl.ac.uk/software/indelible/

Yang, Z. (2015, August 20). *PAML*. Retrieved from UCL: http://abacus.gene.ucl.ac.uk/software/paml.html