Sam Smith August 2024

IMPERIAL

Phylogenetic Backbone Reconstruction using Kmeans Clustering on Coleopteran Mitogenomes

Supervisor: Alfried Vogler
Department of Life Sciences
Imperial College London
SW7 2AZ
UK

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science at Imperial College London - submitted for the MSc in Computational Methods in Ecology and Evolution

Declaration:

The data used in this study was provided by the Site 100 database, hosted by the Natural History Museum, London, in collaboration with Imperial College London. The data required a minimal extent of cleaning; merely just correcting typo errors and changing file type (i.e. GenBank to Fasta). The SigClust algorithm developed and scrutinized further in this study was originally published by Chappell et al., 2017. All analysis, data cleaning, plotting, and thesis writing was entirely independent, unless otherwise explicitly stated in the text.

Contents

1 Abstract

The challenge of building large phylogenetic trees is not new, but increasingly frustrating due to the growing availability of molecular data. This investigation explores the application of kmeans clustering in aiding the reconstruction of large phylogenies. In particular, Coleopteran mitogenomes from The Site 100 database served as the sequences used to develop the method in this investigation. The method proposed in this study aims to cluster sequences and compute the 'cluster centers' that can consequently serve as representatives for phylogenetic backbone reconstruction. Therefore, the phylogenies of individual clusters (subtrees) can be added to their respective tip on the backbone, thus providing additional phylogenetic detail where desired. Two clustering approaches were experimented in this study: a simple kmeans approach and an adapted version of a previously developed algorithm that also takes advantage of the speed and simplicity of kmeans clustering. The simple approach whereby kmeans was executed directly on a distance matrix struggled to find suitable representatives for bacbone reconstruction. However, the implementation of an adapted signal clustering method (SigClust) showed more promise. Whilst the clusters from this method were mostly not taxonomically homogenous and therefore unlikely to be monophyletic and serve as good subtrees, the cluster representatives did manage to capture a lot of the diversity present in the sample. Despite the often disappointing contents of each cluster, there were cases of good taxonomically pure clusters which were used to illustrate the addition of subtrees onto a phylogenetic backbone.

2 Introduction

Coleoptera consists of a quarter of all known animal species on Earth [Yuan et al., 2016]. There are around 400,000 described beetle species in this highly diverse order. The order has been subject to many phylogenetic analyses based on morphology and more recently through the 21st Century, molecular data [Orlov et al., 2021]. Understanding phylogeny has broad implications in ecology, particularly relating to ecosystem services, pathogen-host interactions, and general community structure [Davis et al., 2010]. Moreover, applying a phylogenetic framework to phenological studies provides insight into the behaviour of closely related organisms, particularly important for understanding their response to climate change [Davis et al., 2010]. Coleopterans are found in all climates and habitats except the most extreme latitudes and altitudes [Gressitt, 2024]; they strongly contribute to ecosystem services in almost all terrestrial and freshwater ecosystems [Zhang et al., 2018]. Investigating the complex phylogeny of Coleopterans is fundamental to understanding the processes that has led to this extraordinary diversity [Zhang et al., 2018].

Determining Coleopteran phylogeny has been challenging due the sparse availability of molecular data relative to the species richness of the order [Zhang et al., 2018]. Whilst studies continue to refine the phylogenetic tree and clarify evolutionary relationships between species; previous phylogenetic reconstructions of the order have provided the monophyly of the four described suborders: Archostemata, Adephaga, Myxophaga, and Polyphaga. The extreme species richness of Coleoptera, coupled with the computational limitations of traditional tree building software, suggests the need for innovative phylogenetic methods to be developed. For example, the computational burden placed on phylogenetic software increasingly worsens upon the addition of species [Nabhan and Sarkar, 2011]. This is particularly frustrating in an era of increasingly available molecular data [Sanderson and Driskell, 2003]. However, the challenge of building large trees is not new and methods have been developed to address these challenges. For example, less dense sampling leading to the development of strongly supported phylogenetic backbones, that capture the broad diversity of a clade without including the detail of every species has become popular practice. It provides means for

a consistent classification for further research [Xu et al., 2022]. However, these backbone trees not only fail to capture all of the evolutionary detail, they also do not currently facilitate the addition of such. The novel method proposed in this study is a 'divide-and-conquer' approach, aimed to divide the available phylogenetic detail at the tip of a tree through clustering, and graft onto a backbone phylogeny where appropriate.

Clustering algorithms are frequently employed to classify genetic data into groups based on similarity. These clusters can then be analysed individually, or a meta-analysis can be conducted using representative sequences from each cluster. This method reduces computational demands, making it a more efficient solution to observe broader evolutionary relationships than analyzing an entire dataset. One popular clustering tool that has not found a mainstream application in phylogenetic reconstruction is kmeans. Kmeans is a fast, effective and straightforward algorithm well-employed in other areas of biological research [Lu et al., 2004]; it partitions data through an iterative process into a predefined number of clusters based on shared variables. One of the few examples of kmeans being used in phylogenetic research is by Chappell et al., (2017). The aim of this study was to cluster sequences from highly diverse environmental samples quickly and alignment free, for consequent analyses to take place. This study converted DNA sequences into kmers and then transformed them into an intermediate binary format before kmeans clustering [Chappell et al., 2017]. This innovative approach clustered sequences with high accuracy, outcompeting alternative clustering methods convincingly, thus making a strong case for kmeans clustering to be used more in this field. Whilst the absence of an alignment in this method makes it incredibly fast, this method has its drawbacks. Firstly, reducing the detail in the sequences during the conversion to a binary vector format, suggests the algorithm is unsuited for a closely related sample of sequences. Secondly, the application of the SigClust software is currently unsuitable for backbone reconstruction as it does not provide a means for selecting representatives from each cluster.

This is an exploratory investigation that describes a novel method for building large phylogenetic trees whilst also building on the work of Chappell et al (2017).

