Using Machine Learning Models to Classify Tick Genus Based on Molecular Features

Avery Murphy
ID: 1131559
Master of Bioinformatics

University of Guelph
Guelph, ON
amurph22@uoguelph.ca

ABSTRACT

Ticks can carry a variety of harmful pathogens that can be spread to humans when they are bit leading to severe illness and disease. To manage this threat pest management strategies are implemented to control the boundaries of these parasites. This study aims to provide a tool to separate 4 genera of ticks; *Ixodes, Haemaphysalis, Rhipicephalus* and *Amblyomma*, from the family *Ixodida* by molecular features of their cytochrome C oxidase (COI) gene. This will provide an alternative method to relying on physical features to separate harmful genus of ticks.

To accomplish this machine learning classification methods were used to distinguish the tick genera from one another based on 3 length K-Mer frequencies in the COI gene. The models were trained on a subset of the data where the genus is known and were tested on a separate validation set. Supervised learning models random forest, support vector machine and gradient boosting machine were used. Every model was able to classify each tick sample to its genus with 99% accuracy. A *Rhipicephalus* sample was misclassified as *Amblyomma* by each model. Although all the models preformed the same in terms of accuracy, the random forest provided the simplest execution and was the fastest. Overall, the Tick genus could be separated by their COI gene features with high accuracy.

INTRODUCTION

The patristic arachnid *Ixodida* commonly known as Tick's feed on the blood of hosts including humans and other mammals [1]. The *Ixodidae* family of ticks is characterized by their hard outer shells, and they feed on their hosts during the early stages of development [1]. Despite the small size of these parasites, they possess a very significant threat to human health. Ticks act as vectors for pathogens; viruses and bacteria, that can cause infection in humans when introduced into the blood stream through tick bites [2]. The *Ixodidae* family is a major public health concern due to the pathogens carried that can lead to life threating infections and long-term illness [2]. This includes the genus *Ixodes* that carries the bacterial pathogen responsible for Lyme disease a chronic illness [2,3]. The genus *Haemaphysalis* spreads the virus leading to fatal hemorrhagic fever; the bursting of blood vessels [2,3]. As well the genera *Rhipicephalus* and *Amblyomma* can spread pathogens responsible for serval flus [2].

The best way to protect humans from these harmful illness's is to monitor the vector spreading infection itself by implementing pest management and control strategies [4]. This involves tracking the movement of ticks and identifying the location of their reservoirs where larvae are hatched [5]. This is a very time-consuming process, and distinguishing types of ticks from one another by their morphological features spare no difficulty [6]. Ticks look very similar in their early stages of development when the *Ixodidae* family is predominately feeding on hosts and posing the greatest risk [1,6]. Therefore, a tool that allows for quick tick identification and does not rely on morphological traits is needed. Identifying ticks based on their genetic properties would help address both these issues. This could in turn be used to identify tick habitats through environmental DNA (eDNA) samples such as water even without the presence of a physical tick [6,7]. This could have the potential to detect harmful tick genus's early in their development so that the appropriate management strategies can be implemented quickly ultimately protecting human health.

The objective of this investigation is to determine if machine learning (ML) models can be used to classify the Tick genera *Ixodes, Haemaphysalis, Rhipicephalus* and *Amblyomma* based on their molecular features. The goal is to use ML models to correctly classify a tick based on properties of its COI mitochondrial gene sequence. The COI gene has been made the traditional 'barcode' or 'marker' used to classify species and is heavily sequenced in databases [8]. This gene will be utilized for this investigation due to its fast nature of sequence divergence and therefore it should be effective for separating closely related genus in the *Ixodidae* family from one another [8]. Supervised ML models will be used in this investigation to learn the patterns the govern the features of the COI gene sequence in each genus [9]. The ability to correctly classify the genus ticks belong to will be compared between 3 different ML models.

METHODS

Data Acquisition and Cleaning

The following analyses were conducted in R studio. First the COI gene sequences were gathered directly in R from the National Center for Biotechnology Information (NCBI) nucleotide data base. 396 samples were obtained for each genus. *Amblyomma* had the least number of samples (396) so every genus was set to have this maximum number of samples to ensure a balanced data set. The DNA sequences were then checked for missing and unknown nucleotides. There were no sequences with missing nucleotides but those that consisted of 1% or higher of unknown nucleotides in the sequence were excluded from the study. Ticks typically have a COI gene of 580 base pairs (bp) [10]. The sequence lengths of each genus were examined in **Figure.1** The sequences were filtered to only include those between 500 and 700 bp to account for the variation in genus length and differences in PCR amplification of the gene [10].

