OBJECTIVES

General objectives

To construct a series of phylogenetic trees based on COI barcode sequences of arthropod specimens hosted by BOLD database and establish the phylogenetic distribution of these organisms within East Africa. And to conduct phylogeographic assessment based on their associated geographic coordinates and metadata.

Specific Objectives

1. To identify phylogenetic and phylogeographic distribution of fruit flies (Tephritidae family) and biting flies (Diptera: True flies) within East Africa.
2. To identify insect species that are potential pests or disease vectors from combined phylogenetic similarity and phylogeographic distribution.
3. To identify any potential new classification or species if any.

Within the period from 1st November 2018 to end of February 2019, I managed to get registered in the BOLD systems database, got access to both published and unpublished datasets, a total of 279189 (213195 african arthropoda phylum and 202656 Insecta class) records from all 56 African countries (47550 arthropod records from East Africa: Kenya, Tanzania, Uganda, Rwanda, Burundi, South Sudan and Ethiopia). I managed to set up a GitHub account to enable open and collaborative research in the project.

1) Data Mining:

a) I conducted some data mining on this data: data transformation from XML files to text (TSV) files in Python using BeautifulSoup4, lxml and pandas.

b) Then did a study on summary statistics of the data using R tidyverse flavour (dplyr and magrittr packages).

c) Next I conducted data cleaning: excluding non-arthropoda, non-african and only retaining Cytochrome c Oxidase 1 subunit marker sequences. Informed by the summary statistics, we chose to focus on the class insecta out of 11 classes as of the starting point.

d) Informed by the data study I sorted the 192852 insecta records based on the unaligned sequence length and grouped them into categories as follows: (i) those with less than nucleotides; (ii) those with over 700 nucleotides; (iii) those with over 499 nucleotides; (iv) from 500 to 700 nucleotides and (v) from 650 to 660 nucleotides

2) Bioinformatics pipeline development:

From the six groups of data I made eight random samples, 100 records per sample. This sample would be used in the development and testing of different aspects of the pipeline. Based on Literature review I settled on a number of tools to test for different steps of our analysis:

a) Multiple Sequence Alignment (MSA) tools: Analysis; MUSCLE(Edgar, 2004), T_Coffee (Nogales et al., 2018), MAFFT(Katoh & Standley, 2013), PASTA(Siavash Mirarab et al., 2015), SATé(S. Mirarab, Nguyen, & Warnow, 2012), UPP and HMMER. MSA visualization; SeaView (Gouy, Guindon, & Gascuel, 2010), SuiteMSA and Jalview(Waterhouse, Procter, Martin, Clamp, & Barton, 2009). MSA evaluation; T_Coffee's CORE index and Transitive Consistency Scores (Chang, Di Tommaso, Lefort, Gascuel, & Notredame, 2015)

b) Phylogenetics trees tools: Inference tools; RAxML8(Stamatakis, 2014) and FastTreeII (Price, Dehal, & Arkin, 2010). Visualization tools; Archaeopteryx (Han & Zmasek, 2009), Dendroscope (Huson et al., 2007) and Figtree. Analysis tools; Evolutionary Placement Algorithm (Berger, Krompass, & Stamatakis, 2011) in rooting, Biopython's Phylo package in manipulation (Talevich, Invergo, Cock, & Chapman, 2012).

c) Phylogeographic assessment tool: BEAST2 package; BASTA for Phylogeography (Bouckaert et al., 2014).

d) Other tools used were: PGDSpider for sequence data format conversion.

The project is set up to support reproducibility by maintaining well annotated and structured scripts and a well documented workflow done in Jupyter lab. The project is done collaboratively via GitHub, which allows remote supervision. Version control is done using git.

Reference.

Berger, S. A., Krompass, D., & Stamatakis, A. (2011). Performance, Accuracy, and Web Server for Evolutionary Placement of Short Sequence Reads under Maximum Likelihood. *Systematic Biology*, *60*(3), 291–302. https://doi.org/10.1093/sysbio/syr010

Bouckaert, R., Heled, J., Kühnert, D., Vaughan, T., Wu, C.-H., Xie, D., … Drummond, A. J. (2014). BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Computational Biology*, *10*(4). https://doi.org/10.1371/journal.pcbi.1003537

Chang, J.-M., Di Tommaso, P., Lefort, V., Gascuel, O., & Notredame, C. (2015). TCS: A web server for multiple sequence alignment evaluation and phylogenetic reconstruction. *Nucleic Acids Research*, *43*(Web Server issue), W3–W6. https://doi.org/10.1093/nar/gkv310

Edgar, R. C. (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Gouy, M., Guindon, S., & Gascuel, O. (2010). SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Molecular Biology and Evolution*, *27*(2), 221–224. https://doi.org/10.1093/molbev/msp259

Han, M. V., & Zmasek, C. M. (2009). phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, *10*(1), 356. https://doi.org/10.1186/1471-2105-10-356

Huson, D. H., Richter, D. C., Rausch, C., Dezulian, T., Franz, M., & Rupp, R. (2007). Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics*, *8*(1), 460. https://doi.org/10.1186/1471-2105-8-460

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution*, *30*(4), 772–780. https://doi.org/10.1093/molbev/mst010

Lischer, H. E. L., & Excoffier, L. (2012). PGDSpider: An automated data conversion tool for connecting population genetics and genomics programs. *Bioinformatics*, *28*(2), 298–299.

https://doi.org/10.1093/bioinformatics/btr642

Mirarab, S., Nguyen, N., & Warnow, T. (2012). SEPP: SATé-enabled phylogenetic placement. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 247–258.

Mirarab, Siavash, Nguyen, N., Guo, S., Wang, L.-S., Kim, J., & Warnow, T. (2015). PASTA: Ultra-Large Multiple Sequence Alignment for Nucleotide and Amino-Acid Sequences. *Journal of Computational Biology*, *22*(5), 377–386. https://doi.org/10.1089/cmb.2014.0156

Nogales, E. G., Tommaso, P. D., Magis, C., Erb, I., Laayouni, H., Kondrashov, F., … Notredame, C. (2018). Fast and accurate large multiple sequence alignments using root-to-leave regressive computation. *BioRxiv*, 490235. https://doi.org/10.1101/490235

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE*, *5*(3). https://doi.org/10.1371/journal.pone.0009490

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. https://doi.org/10.1093/bioinformatics/btu033

Talevich, E., Invergo, B. M., Cock, P. J., & Chapman, B. A. (2012). Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics*, *13*(1), 209. https://doi.org/10.1186/1471-2105-13-209

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189–1191. https://doi.org/10.1093/bioinformatics/btp033