The study particularly provides a solution to the challenge of building very large trees by efficiently combining multiple phylogenetic trees that are originally designed to be grafted together - unlike previous methods. As the need for constructing larger phylogenetic trees grows, it is certain that to include more species, increasing the number of sites per species is essential - but not always possible. To combat this, 'supertree' methods such as matrix representation with parsimony (MRP) [Pisani and Wilkinson, 2002], or MinCut [Kupczok et al., 2010] have been used to combine 'input' trees that contain overlapping taxonomic coverage. The method to build large trees proposed in this study is a logical step-by-step process that combines using kmeans clustering for backbone tree reconstruction with consequent tree grafting. The method uses smaller monophyletic 'subtrees' (from samples identified through clustering) that capture less biodiversity to graft onto the broader backbone tree in their appropriate position. Or equally, one could observe a diverse backbone tree alongside subtrees that give more detail. This is a flexible and efficient solution to building large trees, as subtrees can only be built and added to the backbone wherever is desired and optimal.

This study aims to find the most appropriate representatives for phylogenetic backbone reconstruction (PBR) utilizing a novel k-means approach as well as an adapted version of the SigClust algorithm. Choosing representatives for backbone reconstruction that include the broader biodiversity within a sample of organisms is a challenge for phylogeneticists. Representatives may be chosen based on morphology, taxonomy, a genetic basis, or a combination [Orlov et al., 2021]. Alternatively, representatives do not need to be chosen and backbone trees can be proposed as just a pruned complete phylogenetic tree, consisting of only strongly supported nodes and branches [Creedy et al., 2024]. Picking representatives based on morphology requires the close inspection of specimens and assumes their capture; in addition, using organisms taxonomy requires the effort of labeling data, potentially even at the cost of an expert taxonomist. Therefore, identifying species for backbone reconstruction using exclusively molecular data is favourable, and takes advantage of the accessibility of new and innovative methods of sequencing specimens. Nuclear or mitogenomes can both used as the molecular data for phylogenetic reconstruction. However, mitochondrial genomes have become increas-

ingly popular for such analyses for reasons including: maternal inheritance and lack of recombination, coupled with a high mutation rate [Yu et al., 2022]. Furthermore, this study will be conducted on mitogenomes also due to their greater availability than nuclear genome sequences.

The method developed in this study identifies representatives based solely on their molecular make-up. This is a more reusable approach, and would eliminate the bias introduced by disparities in species diversity between equivalent taxonomic ranks. The adaptation to the SigClust algorithm, proposed in this study, sacrifices the impressive computational speed due to the necessity of distance matrices for calculating cluster centers. Despite this, the flexibility of this method means fortunately only small distance matrices consisting of the sequences of particular clusters need to be made. It is important to note that despite strong efforts to benchmark the performance of different clustering techniques, it is accepted that the capabilities of different methods are too sensitive to the input data for comparisons between different 'clustering' studies to be made [Tseng, 2007]. Therefore, it is difficult to predict the performance of this study and consequently compare it to others.

3 Methods

All methods in this study were carried out using a combination of R [R Core Team, 2024] and Python [Foundation, 2024], as well as more specialised software for alignments and phylogenetic analysis.

3.1 Taxon Sampling

In this study, Coleopteran mitogenomes were used for developing the method. Sequences were obtained from the Site 100 database, including samples from all around the world [Bian et al., 2022]. Only complete mitogenome sequences labelled 'Coleoptera' were used, resulting in 4438 sequences being extracted from the database, hosted by the Natural History Museum, London. All four suborders were included in the sample, most abundantly represented were Polyphaga

and Adephaga - though exact abundances are unknown due to incomplete taxa-specimen labelling in the database. 23 superfamilies were included in the sample with the most represented superfamilies being Chrysomeloidea, Curculionoidea and Staphylinoidea (943, 733, 470 respectively) whilst 413 sequences were not identified to a superfamily level.

3.2 Sequence Alignment and Similarity Scoring

A sequence alignment is necessary for both the computation of a distance matrix and for consequent tree building. The execution of the SigClust approach does not require an alignment but may benefit from the distance matrix for the identification of cluster centers. For the sequence alignment, all mitogenome sequences were partitioned into their protein coding genes (PCG) before being translated into their respective amino acid sequences. Consequently, a globalpair MAFFT v7.505 [Katoh et al., 2002] alignment, suitable for closely related sequences of similar lengths was executed. The aligned amino acid sequences were back-translated before clustering and phylogenetic analysis could take place on the aligned nucleotide sequences. Both translation and back-translation were executed using the translate.py and backtranslate.py scripts from the TjCreedy biotools suite [Creedy, 2024].

A standard barcode region of the mitochondrial cytochrome C oxidase subunit I gene (COX1) 658bp long [Pentinsaari et al., 2016] was used for the computation of a distance matrix. This K2P distance matrix would consequently be used for both the simple kmeans and the adapted SigClust approaches. In this study Kimura's two parameter model (K2P) was used as it is a widely employed model of evolution and a standard for computing distance matrices [Collins et al., 2012].

3.2.1 Kimura's two parameter model

The K2P distance between two sequences is calculated as [Kimura, 1980]:

$$K2P = -0.5 \log((1 - 2P - Q)\sqrt{1 - 2Q}) \tag{1}$$

where P is the proportion of transitions and Q is the proportion of transversions between two sequences.

3.3 Clustering and Statistical Analyses

Both clustering approaches in this study require a distance matrix. This distance matrix was used in different stages of the two clustering methods experimented in this study. In the first simple kmeans approach, kmeans was executed on the entire distance matrix after it was scaled, treating columns as indepedent variables. The base R kmeans function output the contents of each cluster as well as coordinates of the cluster centers. The observations found to be closest to the central coordinates (by Euclidian distance) would qualify as representatives for the backbone reconstruction. This method was repeated for a different arbitrarily chosen number of clusters. In each case, a pairwise sub-matrix consisting of the central observations was made to consequently facilitate the comparison of K2P distributions of the cluster centers compared to the entire distance matrix.

The adapted SigClust algorithm also required a predefined number of clusters, in addition to a kmer length parameter. Multiple kmer length parameter values were chosen to assess the optimal value for this dataset - as is suggested by Chappell et al., (2017). Similar to the simple kmeans approach, the algorithm assigns each sequence to a particular cluster, of varying sizes. The observations in each cluster are then used to make multiple subsetted distance matrices to find the central observation in each cluster. The cluster centers are found to be the observations in the K2P distance matrix that have the smallest sum of Euclidean distances to all other observations in that cluster. The K2P similarity scores between cluster centers where again compared with the distribution of the entire distance matrix to observe the performance of the clustering algorithm.