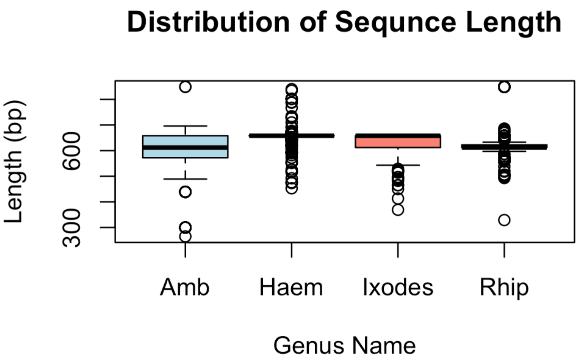


Figure 1: The sequence lengths of each sample in base pairs (bp) is presented per genus: Amb; *Amblyomma,* Haem; *Haemaphysalis,* Ixodes and Rhip; *Rhipicephalus.*

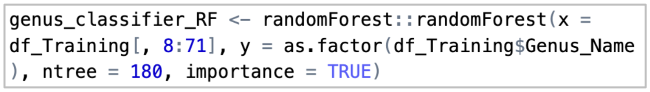
Making the data sets for ML

Once all unwanted sequences were removed, the frequency of each K-Mer was calculated for each sample. A 3-Mer is a short substring made of 3 nucleotides. With 4 different bases there is 64 possible 3-Mer combinations. The frequency of each was determined by the number of times each 3-Mer appeared in the COI sequences relative to its total length. After the frequencies were calculated and added to the data frame it was split into a training and validation set for the ML models. 235 samples per genus; 76% of the total were used in the training set. The remaining 75 samples per genus; 24% of the total were used as the validation data set. This ensured that the number of samples in each genus were even to avoid any bias to one class during the model training [11]. The same training and validation set was used for each ML model to ensure consistency. The models were first trained using the training data set where the model will learn the 3-Mer frequency patterns associated with each tick genera. After the training is complete the models will be tested on the validation set it did not see during the training process. Each model will be trained with parameters to maintain generalizability and reduce over fitting to this specific data set. This is with the intent that these models can be used outside of this data set to classify ticks in the field.

ML Models

1 - Random Forest

The random forest (RF) ML algorithm creates multiple decision trees and merges them together to vote on how to classify a data point, in this case by a genus [12,13]. The algorithm takes a random subset of the data to build each tree to avoid specific or bias learning and keeps the model generalizable [12,13]. As the tress grow and learn the patterns it will determine which features (3-Mer's) are the most influential [12]. Due to these features the algorithm provides randomness to reduce overfitting while the multiple trees provide high accuracy [13]. The code for the model was set up as follows:

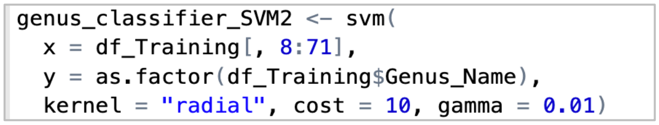


```
genus_classifier_RF <- randomForest::randomForest(x =
df_Training[, 8:71], y = as.factor(df_Training$Genus_Name
), ntree = 180, importance = TRUE)
```

The model was set up to train on the 3-Mer frequencies columns 8 to 71 in the training data set and the genus name was factored out to specify the categories being predicted. There are 63 predictors or K-mer frequencies being used in this model. This is a relatively high number and in turn the model was set to make 180 decision trees [14]. A greater number of descension trees promotes accurate predictions and ensures that each sample is left out of the tress enough times so that it has the chance to be predicted [14,15]. 8 randomly selected features or 3-Mer frequencies were used to split the data at each node of the tree [15].

