Megan Nguyen
AS.410.712.82.FA23

Project Report for the dN/dS Ratio Analysis Tool

Project Background

Genetic mutations to the gene's DNA sequence can result in permanent changes to the proteins produced, and can either be harmless or detrimental. The ratio of nonsynonymous to synonymous substitutions (dN/dS) serves as a valuable measure for assessing the intensity and mode of natural selection operating on protein-coding genes. It helps researchers understand whether natural selection favors or acts against changes in protein sequences, essentially reflecting natural selection strength. Typically, scientists assume that synonymous changes are selectively neutral, meaning they neither provide an advantage nor a disadvantage to the organism's survival and occur in a population over time. Therefore, the ratio is expected to equal 1 if there is no selection against nonsynonymous substitutions. Ratios below 1 suggest selective constraint (i.e., negative selection), while ratios above 1 indicates an excess of nonsynonymous substitutions.

Currently, there are a few tools that can be found online that may perform these calculations, such as MEGA or the dN/dS ratio R function in R programming. The dN/dS ratio can be calculated on pairwise or multiple sequence alignments depending on research purposes. In pairwise analysis, the ratio is computed by comparing the sequences of two distinct species or genes. More precisely, our objective is to determine whether the dN/dS ratio is significantly greater than 1. Depending on which alignment is performed, the results can be assessed in a variety of different interpretations.

For this project, the downstream analysis will be performed on pairwise sequence alignments on two individual species under one gene. Users can assess the types of selection acting on the compared sequences from the analysis. A pre-aligned FASTA file in DNA/nucleotide format of the two species of choice is required to be uploaded to the web interface. There, the user can view results for the dN/dS ratio, and the amount of synonymous and nonsynonymous substitutions and assess results based on the dN/dS key.

Project Proposal: Initial

I am interested in the methods used in evolutionary biology, specifically synonymous and nonsynonymous substitutions. The original project proposal was oversimplified; requiring just a single DNA sequence from the user initially seemed ideal, but ultimately, for dN/dS ratio calculations, a lot more is needed. In particular, I reframed the tool around pairwise sequence alignments. I have not found official tools for calculating the dN/dS ratio that specifically handle pairwise sequence alignment files. The revised proposal follows the same concept; however, it includes other steps and calculations necessary to acquire a meaningful dN/dS ratio.

Project Proposal: Revised

Typically, dN/dS ratio analysis are performed on pairwise sequence alignment or multiple sequence alignments. Ultimately, I focused the tool on pairwise alignments. I had to determine whether the user could either upload their own pre-aligned file, or I include that alignment ability in my tool. If I was allowing the user to submit their own DNA sequences for alignment, the alignment would prove to be difficult as some sequences were too large and memory intensive (especially when using Biopython's pairwise alignment module). This was a huge limitation, as I wanted the user to do the least amount of work when using the tool. This is something I may incorporate later.

I revised the project where the user would upload their own pairwise alignment and downstream analysis would be provided. To keep things consistent, I specifically list NCBI GeneBank as a source to retrieve DNA sequences. I also included instructions for the user to use alignment tools from https://

www.ebi.ac.uk/Tools/psa/, though they may align their sequences elsewhere. The goal of this project is to calculate the dN/dS ratio per site of protein-coding genes on pairwise sequence alignments so that users may compare the rates of nonsynonymous (dN) and synonymous (dS) substitutions and determine the strength and mode of natural selection acting on the protein-coding gene of their choice.

Project Technologies and Methods

The dN/dS ratio is calculated following the Nei & Gojobori method and the Jukes & Cantor model (JC69), where the latter attempts to correct for the multiple substitution pathways. The JC69 model assumes equal base frequencies and equal mutation rates. Thus, the rate of nucleotide substitution is the same for all pairs of the four nucleotides A, T, C, and G. I also referenced the methodology used here: https://bioinformatics.cvr.ac.uk/calculating-dnds-for-ngs-datasets/.

I tested a series of pairwise alignment FASTA files from EMBOSS Needle:

/var/www/html/mnguye87/final/test_pwa_files

Alignments were performed using tools from https://www.ebi.ac.uk/Tools/psa/ and were output in Pearson/FASTA file format. The alignments files were conducted on genes I found on NCBI GenBank, where I selected two different species (usually homo sapiens and another species), and performed the alignment on the sequence of each species.