This method performs a new technique for estimating the optimal kmer value. Firstly, the kmer value whose cluster centers distribution is most present at larger K2P values were deemed optimal. This indicates a set of distantly related sequences, ideal for backbone reconstruction that captures all the biodiversity in the order. Secondly, the optimal kmer values can be chosen via examining the

taxonomy of each cluster's sequences. The kmer value that's clusters were overall most taxonomically homogenous, and whose cluster centers were most diverse, is chosen for phylogenetic analyses. Taxonomically homogenous clusters are more likely to be monophyletic and therefore serve as good subtrees to observe alongside the cluster center backbone for additional detail.

Three different statistical tests were used to check for significant differences between the means of the subsetted cluster center's similarity scores, and the distance matrix. Firstly, F tests were carried out to check for equal variance between distributions before indicating the necessity for Welch's T tests to identify significant differences between means. Secondly, Cohen's d test was used to clarify the size and practicality of any significant differences [Aoki, 2020].

The overarching adapted SigClust method workflow described in this study is illustrated in a flow diagram in Figure 1.

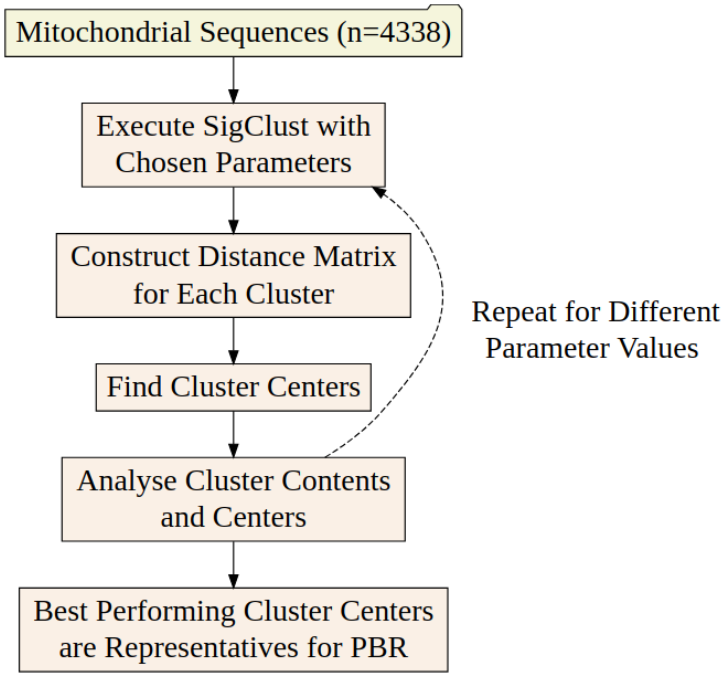


Figure 1: **Flow diagram demonstrating the adapted SigClust method before phylogenetic reconstruction**

3.4 Phylogenetic Analysis

Upon the completion of the clustering methods mentioned, the representatives for backbone reconstruction can be used to build the phylogeny. Phylogenetic trees were constructed using IQ-TREE (multicore version 2.0.7) [Nguyen et al., 2014], a software that utilises maximum-likelihood based phylogenetic inference. The -m MF+MERGE parameter was utilised for finding the optimal substitution model for each gene partition, and concatenating partitions that share the same model for efficient tree building. The molecular data was partitioned by gene using the partitioner.py script from the TjCreedy phylostuff suite [Creedy, 2024]. The representatives for the backbone tree were aligned with 120 sequences with a well established topology which was used as a tree constraint when executing IQ-TREE. Secondly, upon the construction of a backbone tree, taxonomically homogenous and monophyletic clusters can be grafted onto their respective tip on the backbone. These subtrees are reconstructed again using IQ-TREE with identical parameters, but without any constraint. The figures containing phylogenetic trees included in this write-up were formatted using the free online Interactive Tree of Life (iTOL) v6 software [Letunic and Bork, 2024].

4 Results

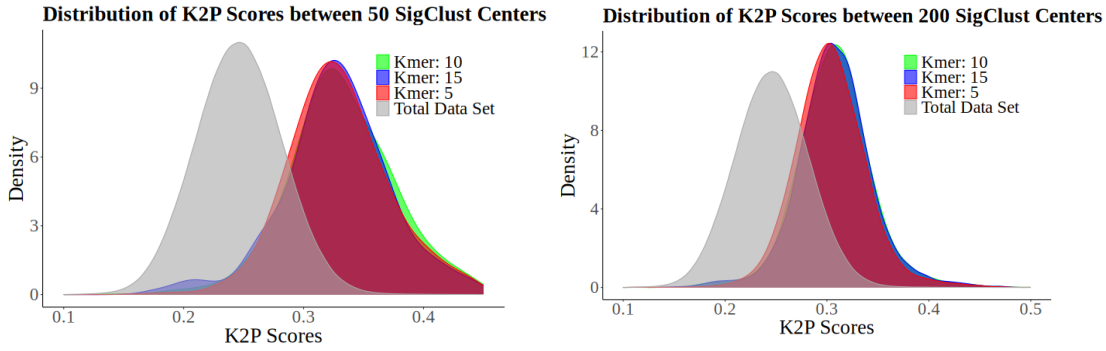
4.1 Simple Kmeans Approach

Executing the simple kmeans approach on the entire distance matrix, for a differing number of clusters, output a varying non-zero number of sequences in each cluster. The distributions for the 50 and 100 cluster centers are only marginally shifted towards larger K2P scores, illustrating that the kmeans clustering poorly identified representatives that were most dissimilar to each other. The 50 cluster centers had a mean pairwise K2P score of 0.258, while the 100 cluster center's mean of 0.252 was even closer to the mean of the entire distance matrix (0.246). F-test results show that the variances between the cluster centers and total dataset distributions are significantly different ($p < 2.2 \times 10^{-16}$) for both 50 and 100 clusters, thus justifying the use of a Welch's t-test for unequal variance. Whilst these

t-tests suggested significant differences between the means ($p < 2.2 \times 10^{-16}$) of the distributions for both cluster sets, a further Cohen's d test indicated small and negligible practical differences ($d = -0.318$, $d = -0.155$) between 50 and 100 clusters respectively and the distance matrix. This method showed a poor ability at finding representatives that had pairwise K2P scores at the larger end of the entire distance matrix. Therefore, it should not be developed further.