2 – Support Vector Machine

In support vector machine (SVM), the algorithm separates the data across a boundary in 3D space to find the greatest distance to separate these groups [16]. The algorithm will find the optimal boundary to separate the points that lie close together and form different clusters [17]. The SVM model was run using a radial shape:

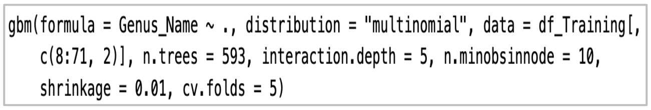


```
genus_classifier_SVM2 <- svm(
  x = df_Training[, 8:71],
  y = as.factor(df_Training$Genus_Name),
  kernel = "radial", cost = 10, gamma = 0.01)
```

The radial shape is standard for data that is not linear so the decision boundary can find its way through the data without having to take on the shape of a straight line [16]. The x values in the model were defined as the 3-mer values in columns 8-71 and y values were the factored-out genus names. The gamma value defines how much influence each data point has on the model during the training process [16]. This value was set to 0.01 to ensure the decision boundary was not too specific to these data points and keep it generalizable [16]. A lower value also lowers the complexity of the decision boundary and in turn lowers the number of points that lie on the decision boundary that are difficult to separate [16]. The cost value represents the deduction implemented when a data point is classified incorrectly across the decision boundary during training [18,19]. A value of 10 was used in this model to ensure accurate prediction of a tick's genus was prioritized.

3 – Gradient Boosting Machine

The gradient boosting machine (GBM) algorithm like FR builds a series of decision trees [20]. However this model builds a tree then attempts to make a prediction, the following tree built will learn and correct for the errors in the prior prediction [20]. This self-correcting tool allows the model to learn as it goes providing high accuracy [20]. GBM also offers cross validation during the training stage where it splits the data into sets to train and test its predictability [21]. The GBM models code was set up as follows:



```
gbm(formula = Genus_Name ~ ., distribution = "multinomial", data = df_Training[,
    c(8:71, 2)], n.trees = 593, interaction.depth = 5, n.minobsinnode = 10,
    shrinkage = 0.01, cv.folds = 5)
```

The distribution of the model is classified as multinomial due to the 4 different genus it will be classifying [21]. The 3-mer frequencies and genus were extracted from the training set. The number of trees was set to 593 as suggested by the algorithm as this was the optimal value, any trees created past this value were not significant and could create overfitting to the data [20]. The interaction depth was set to 5, specifying that a maximum of 5 features can interact in each tree [21]. A moderate value was chosen as higher values increase the complexity of interactions and can again lead to over fitting [21]. The parameter n**.**minobsinnode represents the number of observations that must be found at the last node of each tree [19]. 10 observations were used as opposed to the lower default to generate more simple trees that would not be specific to a few samples [20,21]. Shrinkage refers to the time each tree will take to learn, a smaller value often yields more accurate results [19,20]. This value was set to a small value 0.01, although typically this would require more trees in the model [20,21]. Finally the cross validation cv.fold was set to 5. This mean the data was trained on 4 subsets of the data and tested on the remaining subset, this was then repeated 4 more times.

RESULTS

Once each model went through the training process it was tested on the validation set. It was found that the tick samples could be classified as their correct genus 99.6% of the time using each model. All 3 models preformed the same and the results are presented in **Figure 2**. It is seen in **Figure 2** that there was one misclassification, a *Rhipicephalus* sample was incorrectly classified as *Amblyomma.*

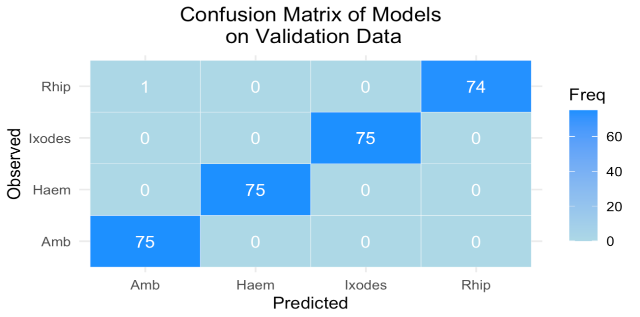


Figure 2: The classification results when the ML models; RF, SVM and GBM, were used on the validation dataset. Amb; *Amblyomma,* Haem; *Haemaphysalis,* Ixodes and Rhip; *Rhipicephalus.*

As a result, the genus *Rhipicephalus* had 98.7% sensitivity, the percentage it could be correctly identified. The remaining genus showed 100% sensitivity consistent across all 3 models. Each model preformed the same, the effects of different parameters on the model's accuracy will be further discussed.

1.1 – Random Forest

As previously stated, RF only classified one sample incorrectly, a *Rhipicephalus* sample as *Amblyomma*. Initially the RF model was run with 130 decision trees a smaller number than presented. With this number of trees there was a 0.96% estimated error rate when the trees were used to vote on the class of the sample during training. Although this is very low increasing the number trees to 180 reduced the error to 0.85%. However, the number of trees did not impact the final accuracy of this model as it was the same for both values. The effects of these parameters are summarized in **Table 1**. The average time a sample could be predicted during training of the RF model was 66 times. Training this model and validating it only took a few seconds.

