0.1 Data analysis

The mean time to completion and proportion of species correctly identified by each classifier was analysed in R for all RefSeq and articifical datasets. All data analysis and preparation were carried out in R (4.2.1) [1] using R-Studio (2022.7.2.576) [2] with the package data.table (1.14.6) [3] and visualisations were constructed using ggplot2 (3.4.0) [4] and ggpubr (0.6.0) [5].

In order to assess significance, mean time taken was modelled using a negative-binomial generalised linear model (GLM) from the package MASS (7.3.60) [6]. A negative-binomial model was chosen due to its ability to handle over-dispersed count data like a quasi-poisson model and produce extremely similar estimates. However, unlike a quasi-poisson, a negative binomial model is a full maximum likelihood model meaning that the standard model selection tests like AIC, BIC... etc are still valid making it an easier model to work with. The saturated GLM was constructed and then reduced step-wise until a minimally adequate model was obtained (see equation 1).

$$\log(E(\text{mean\_time(ms)})) = \alpha + \beta_1(\text{classifier}_{\text{kaiju}}) + \beta_2(\text{classifier}_{\text{kraken2}}) + \beta_3(\text{homopolymer\_errors}) \tag{1}$$

Likelihood ratio tests were conducted to obtain p-values and compare the full model to the reduced model. Estimated marginal means were used as a Post hoc test using the emmeans package (1.8.6) [7]. The tests were carried out on the log scale using the Tukey method for comparing families of estimates for p-value adjustment. While the proportion of species identified was modelled using a quasi-binomial GLM; This model was chosen due to the nature of the data being under-dispersed bounded count data, making a quasi-binomial distribution the obvious choice because of its ability to handle under-dispersed proportional data. The saturated GLM was then reduced step-wise until a minimally adequate model was obtained (see equation 2).

$$\begin{aligned}\log\left[\frac{E(\text{proportion\_identified\_species})}{1-E(\text{proportion\_identified\_species})}\right] = {} & \alpha + \beta_1(\text{classifier}_{\text{kaiju}}) + \beta_2(\text{classifier}_{\text{kraken2}}) + \beta_3(\text{level\_of\_difference}_{0.05}) + \beta_4(\text{level\_of\_difference}_{0.1}) + \beta_5(\text{level\_of\_difference}_{0.15}) + \\ & \beta_6(\text{level\_of\_difference}_{0.2}) + \beta_7(\text{level\_of\_difference}_{0.25}) + \beta_8(\text{level\_of\_difference}_{0.3}) + \beta_9(\text{classifier}_{\text{kaiju}} \times \text{level\_of\_difference}_{0.05}) + \\ & \beta_{10}(\text{classifier}_{\text{kraken2}} \times \text{level\_of\_difference}_{0.05}) + \beta_{11}(\text{classifier}_{\text{kaiju}} \times \text{level\_of\_difference}_{0.1}) + \beta_{12}(\text{classifier}_{\text{kraken2}} \times \text{level\_of\_difference}_{0.1}) + \\ & \beta_{13}(\text{classifier}_{\text{kaiju}} \times \text{level\_of\_difference}_{0.15}) + \beta_{14}(\text{classifier}_{\text{kraken2}} \times \text{level\_of\_difference}_{0.15}) + \beta_{15}(\text{classifier}_{\text{kaiju}} \times \text{level\_of\_difference}_{0.2}) + \\ & \beta_{16}(\text{classifier}_{\text{kraken2}} \times \text{level\_of\_difference}_{0.2}) + \beta_{17}(\text{classifier}_{\text{kaiju}} \times \text{level\_of\_difference}_{0.25}) + \beta_{18}(\text{classifier}_{\text{kraken2}} \times \text{level\_of\_difference}_{0.25}) + \\ & \beta_{19}(\text{classifier}_{\text{kaiju}} \times \text{level\_of\_difference}_{0.3}) + \beta_{20}(\text{classifier}_{\text{kraken2}} \times \text{level\_of\_difference}_{0.3})\end{aligned} \tag{2}$$

Likelihood ratio tests were conducted to obtain p-values, and compare the full model to the reduced model. Estimated marginal means were used as a Post hoc test using the emmeans package (1.8.6) [7]. The tests were carried out on the log scale using the Tukey method for comparing families of estimates for p-value adjustment.

1 References

[1] R Core Team. R: A language and environment for statistical computing, 2022.

[2] RStudio Team. *RStudio: Integrated Development Environment for R*. RStudio, PBC, Boston, MA, 2022.

[3] Matt Dowle and Arun Srinivasan. data.table: Extension of 'data.frame', 2023. R package version 1.14.8.

[4] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[5] Alboukadel Kassambara. ggpubr: 'ggplot2' based publication ready plots, 2023. R package version 0.6.0.

[6] W N Venables and B D Ripley. *Modern Applied Statistics with S*. Springer, fourth edition, 2002. ISBN 0-387-95457-0.

[7] Russell V Lenth. emmeans: Estimated marginal means, aka least-squares means, 2023. R package version 1.8.6.