4.2 Adapted SigClust Algorithm

The SigClust algorithm executed on the COX1 gene output the members of each cluster. Subsetted distance matrices for each cluster were used to find the cluster centers to facilitate comparison in K2P score distributions. Figure 2 shows the distribution of K2P scores between the total dataset and the cluster centers; in addition, density curves of differing colours show the effect of changing the kmer value.



(a) 50 Cluster Centers

(b) 200 Cluster Centers

Figure 2: **Density plots showing the distribution of pairwise K2P scores for 50 (a) and 200 (b) SigClust cluster centers, relative to the entire distance matrix.**

The distribution of the 50 cluster centers (shown in Figure 2(a)) are clearly significantly shifted to greater K2P scores for every kmer value tested and plotted. Secondly, the adapted SigClust approach also evidently found cluster centers that were dissimilar from each other for 200 clusters (shown in Figure 2(b)), likely also capturing much of the diversity in the total sample. The mean of the 50

cluster center's pairwise K2P scores was 0.330, and the 200 cluster centers had a mean of 0.303. Both cases showing more of a shift to higher K2P values than the in the simple kmeans approach. Increasing the number of clusters intuitively reduces the likelihood of the average pairwise similarity score deviating far from the mean of the entire distance matrix. This is backed up by the two means and also additional Cohen's d scores. For all kmer values used for both sets of cluster numbers the Cohen's d scores suggest a large practical difference. Furthermore, using kmer value 5 for example, the 50 clusters gave a larger effect size than the 200 ($d = -2.314$, $d = -1.580$ respectively). Ultimately, Figure 2 shows that the adapted SigClust approach shows vast improvement than the simple kmeans approach, and will be further explored henceforth.

However, the density plots in Figure 2 clearly have not addressed the question on the best performing kmer value. Therefore, a line graph showing the mean number of taxa in each cluster against different kmer values addresses the ambiguity of the best performing kmer. Figure 3 shows a line graph for the mean number of superfamilies inside each outputted cluster.

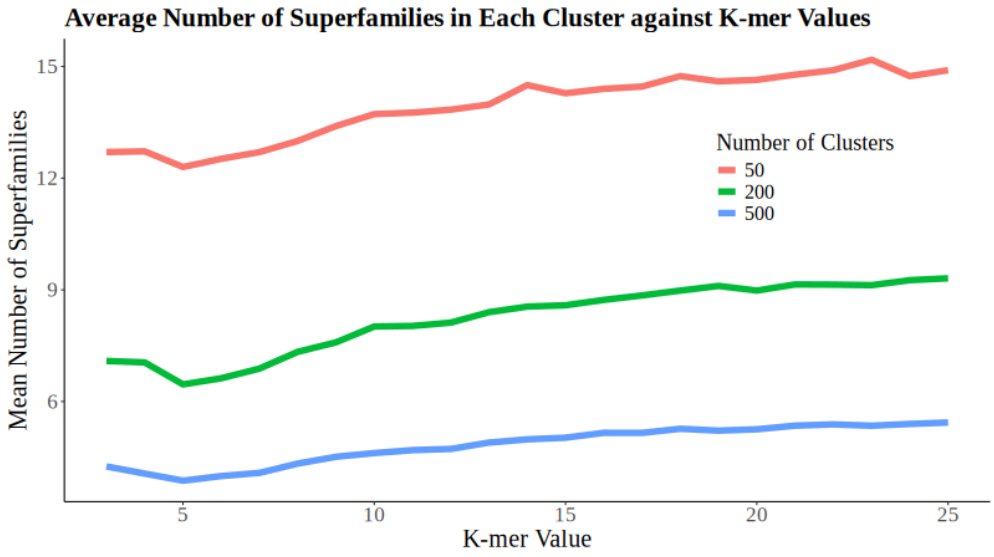
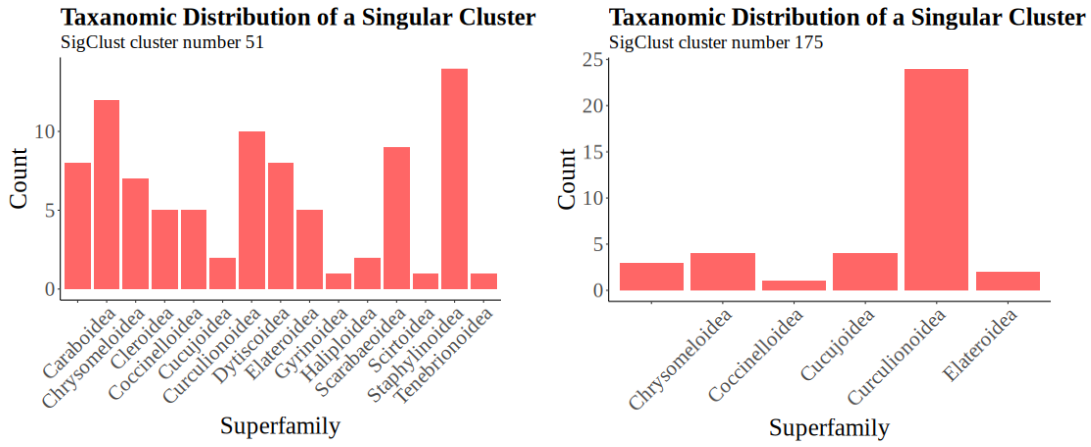


Figure 3: **Line graph showing the mean number of superfamilies present in each cluster for a differing number of clusters.**

Consistent for each different number of clusters chosen (represented by different coloured lines), kmer value 5 was best performing. At this value the mean number of different superfamilies represented in each cluster was lowest. This suggests a kmer value of 5 is best at identifying clusters that are most taxonomically homogenous at a superfamily level - indicating their potential for building monophyletic subtrees, post backbone reconstruction.