	Number of Trees	
	Less Trees (130)	More Trees (180)
Classification Error During Training	Lower	Higher
Accuracy	No impact	No impact

Table 1: The effect of changing parameters in the RF algorithm on model error during training and final accuracy.

2.1 – Support Vector Machine

The SVM model had the highest accuracy (99%) when it was trained with a radial shape as shown in the methods. The model was also run using a polynomial shape with the same parameters, but this resulted in high misclassification of samples from the *Rhipicephalus* genus indicting this was not the best shape to form the decision boundary. With a gamma value of 0.01 there was 156 data points lying close to the decision boundary that the model had to distinguish. When the gamma value was increased to 0.1 putting more emphases on each point this value increased to 249. Increasing the gamma value made the decision boundary more complex and in turn there was more misclassification of the *Rhipicephalus* genus when the model was validated bringing its sensitivity down to 97%. The gamma value at 0.01 gave optimal results for this model. The cost value did not have an impact on the model when both lowered or raised from the value 1. The effects of the parameters are summarized in **Table 2**. This model preformed best when the decision boundary was kept simple and could be trained and validated quickly.

	Gamma Value		Cost Value		Shape	
	Low 0.001	High 0.1	Low 1	High 10	Radial	Polynomial
Accuracy	Higher	Lower	No impact	No impact	Higher	Lower

Table 2: The effect of changing parameters in the SVM algorithm on model accuracy

3.1 – Gradient Boosting Machine

Implementing the GBM models suggested number of optimal of decision trees 593 the accuracy remained 99% with the parameters shown in the model's code providing the optimal results. When the cv.fold was increased to a value of 10 it did not change the accuracy of the model but increased the time it took to train the model. Lowering the shrinkage value from 0.01 to 0.001 to increase the learning time lowered the accuracy only slightly to 99.3% contrary to what usually occurs with this parameter. The effects of the parameters for this model are summarized in **Table 3**. This model took much longer to train and validate compared to random forest.

	Cross	Validation	Shrinkage	
	Lower 5	Higher 10	Lower 0.001	Higher 0.1
Accuracy	No impact	No impact	Lower	Higher

Table 3: The effect of changing parameters in the GBM algorithm on model accuracy.

DISSCUSSION

It is evident that the tick genus in the *Ixodidae* family could be effectively separated by the K-Mer frequencies of their COI gene. This method was highly accurate using any of the demonstrated ML models. This provides a new tool that removes the reliance on taxonomic key cards to decipher different genus [22]. This proves to be very challenging due to ticks' similar appearance and taxonomic cards are often limited in reporting distinguishing features of ticks during early development [22]. Using these ML models also abolishes the traditional identification methods that require alignment [22]. Prior work to identify tick genus by their COI gene has required alignment of the sequences to reference data bases [22]. Alignment based methods such as BLAST (Basic Local Alignment Search Tool) require long running times and computation [23]. This is not suitable for large data sets that would be generated through tick surveillance strategies and eDNA sample collection. Instead, the trained ML models only require 3-Kmer frequencies, and it can quickly classify many samples. This tool by passes the alignment requirement leading to faster results and the timely implementation of control strategies to manage these parasites.

Each ML model was trained on the data with parameters to increase generalizability and to reduce over-fitting to this specific data set. This was to ensure that these models could be used as a classification tool outside of this data and in the field. In the end each model preformed the same and they all shared the same misclassification of a *Rhipicephalus* sample. Although no model can be chosen to produce more accurate results, each had their unique strengths and weaknesses. SVM and GBM each have parameters that could be directly manipulated to reduce overfitting to the data set. Although the RF algorithm is random by nature it is heavily influenced by the number of trees used and could not be manually adjusted like the others. GBM offers high confidence in its results due to its error correction algorithm and cross validation. However, because of this model took significantly longer compared to RF and SVM. For this classification problem RF sufficed as the optimal method choice as it is the most user friendly to set up and the model and was extremely fast. The random subsets of data taken at each split tree ensures it is generalizable despite directly altering parameters. As well the contribution of all trees to vote on the classification of each sample offering high accuracy.

