7-21-2016

Ciliate Codon Translator Program Manual

Quentin D. Altemose
Ursinus College, qualtemose@ursinus.edu

Ciliate Codon Translator Program Manual

By Quentin Altemose

Program Overview

The Ciliate Codon Translator functions as a method of producing DNA to amino-acid translations based on the ciliate codon table. In addition, this program has matching capabilities programed into the functionality (discussed in more detail in the "match function" section). The program has lines in the code where the user inputs the text file containing the properly formatted DNA sequences. The user must also provide a destination file for the output to be written to, which is usually in the form of a basic text file. Finally, if using the matching feature, the user must input the name of the file (also in proper format) that contains the other amino acid sequences that will be read in and compared to the translation of the given DNA. When completed, the program will produce an output file that has all of the information regarding the name of the sequence, the number of nucleotides per sequence, the number of codons, the translated amino-acid sequence, and (if using the match feature), the name of the match and the corresponding sequence for verification.

Functionality of the Program

The ciliate translator serves a vital role in any genetic/bio-information field: translating DNA and checking for matches against a database to ensure quality. Through doing this, we are able to better understand the genes that we are working with at the time, and limit any potential errors that may be present in the data (as these errors could influence the results of the overall experiment being undertaken). Many translators exist that are able to translate DNA to amino-acids using the standard genetic code (present in humans for example), however fewer exist that are able to work with other organisms that do not share the same codon table. As such, this program serves to produce translations

for these alternative organisms that have differing codon sequences. As of right now, the program is only able to translate using the ciliate codon table, however this can be improved with a few simple code changes and input from the user.

The Match Feature

The match feature is an optional function of the program that the user may select to turn on or off, however it is highly recommended that it remains on for the purposes of validating sequence quality. The match feature takes a file from the user that contains properly formatted amino-acid sequences that will serve as a database. The program will then look through the database attempting to find a 100% match between the translated sequences and the sequences located in the database. At the moment, the program is only capable of detecting 100% matches, and while this represents the ideal situation, due to random genetic mutation and variation between organisms, this will often not be the case. Future builds of this program will need the additional functionality of comparing all translated sequences to all database sequences and producing the percent identity match, which provides a better understanding of how closely related two sequences are. Additionally, more sequence validation features should be added in order to ensure that the sequences are truly related. For example, the addition of a query cover value would allow the program to inform the user about how much of the database sequence matched the translated sequence, allowing them to better understand how close of a match the sequences really are. Additionally, adding a statistical likelihood feature would better allow the user to understand what the probability is that the differences between the given and archived sequences are due to random chance. These features would allow the user to better understand the data that they are working with, and as such they could ensure that the sequences that they are working with are of high quality and will produce meaningful results.

Nucleotide Getter Program

The nucleotide getter program functions as a method of finding any available nucleotide sequence and retrieving them from the NCBI sequence database. The program is a Perl script and functions as a simple getter that retrieves information based off of the input from the user. To use the program, you must run the Perl start command from a command line in UNIX. Then provide the path to a file containing all of the GI identification numbers for all of the sequences that are of interest in list form (that is, one number per line). Finally, provide the name of an output file that will be written to during the completion of the run. The program will connect to the NCBI database online and look up all of the nucleotide sequences provided. Any sequences that are available will be written to the output file.

Big Picture

So what does all of this data mean in the big picture? In order to better understand this question, we will look at the field of bioinformatics and the process of tracing back phylogenies of ciliate genes. In order to trace back the history of some gene in an organism, we must first find a model that best fits the DNA and its potential to change over time. Programs exist (such as ProtTest) that are able to take DNA input and find the model of best fit for it. These models are based off of known genetic mutation probabilities, and hundreds have been created in order to find the best model for different situations. In order to determine the best model, the DNA is run through each possibility, and scored based off of separate criteria each with its own importance value. The highest scoring model is considered to be the model of best fit, and while it may not reflect the evolutionary history for the entire timeline of the experiment, it provides the best likelihood of recreating the history of the organisms under study.

After finding the best model, we must actually use the model to produce the history of the organism. Historical phylogeny programs (such as PhyML) are able to take amino-acid sequences and specific models, and produce a series of potential models showing the evolutionary history of the gene. Bootstrapping at this point is essential, as it repeats the run multiple times and results in the one tree that has the highest likelihood of being the correct tree out of all of the others tested. Additionally, these programs are able to measure the likelihood of each gene being located where it was placed on the final tree, allowing the user to see where the model functions, and where it falls short.

Finally, programs such as PAML are able to test the data for selective pressures occurring on the genes throughout history. This is most often done through finding the dN/dS ratio (that is, the ratio of changes in DNA that results in a new protein over the changes that produce no change in the protein). Results greater than 1 indicate that there is a positive selective pressure occurring on these genes for these organisms (i.e. that selection is favoring genetic change in the population). Ratio values equal to 1 however generally indicate that there is no strong selective pressure for change or stagnation in the population, and ratio values less than 1 indicate a strong stabilizing pressure acting on the population (that is, selection is favoring the population to remain genetically the same).

These results can lead to a variety of inferences and understandings regarding the history of the organism and its genetic code, however this data stems from the amino-acid and DNA data, and the accuracy of that data. Everything that has occurred in this thought experiment was completely dependent on the amino-acids and DNA, and that is where these programs assist. First off, if only given one piece of the code (that is, you are only given DNA and no amino-acid sequences or vice versa), these programs allow the user to acquire the other pieces of information that they need in order to finish their experiment. Additionally, these programs serve a vital role in the process of understanding the genetic history of organisms, and that is ensuring the quality of the DNA and amino-acid sequences. Most of the work done in this thought experiment is based off of likelihood statistics, and statistics are only as

accurate as the data used. If the sequences were incorrect or had errors in them, the resulting phylogeny would most likely be incorrect. Through data checking, we ensure that the sequences have been previously found and recorded, and that the sequences that are used in the statistical calculations are logical and will produce a highly likely genetic history for the organism. This work is essential to understand the history of our plant and the organisms we share it with, and additionally allows us to solidify our knowledge of genetics and the expression of genes in populations. Only through understanding and ensuring the quality of the most basic building blocks of genetics (DNA and amino-acids) can we successfully continue the experiments performed in the field of bioinformatics.