As the desirable kmer value has been established (5), the contents of the clusters themselves, computed using this parameter value can be scrutinized. Figure 4 shows case studies for the contents of different clusters with the SigClust parameters set to kmer value 5 with 200 clusters.



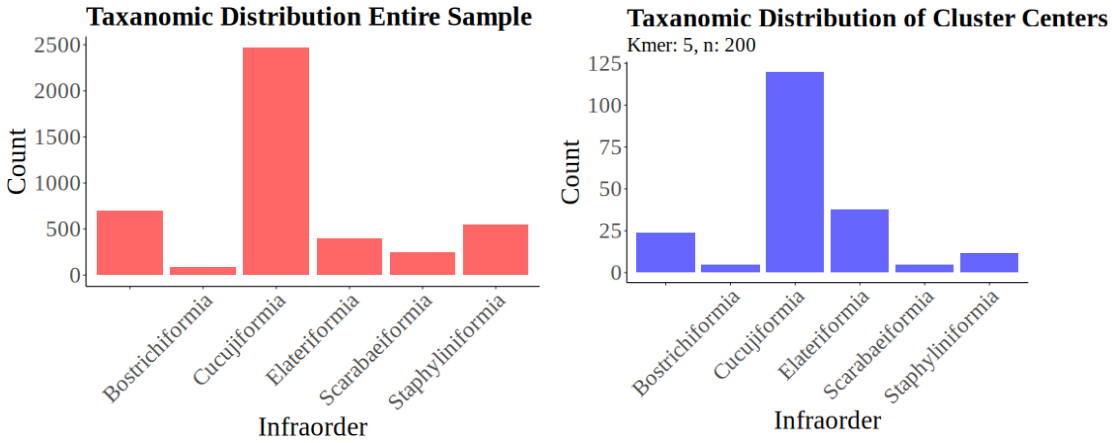
(a) Case study showing a cluster with diverse contents

(b) Case study showing a cluster with a clearly dominant taxa present

Figure 4: **Bar graphs showing the different taxonomic distributions of case study SigClust clusters.**

Figure 4 illustrates case studies for the typical contents of a SigClust cluster. Figure 4(a) shows a large cluster consisting of a broad sample of the overall biodiversity, illustrating a poor SigClust performance. Whereas, Figure 6(b) shows a better performing cluster consisting of a clearly dominant taxonomy (superfamily Curculionoidea). In addition, 41 out of the 200 clusters (20.5%) performed perfectly to a superfamily level consisting of sequences belonging to only one superfamily. Next, the cluster center's taxonomic distribution can be observed and

is shown in figure 5.



(a) Entire sample (b) Cluster centers

Figure 5: **Bar graphs showing the taxonomic distributions for the total sample (a) and the cluster centers (b)**

Figure 5 illustrates that the cluster centers strongly replicate the taxonomic distribution of the sample - suggesting a roughly uniform and desired selection. These cluster centers, alongside sequences of the constrained topology were used to build the backbone phylogenetic tree.

4.3 Phylogenetic Reconstruction

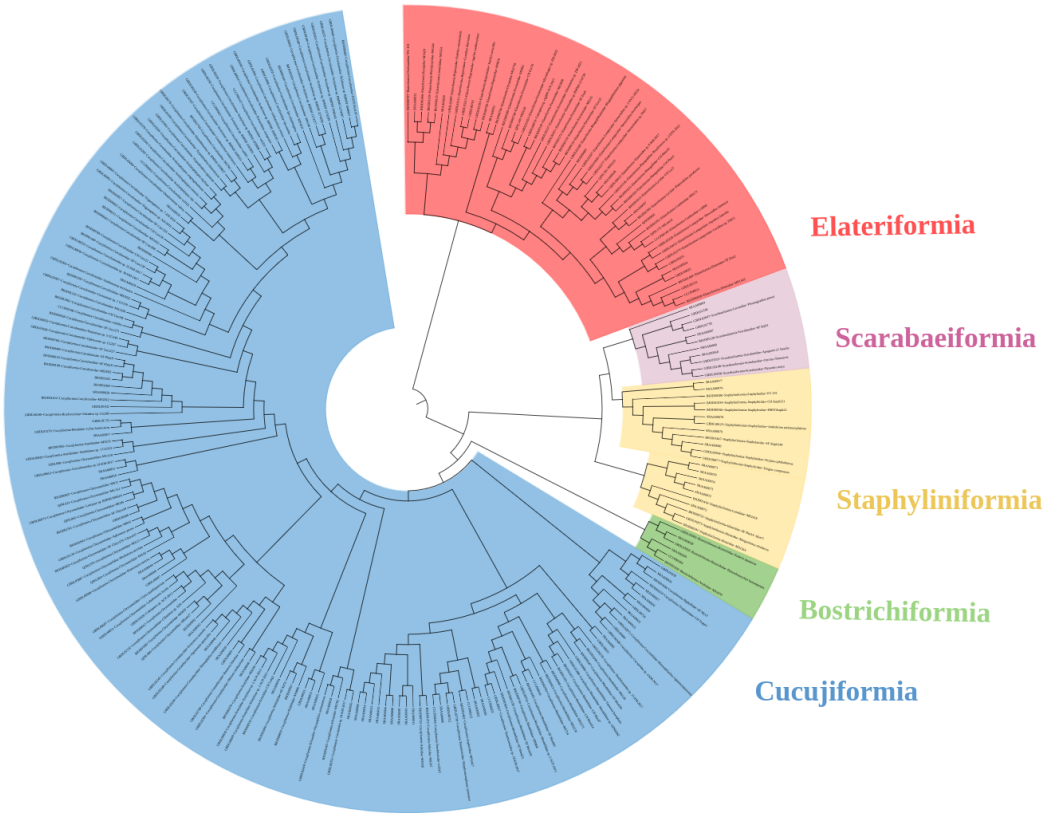


Figure 6: **Phylogenetic cladogram illustrating the relatedness between Coleopteran infraorders.**

Figure 6 is a formatted phylogenetic backbone tree illustrating the diverse Coleopteran phylogeny highlighting the present infraorders. The taxonomic distribution of the original 4438 sequences is further showed in this tree. The 200 taxonomically labelled tips make up the cluster center representatives, whilst 120 non-labelled tips in Figure 6 are those of the constrained tree used to increase the accuracy of the topology found in this study. The original tree file including the family and genus of all species tips in this figure are available and accessible following the guidance in the code availability statement concluding this study. Once the backbone was constructed a strong performing taxonomically homogenous cluster is chosen as a case study for demonstrating how the method in this study can be

performed.