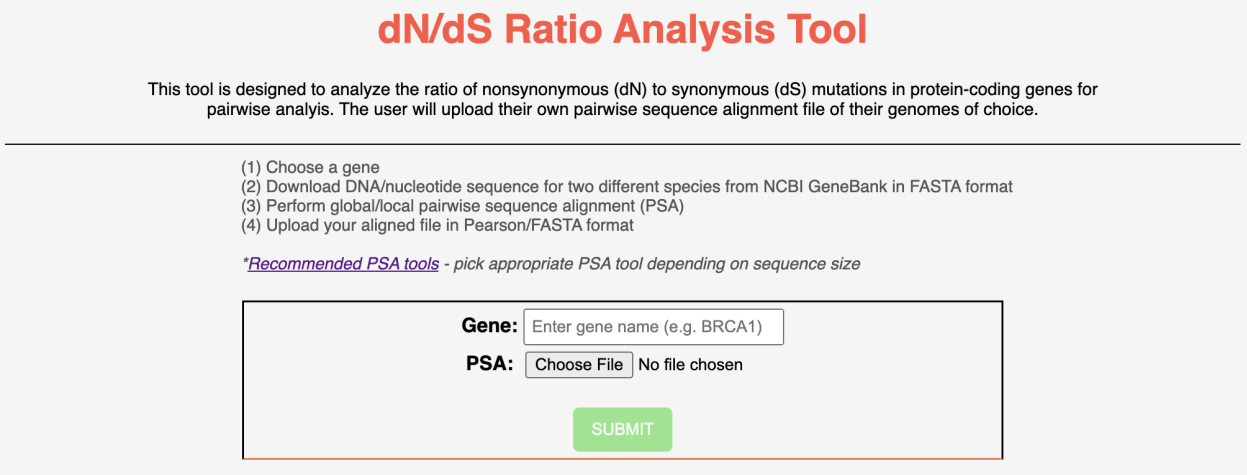
Discussion: Validation and Challenges

The completed project is a web interface that accepts a FASTA file that has been aligned via pairwise sequence alignment tools. The tool outputs the dN/dS ratio of the aligned sequences.

The accuracy and reliability of the tool is to be determined. The interpretation of dN/dS calculations appear to not be standard or universal, despite the many references to the Nei & Gojobori method online and numerous models created by researchers. I followed the method that was most frequently reference online. Theoretically, the tool should be able to handle relatively larger sequences, but that has not been put to the test (specifically extremely large sequences).

While running tests on PWA files, I obtained multiple dN/dS values greater than 1. These values suggest an excess of nonsynonymous substitutions. In the testing of the BRCA1 alignment between chimpanzee and human, I anticipated finding more synonymous mutations due to their close relationship as species. However, the results yielded a dN/dS value greater than 1. Nonetheless, the tool still functions as I intended, providing a dN/dS ratio based on the types of mutations present. A dN/dS value greater than 1 indicates a higher occurrence of nonsynonymous substitutions and sites as expected, which confirms positive selection, where nonsynonymous mutations are favored.

Below are images of the tool: the input, results and the mySQL table that stores the calculations.



Results

Gene: BRCA1

dN/dS Key

dN/dS < 1: Negative (purifying) selection, where synonymous mutations are favored over nonsynonymous mutations
dN/dS = 1: Neutral selection, where synonymous and nonsynonymous mutations occur at the same rate
dN/dS > 1: Positive (adaptive) selection, where nonsynonymous mutations are favored

Sites	Seq1	Seq2	Seq1 and Seq2
Synonymous Sites	17889.0	17889.0	35452.333
Nonsynonymous Sites	62907.0	63232.667	126139.667

Synonymous and nonsynonymous substitutions are totaled between the two sequences.

Synonymous Substitutions	Nonsynonymous Substitutions
1025	25405

pN	pS	dNdS Ratio
0.201	0.029	7.911

```
MariaDB [mnguye87_chado]> SELECT * FROM final_project5;
+----+---------+--------------+-------------------+---------------+-----------------+---------------+-----------------+-----------------+-------------------+-------+-------+------------+
| ID | Gene_ID | Syn_subs_total | Nonsyn_subs_total | Syn_sites_seq1 | Nonsyn_sites_seq1 | Syn_sites_seq2 | Nonsyn_sites_seq2 | Total_syn_sites | Total_nonsyn_sites | pn    | ps    | dNdS_Ratio |
+----+---------+--------------+-------------------+---------------+-----------------+---------------+-----------------+-----------------+-------------------+-------+-------+------------+
|  1 | INS     |           22 |               352 |       310.667 |         838.333 |       310.667 |         858.667 |             601 |              1697 | 0.207 | 0.037 |      6.384 |
|  2 | ALB     |          229 |              5230 |       3523.67 |         13123.3 |       3523.67 |         13227.3 |         6943.33 |             26350.7 | 0.198 | 0.033 |      6.812 |
|  3 | H1-1    |            7 |               171 |           181 |             575 |           181 |             595 |             342 |              1170 | 0.146 |  0.02 |       8.01 |
+----+---------+--------------+-------------------+---------------+-----------------+---------------+-----------------+-----------------+-------------------+-------+-------+------------+
```

Project File Structure

/var/www/html/mnguye87/final

/css/ CSS styling for the entirety of HTML code used

/img/ images used for the project

/js/ javascript code that I ended up not using

/templates/ template used to display results (results.html)

/test_pwa_files/ pairwise sequence alignments files I aligned for testing

/test_scripts/ scripts that I used for testing the code before implementing HTML

/user_upload_files/ user uploaded files are stored here

codon_table.py codon map dictionary used for the calculations

dnds_tool_input.html template used for user input/FASTA file upload

dnds_tool.cgi performs the dN/dS calculations, sends to GUI and displays results to user

README.txt documentation for tool usage

Sources

Nei M & Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular Biology and Evolution 3:418-426.

Jukes TH, Cantor CR (1969). Evolution of Protein Molecules. New York: Academic Press. pp. 21–132
https://bioinformatics.cvr.ac.uk/calculating-dnds-for-ngs-datasets/