As stated, every model misclassified a single *Rhipicephalus* as *Amblyomma*. With further investigation it was discovered that it was the same exact *Rhipicephalus* sample that was misclassified by each model. It was thought that this sample may have an abnormal COI gene length, but it was 612 bp with the average in the *Rhipicephalus* family being 616 bp and *Amblyomma* 631 bp. As well the 3-Mer frequencies were compared with other *Rhipicephalus* samples, and its values were very consistent. A local alignment was done between the misclassified *Rhipicephalus* sample and a *Amblyomma* sample to determine their similarity [24]. There was an 81% similarity between the 2 sequences. This was re-run with a few other *Rhipicephalus* samples, and they also had 81% similarity. This suggests that there was not increased similarity of this specific sample to the *Amblyomma* genus. Therefore the misclassification may be a result of the *Rhipicephalus* and *Amblyomma* being the most closely related genera between the 4 in this study [25]. Although it was only 1 sample misclassified in this study it did not appear to be an outlier. It was not removed as it represents real data that could be encountered in the field. A future direction of this study could be to re-run the models with K-mers of length 4. This would increase the number of features included in the model use to separate the classes and could increase the genetic variation captured.

There were some limitations in this study that could have had impacted the results. Having balanced datasets for the ML models were prioritized but in turn that made the size of the dataset smaller as it was limited to the genus with the minimum number of samples which in this case was *Amblyomma.* Although the total sample size in the training set was 10 times that of the amount of features the suggested amount, more samples per genus would over all feed more data to the models [26]. As well when samples were chosen from NCBI for each genus, the number of different species that were apart of each genus were not specified for. Therefore, many samples in a genus could be predominantly from one species, which could lead to an inaccurate representation of 3 K-Mer frequencies for the COI gene in that genus. This could in turn make the classifier less versatile if the model was trained on sequences from only a few members of each genus. To diminish the possibility of this, a future step should be to get the species directly from NCBI and confirm even amounts came from each species and then combine them into a data frame. To further access the accuracy of the classifier the models could be run on new data, for instance tick COI gene sequences from the (Barcode of Life Data Systems) BOLD. If the same results were found this would ensure the model is reproducible and offer insight into its generalization. None the less this study showed that a variety of ML classification models offer a promising tool to separate tick genus based on COI K-mer frequencies. This also provided insight into the impact of different parameters on these algorithms.

REFERENCES

[1] Watson, S. (2024, October 13). *What are ticks an what* diseases *do they spread?*.WebMD.https://www.webmd.com/skin-problems-and-treatments/ticks-and-the-diseases-they-spread

[2] CDC (Centers for Disease Control and Prevention). (2017, December 31). *CDC - dpdx - ticks*. https://www.cdc.gov/dpdx/ticks/index.html

[3] Dye-Braumuller, K. C., Gual-Gonzalez, L., Abiodun, T., Rustin, L. P., Evans, C. L., Meyer, M. M., Zellars, K., Neault, M. J., & Nolan, M. S. (2023). Invasive Haemaphysalis longicornis (Acari: Ixodidae) investigation in South Carolina: new records of establishment, pathogen prevalence, and blood meal analyses. Journalvof medical entomology, 60(6), 1398–1405. https://doi.org/10.1093/jme/tjad119

[4] World Health Organization. (n.d.). *Vector control*. World Health Organization. https://www.who.int/teams/control-of-neglected-tropical-diseases/interventions/strategies/vector-control

[5] Michelitsch, A., Wernike, K., Klaus, C., Dobler, G., & Beer, M. (2019). Exploring the Reservoir Hosts of Tick-Borne Encephalitis Virus. Viruses, 11(7), 669. https://doi.org/10.3390/v11070669

[6] Intirach, J., Lv, X., Han, Q., Lv, Z. Y., & Chen, T. (2023). Morphological and Molecular Identification of Hard Ticks in Hainan Island, China. Genes, 14(8), 1592. https://doi.org/10.3390/genes14081592

[7] Iacaruso, N., Kopsco, H., Gronemeyer, P., Merkelz, S., Smith, R., & Davis, M. (2024). Design and partial validation of novel edna qpcr assays for three common North American tick (Arachnida: Ixodida) species. Environmental DNA, 6(2). https://doi.org/10.1002/edn3.537