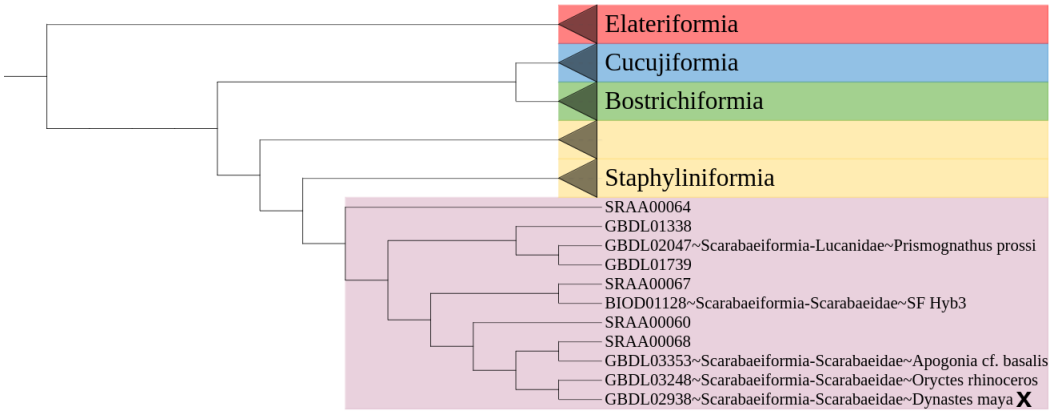


Figure 7: **A partially collapsed tree highlighting a particular cluster center (Dynastes maya), marked with the X.**

Figure 7 shows an identical tree to Figure 6 with certain condensed infraorders to highlight the chosen cluster center - marked with an X. *Dynastes* maya happened to be the cluster center for an identified taxonomically homogenous cluster. A separate phylogenetic subtree is consequently built with the sequences in this cluster for refining and adding detail to the backbone.

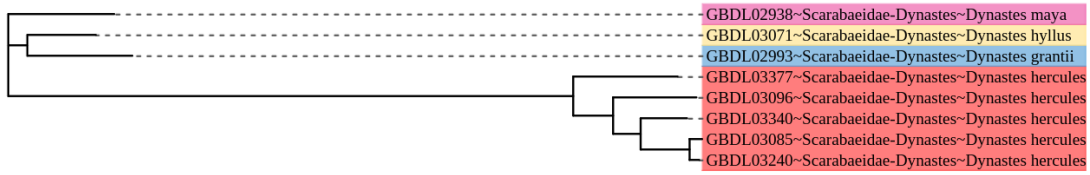


Figure 8: **Phylogram showing the evolutionary relationship between species of the *Dynastes* genus present in the sample.**

Figure 8 shows an example homogenous and monophyletic subtree that adds detail to its respective cluster on the backbone. It provides phylogenetic inference to many more species that were not originally used in the process of already challenging backbone reconstruction. Secondly however, it also highlights the shortcomings of the process used to identify cluster centers. This is because *Dynastes* maya is clearly not the most similar to the majority of sequences in this

cluster. This is illustrated in Figure 8, as *Dynastes* maya is the most distantly related sequence to all others in its cluster, despite it being found to be the cluster center.

5 Discussion

The results in this study illustrate the strengths and weaknesses of kmeans clustering in phylogenetic analyses. In addition, poor performing elements in this study may point to limitations that can be addressed in future studies. Firstly, whilst the cluster centers from the simple kmeans approach showed a pairwise K2P distribution significantly different to the total distance matrix; it is clear by the Cohen's d scores, and in comparison to consequent SigClust methods that they did not adequately capture the broad diversity of the dataset. Furthermore, the decrease in performance upon the increase in the number of clusters showed attempting this approach with more clusters would not be appropriate. This is because increasing the number of clusters creates more similarity between cluster centers as smaller clusters represent more localized portions of the data - leading to an overall increase in similarity between clusters. Faults in this simple kmeans technique must lie within either the distance matrix or the execution of kmeans clustering itself. The results from the simple kmeans approach further validate kmeans' struggle with distance matrices. It also hints at other clustering methods to be attempted on the K2P pairwise matrix. For example, another clustering method, partitioning around medoids (PAM) is a similar technique to kmeans, proven to work well with gene expression data, and is better designed for distance matrices [Huang and Pan, 2006].

The pairwise distributions of the cluster centers from the adapted SigClust method convincingly outperformed, thus making redundant, the simple kmeans approach. It is clear from Figure 2 that the SigClust cluster centers for all kmer values captured similar sequence similarity scores at the high extreme of the total distance matrix distribution. Thus suggesting that the cluster centers may represent sequences that are most dissimilar from each other and therefore capture the broad

diversity of the dataset. Another positive from the adapted SigClust method is that whilst it still requires a distance matrix, computational demands are still reduced in comparison to other clustering methods, as only subsetted matrices are required (for cluster center calculation) rather than a larger distance matrix consisting of all pairwise similarity scores in the sample.

Whilst it is clear that dissimilar sequences have been selected as desired, more nuanced analysis including the taxonomy of the outputted sequences gives a greater understanding on how the clustering is performing with different parameter values. Clusters composed of sequences with minimal differences in their taxonomy are preferred. Therefore, Figure 4 illustrates how the number of different taxa (superfamilies) in each cluster changes as the kmer value increases. It shows the effective 'error' of the clustering against different kmer values. The graph therefore shows that kmer value 5 should be the set parameter for any following analysis for all number of clusters. Although Chappell et al., (2017) noted that different kmer values are suitable for different sets of sequences, this result was anticipated, as a kmer value of 5 is the default parameter of the algorithm and performed best in their study [Chappell et al., 2017]. The contents of example clusters are illustrated as varied in Figure 4. Whilst a taxonomically homogenous set of clusters would have been favourable, it is the accurate identification of cluster centers that is integral to the backbone reconstruction. However, the addition of subtrees onto the backbone would be a struggle without taxonomically homogenous clusters; therefore, some clusters are far better for tip adding than others.

The SigClust algorithm output clusters with taxonomically curious contents, with only 20.5% being homogenous to a superfamily level. Furthermore, although the cluster center's taxonomic distribution is similar to the overall samples', the adapted SigClust approach struggled to identify cluster centers that were part of the dominant taxonomy in their cluster. Whilst the approach used to find cluster centers was not effective at selecting the dominant superfamily in the cluster, its consistency ultimately lead to the cluster centers still strongly representing the overall taxonomic distribution in the original sample - shown in Figure 5 and 6. Overall then, the experiment identified representatives capturing a similar

taxonomic distribution of the infraorders than in the total sample. This illustrates that representatives were relatively uniformly picked from the sample as intended and that kmeans has potential to be used for selecting backbone representatives in future studies.

However, the SigClust algorithm did not perform as well as intended regarding the contents of each cluster, relative to the original study. The online address of the sequences used in the original study was declared, but has since been changed, thus there is no indication on the taxonomy or relatedness between sequences or groups of sequences used in the development of the SigClust method. However, the method was initially designed for environmental samples with a likely highly diverse range of taxa, which was probably reflected in the input sequences used during its development. One can theorise that the clustering has performed poorly in this study for a couple of reasons. Firstly, the mitogenome sequences used in this study may generally be all too similar to each other. The conversion of the sequences into kmers and their transformation into a binary format may have lost too much detail for their consequent assignment into accurate clusters [Chappell et al., 2017]. Another explanation may be that there is enough diversisty within Coleoptera however there is a lack of distinct groups, synonymous with the downfall in other clustering studies. When a diverse set of sequences within one order is used, the diversity among the sample exists on a gradient rather than in distinct groups, as is often seen with arbitrary taxonomic labels. SigClust would likely have perfomed better finding clusters from a diverse sample where there is greater discreteness between taxa in the sample. Furthermore, extreme differences between molecular mutation and speciation rates among different lineages suggests validating the clusters taxonomically may not have been suitable to begin with.

In addition, the method to identify cluster centers is likely flawed as one assumes the most central observation of a cluster, from its respective distance matrix, would be of the most dominant taxa in that cluster. This was often not the case highlighting the necessity for an improved method for identifying cluster centers. It seems more sophisticated care is required when choosing how to com-

pute this distance matrix. Firstly, K2P is a strongly backed model, however, it's execution on only the first couple of codon positions may have been more suitable. This is because the third position has a tendency to saturate from massive divergence that renders it useless, and often detrimental to phylogenetic analyses [Orlov et al., 2021]. Saturation is an important factor in phylogenetic analysis and is known to be particularly prevalent in the COX1 gene [Wang and Lv, 2014]. Saturation is a problem as it leads to long-branch attraction due to sequences' evolutionary convergent biases [Yuan et al., 2022]. Extra care could have been taken by testing for saturation [Yuan et al., 2016] before deciding if any codon positions should be removed before phylogenetic analysis in this study. This potential bias may even have been amplified by the use of only one mitochondrial PCG (COX1) in the clustering stage of this study. The use of only a single gene for the distance matrix suggests the study could have employed software for finding the optimal model of evolution before computing the distance matrix. Though K2P is often appropriate, other models of evolution may have been more suited to the COX1 genes in the dataset. Furthermore, alternative resolutions to this issue may include executing the SigClust algorithm on the entire mitogenomes, or their concatenated PCGs, rather than just a single gene.

In hindsight, another potential limitation of the study was incorrectly prioritising the kmer value parameter over the number of clusters parameter. The cluster contents for a changing number of clusters were analysed for comparison; however, ultimately 200 clusters was an arbitrarily chosen parameter value predicted to be suitable for the backbone reconstruction task. More clusters would have resulted in a more diverse backbone, but at a computational cost. Furthermore, increasing the number of clusters also increases the likelihood of overfitting. A too great number of clusters cause the algorithm to form clusters encapsulating the noise in the data, rather than the underlying patterns. An improvement in future studies would be to perform the 'elbow method' for finding the optimal number of kmeans clusters [Saadeh et al., 2020]. This would address the trade-off between the amount of error (mean number of different taxa) in each cluster and the overall number of clusters, helping to identify a parsimonious balance that maximizes the accuracy of the model whilst minimising the likelihood of overfitting.

6 Conclusion

In conclusion, the results in this exploratory investigation show the potential for using kmeans clustering for identifying representatives for backbone reconstruction. The computation of cluster centers that have an associated monophyletic or taxonomically homogenous cluster, is a novel idea that should be explored further in future studies that aim to solve the challenge of large phylogenetic tree building. Whilst there were few taxonomically pure clusters, this study still showed an example subtree, whose phylogeny could be more broadly observed alongside the backbone. Furthermore, the method in this study is a flexible and efficient solution to building large trees. This is because extra detail may be added via subtrees wherever is desired, instead of building all the subtrees - thus saving computational time.

7 Data and Code Availability

All code written and information on the packages used for the analysis and write-up in this thesis can be found in the following GitHub repo:
https://github.com/SamSmithImperial/CMEECourseWork/Project

More information regarding the Site 100 database used in this study can be found at: www.site100.org

8 Acknowledgements

I would firstly like to thank Professor Alfried Vogler and his lab group for welcoming me into the Natural History Museum and guiding me through this MSc thesis. Thank you for the opportunity to work in this team and contribute to weekly lab meetings. Secondly, I would like to thank Professor Samraat Pawar and Professor James Rosindell for co-organising, delivering and supporting me through this MSc in Computational Methods in Ecology and Evolution.