[8] Rodrigues, M.S., Morelli, K.A. & Jansen, A.M. Cytochrome c oxidase subunit 1 gene as a DNA barcode for discriminating Trypanosoma cruzi DTUs and closely related species.Parasites Vectors 10, 488 (2017). https://doi.org/10.1186/s13071-017-2457-1

[9] Xu, C., & Jackson, S. A. (2019). Machine learning and complex biological data. Genome Biology, 20(1). https://doi.org/10.1186/s13059-019-1689-0

[10] Gou, H., Xue, H., Yin, H., Luo, J., & Sun, X. (2018). Molecular Characterization of Hard Ticks by Cytochrome c Oxidase Subunit 1 Sequences. *The Korean journal of parasitology*, *56*(6), 583–588 https://doi.org/10.3347/kjp.2018.56.6.583

[11] Stewart, M. (2020, July 29). *Guide to classification on Imbalanced Datasets*. Medium. https://medium.com/data-science/guide-to-classification-on-imbalanced-datasets-d6653aa5fa23

[12] Donges, N. (2024, November 26). *Random Forest: A complete guide for machine learning*. Built In. https://builtin.com/data-science/random-forest-algorithm

[13] Khushaktov, F. (2023, August 26). *Introduction random forest classification by example*. Medium. https://medium.com/@mrmaster907/introduction-random-forest-classification-by-example-6983d95c7b91

[14] Liaw, A. and Wiener, M. (2002) Classification and Regression by Random Forest. R News, 2, 18-22. https://journal.r-project.org/articles/RN-2002-022/RN-2002-022.pdf

[15] Breiman, L. (2001), *Random Forests*, Machine Learning 45(1), 5-32.

[16] Singh, R. (2024, October 17). *Support Vector Machines (SVM)*. Medium. https://medium.com/@RobuRishabh/support-vector-machines-svm-27cd45b74fbb

[17] IBM. (2024, December 19). What is support vector machine? https://www.ibm.com/think/topics/support-vector-machine

[18] UC Business. (n.d.). *Support Vector Machine*. Support Vector Machine · UC Business Analytics R Programming Guide. https://uc-r.github.io/svm

[19] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., & Leisch, F. (2023). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien* (R package version 1.7-13). https://CRAN.R-project.org/package=e1071

[20] Baladram, S. (2024, November 30). *Gradient boosting regressor, explained: A visual guide with code examples*. Medium. https://medium.com/data-science/gradient-boosting-regressor-explained-a-visual-guide-with-code-examples-c098d1ae425c

[21] Ridgeway, G., Edwards, D., Kriegler, B., Schroedl, S., Southworth, H., Greenwell, B., Boehmke, B., Cunningham, J., & GBM Developers. (2024). *gbm: Generalized boosted regression models (Version 2.2.2)*.https://cran.r-project.org/web/packages/gbm/gbm.pdf

[22] Che Lah, E. F., Yaakop, S., Ahamad, M., George, E., & Md Nor, S. (2016). Preciseidentification of different stages of a tick,Ixodes granulatus Supino,1897(Acari:Ixodidae). *Asian Pacific Journal of Tropical Biomedicine*, *6*(7), 597–604. https://doi.org/10.1016/j.apjtb.2016.05.003

[23] Sievers, A., Bosiek, K., Bisch, M., Dreessen, C., Riedel, J., Froß, P., Hausmann, M., & Hildenbrand, G. (2017). K-mer Content, Correlation, and Position Analysis of Genome DNA Sequences for the Identification of Function and Evolutionary Features. *Genes*, *8*(4), 122. https://doi.org/10.3390/genes8040122

[24] EMBI. (n.d.). *EMBOSS Water - Pairwise Sequence Alignment (PSA)*. EMBL-Ebi Homepage. https://www.ebi.ac.uk/jdispatcher/psa/emboss_water

[25] Burger, T. D., Shao, R., Beati, L., Miller, H., & Barker, S. C. (2012). Phylogenetic analysis of ticks (Acari: Ixodida) using mitochondrial genomes and nuclear rRNA genes indicates that the genus Amblyomma is polyphyletic. *Molecular Phylogenetics and Evolution*, *64*(1), 45–55.https://doi.org/10.1016/j.ympev.2012.03.004

[26] Smolic, H. S. (2024, May 30). How much data do you need for machine learning. Graphite Note. https://graphite-note.com/how-much-data-is-needed-for-machine-learning/#:~:text=The%20rule%2Dof%2Dthumb%20rule,enough %20high%2Dquality%20input%20exists.