References

[Aoki, 2020] Aoki, S. (2020). Effect sizes of the differences between means without assuming variance equality and between a mean and a constant. *Heliyon*, 6(1):e03306.

[Bian et al., 2022] Bian, X., Garner, B. H., Liu, H., and Vogler, A. P. (2022). The site-100 project: Site-based biodiversity genomics for species discovery, community ecology, and a global tree-of-life. *Frontiers in Ecology and Evolution*, 10.

[Chappell et al., 2017] Chappell, T., Geva, S., and Hogan, J. (2017). K-means clustering of biological sequences. In *Proceedings of the 22nd Australasian Document Computing Symposium*, ADCS 2017. ACM.

[Collins et al., 2012] Collins, R. A., Boykin, L. M., Cruickshank, R. H., and Armstrong, K. F. (2012). Barcoding's next top model: an evaluation of nucleotide substitution models for specimen identification. *Methods in Ecology and Evolution*, 3(3):457–465.

[Creedy, 2024] Creedy, T. (2024). biotools and phylostuff. *https://github.com/tjcreedy/Biotools*.

[Creedy et al., 2024] Creedy, T. J., Ding, Y., Gregory, K. M., Swaby, L., Zhang, F., and Vogler, A. P. (2024). A nuclear-mitochondrial phylogeny of coleoptera. Unpublished manuscript.

[Davis et al., 2010] Davis, C. C., Willis, C. G., Primack, R. B., and Miller-Rushing, A. J. (2010). The importance of phylogeny to the study of phenological response to global climate change. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1555):3201–3213.

[Foundation, 2024] Foundation, P. S. (2024). Python (version 3.12.3). Programming language.

[Gressitt, 2024] Gressitt, J. L. (2024). Coleopteran. *Encyclopedia Britannica*. Accessed: 2024-08-21.

[Huang and Pan, 2006] Huang, D. and Pan, W. (2006). Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. *Bioinformatics*, 22(10):1259–1268.

[Katoh et al., 2002] Katoh, K., Misawa, K., Kuma, K., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, 30(14):3059–3066.

[Kimura, 1980] Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16(2):111–120.

[Kupczok et al., 2010] Kupczok, A., Schmidt, H. A., and von Haeseler, A. (2010). Accuracy of phylogeny reconstruction methods combining overlapping gene data sets. *Algorithms for Molecular Biology*, 5(1).

[Letunic and Bork, 2024] Letunic, I. and Bork, P. (2024). Interactive tree of life (itol) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research*, 52(W1):W78–W82.

[Lu et al., 2004] Lu, Y., Lu, S., Fotouhi, F., Deng, Y., and Brown, S. J. (2004). Incremental genetic k-means algorithm and its application in gene expression data analysis. *BMC Bioinformatics*, 5(1).

[Nabhan and Sarkar, 2011] Nabhan, A. R. and Sarkar, I. N. (2011). The impact of taxon sampling on phylogenetic inference: a review of two decades of controversy. *Briefings in Bioinformatics*, 13(1):122–134.

[Nguyen et al., 2014] Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., and Minh, B. Q. (2014). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, 32(1):268–274.

[Orlov et al., 2021] Orlov, I., Leschen, R. A., Żyła, D., and Solodovnikov, A. (2021). Total-evidence backbone phylogeny of aleocharinae (coleoptera: Staphylinidae). *Cladistics*, 37(4):343–374.

[Pentinsaari et al., 2016] Pentinsaari, M., Salmela, H., Mutanen, M., and Roslin, T. (2016). Molecular evolution of a widely-adopted taxonomic marker (coi) across the animal tree of life. *Scientific Reports*, 6(1).

[Pisani and Wilkinson, 2002] Pisani, D. and Wilkinson, M. (2002). Matrix representation with parsimony, taxonomic congruence, and total evidence. *Systematic Biology*, 51(1):151–155.

[R Core Team, 2024] R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[Saadeh et al., 2020] Saadeh, H., Fayez, R. Q. A., and Elshqeirat, B. (2020). Application of k-means clustering to identify similar gene expression patterns during erythroid development. *International Journal of Machine Learning and Computing*, 10(3):452–457.

[Sanderson and Driskell, 2003] Sanderson, M. J. and Driskell, A. C. (2003). The challenge of constructing large phylogenetic trees. *Trends in Plant Science*, 8(8):374–379.

[Tseng, 2007] Tseng, G. (2007). Penalized and weighted K-means for clustering with scattered objects and prior information in high-throughput biological data. *Bioinformatics*, 23(17):2247–2255.

[Wang and Lv, 2014] Wang, M. and Lv, S. (2014). [substitution saturation analysis of mitochondrial cytochrome c oxidase subunit 1 (cox1) gene of angiostrongylus cantonensis]. *Chinese journal of parasitology and parasitic diseases*, 32:205–9.

[Xu et al., 2022] Xu, X., Li, X., and Wang, D. (2022). New insights into the backbone phylogeny and character evolution of corydalis (papaveraceae) based on plastome data. *Frontiers in Plant Science*, 13.

[Yu et al., 2022] Yu, X., Yang, H., Liu, J., Qi, Y., Sun, L., and Tian, X. (2022). A strategy for a high enrichment of insect mitochondrial dna for mitogenomic analysis. *Gene*, 808:145986.

[Yuan et al., 2022] Yuan, L., Liu, H., Ge, X., Yang, G., Xie, G., and Yang, Y. (2022). A mitochondrial genome phylogeny of cleridae (coleoptera, cleroidea). *Insects*, 13(2):118.

[Yuan et al., 2016] Yuan, M.-L., Zhang, Q.-L., Zhang, L., Guo, Z.-L., Liu, Y.-J., Shen, Y.-Y., and Shao, R. (2016). High-level phylogeny of the coleoptera inferred with mitochondrial genome sequences. *Molecular Phylogenetics and Evolution*, 104:99–111.

[Zhang et al., 2018] Zhang, S.-Q., Che, L.-H., Li, Y., Liang, D., Pang, H., Ślipiński, A., and Zhang, P. (2018). Evolutionary history of coleoptera revealed by extensive sampling of genes and species. *Nature Communications*, 9(1).

End of Document