Using K-Means Clustering to Investigate Phylogenetic Relationships in Modern Cnidarians

AOS C204 Final Project

Kira Fish

Abstract:

Element-to-calcium (X/Ca) ratios, measured in marine organisms such as foraminifera, coral, and mollusks, are frequently used as geochemical proxies for paleoclimate reconstructions. Whereas X/Ca ratios in abiotic calcium carbonate reflect the abundance of those elements in ambient seawater (as well as water temperature), X/Ca ratios in biominerals are often distinct from the surrounding seawater. These so-called “vital effects” describe the biological control that calcifying organisms exhibit over their internal environment during the calcification process (Blamart et al., 2007; Cohen and Gaetani, 2011). While vital effects can elucidate interesting properties of the biomineralization process, they often obstruct the use of X/Ca ratios as paleoclimate proxies. Some studies have attempted to circumvent this issue by quantifying species-specific calibrations between X/Ca ratios and the environmental variable in question, however this can only be accomplished if the species in question is extant (Trotter et al., 2011). The fact that these calibrations are species-specific implies that phylogeny impacts X/Ca ratios and, by extension, suggests that the trace elements incorporated into biominerals can be correlated with an organism’s taxonomy and evolutionary history (Edgar et al., 2017). The present study aims to examine a suite of X/Ca ratios in extant cnidaria in order to test the hypothesis that X/Ca ratios in biomineralizers are linked to an organism’s evolutionary history.

Introduction:

Marine biomineralizers utilize a variety of minerals to build their skeletons. For example, many foraminifera and radiolarians have silica-based skeletons, whereas corals more commonly build their skeletons using the polymorphs of calcium carbonate: aragonite and calcite. There is ample evidence suggesting that the evolution of skeletal mineralogy was heavily influenced by shifting global seawater composition, specifically the shifting between calcite and aragonite seas in the Ediacaran-Cambrian periods (Porter, 2007; Gilbert et al., 2022). But even as bulk ocean composition transformed over time, skeletal mineralogy of various phyla did not evolve to reflect the new thermodynamically preferential polymorph, demonstrating what some have called ‘deep resilience’ (Gold and Vermeij, 2023).

Deep resilience is thought to be influenced by an organism’s energy budget, or how much energy that organism must expend in order to perform various functions required for survival (Vermeij, 2020). Changing ocean conditions impact the energy budget required for survival, and it has been suggested that species with enhanced resilience may be able to more easily offset energetic costs associated with changing conditions such as acidification (Gold and Vermeij, 2023). It stands to reason, then, that different groupings of organisms could have evolved calcification mechanisms that compliment the unique ocean conditions at the time of their evolution. This concept was examined by Edgar et al. (2017), who analyzed planktonic foraminiferal δ^{13}C and δ^{18}O values spanning 107 million years. These authors demonstrated that the species-specific vital effects of δ^{13}C are impacted by evolutionary history of the macroperforate foraminiferal species of the Cenozoic. It seems probable, then, that vital effects of trace element ratios could be a result of residual evolutionary pressures that impacted an organism's biomineralization mechanisms at the time of evolution (Edgar et al., 2017). In order to robustly test this hypothesis, we must determine the degree to which we can link skeletal trace element composition to taxonomy in modern taxa.

It has been demonstrated that X/Ca ratios can be linked to taxa at the phylum level (Ulrich et al., 2021), but it is unknown if finer scale taxonomic resolution can be achieved. Ultimately, we aim to utilize a large geochemical dataset with 8-10 X/Ca ratios measured in 12 species of modern cnidarians spanning 3 coral types - scleractinians, octocorals, and hydrocorals. While the preparation of that full dataset is underway, we can utilize an unsupervised machine learning approach (*KMeans)* on a smaller set of coralline geochemical data to explore if the data are clustering according to phylogeny within one coral type – scleractinians. This course project serves as a small part of what will hopefully later be developed into a supervised learning classification problem, where we will see if geochemical composition of coral skeletons can be used as a predictive tool of phylogeny.

The data used here was collected several years ago by some members of the Eagle-Tripati lab, namely Ilian DeCorte and Maxence Guillermic, and was measured using inductively coupled plasma optical emission spectroscopy (ICP-OES). The dataset includes six coral species that can be divided into in five separate families representing 2 evolutionary "clades" (note that, while Siderastreidae is known to have representatives in both clades, the species *Siderastrea siderea* in this dataset is considered to be Complex). Ten X/Ca ratios are represented (Li/Ca, B/Ca, Mg/Ca, Sr/Ca, U/Ca, Cd/Ca, Ba/Ca, Na/Ca, Mn/Ca, and Al/Ca).

Table 1. Summary of dataset taxa

Coral Type	Clade	Family	*Species*
Scleractinian	Robust	Pocilloporidae	*P. damicornis, S. pistillata*
		Mussidae	*Ps. strigosa*
	Complex	Poritidae	*Po. astreoides*
		Agariciidae	*U. tenuifolia*
		Siderastreidae	*S. siderea*

Modeling:

Dataset Preprocessing

The dataset had a total of 279 NaN values, 256 of which came from the Al/Ca, Mn/Ca, and Na/Ca features. The large swaths of missing values for these features was likely due to machine error, which is common in geochemical analyses. For this reason, the ratios Al/Ca, Mn/Ca, and Na/Ca were dropped from the analysis, leaving a total of 7 features and 23 NaN values. In order to avoid losing any more statistical power, I chose to estimate the missing values using the K-Nearest Neighbors Imputation from *sklearn.impute* with n = 3.

I first ran through the entire experiment without removing any outliers, but it quickly became clear that outliers were severely impacting the clustering algorithm (see supplemental figures). I then identified outliers using z-scores. When outliers were removed from the dataset, the number of samples was reduced from 236 to 215, leaving us with 1,505 total values in the dataset.

Because the elements are incorporated into the coral skeletons in such drastically different quantities (ranging from 0.01 µmol to 10,000 µmol), it was necessary to normalize the data. I chose to test 3 different scalers from *sklearn.preprocessing* - *MinMaxScaler, StandardScaler,* and *RobustScaler.* I ultimately chose to use the *MinMaxScaler* due to it producing the highest silhouette score upon analysis. The silhouette scores of the other scalers can be viewed in the supplemental figures.

KMeans Clustering, Elbow Method, and Silhouette Scores

Next, I explored the data using the unsupervised clustering algorithm *KMeans*. In order to determine the ideal number of clusters for each scaler, I employed both the Elbow Method and Silhouette Scores. Finally, in order to evaluate whether or not the *KMeans* algorithm was clustering as we would expect, I calculated the percentage of clades represented in each cluster.

Results and Discussion:

It was initially assumed that opting for k=2 clusters (Figure 1) would result in a satisfactory visualization, given that the dataset encapsulates two evolutionary clades – complex and robust. It is important to note that, for this dataset, Feature1 and Feature2 do not correspond to any specific elemental ratio, as we are condensing 7 dimensions into 2. Nonetheless, the data exhibits relatively cohesive clustering considering the complexity of the dataset.

In order to investigate whether k=2 is the optimal number of clusters, I utilized the Elbow Method and examined the Silhouette Scores for the different number of clusters. The elbow method suggested that the ideal number of clusters might be 2 or 3, as the distortion exhibited the most significant decline between these values (Figure 2). Acknowledging the somewhat heuristic nature of the elbow method, I opted to calculate Silhouette Scores for a more quantitative insight (Table 2, Figure 3).

The silhouette scores indicate three clusters as optimal (Table 2 and Figure 3). However, when *KMeans* was re-run with k=3 and the clusters plotted into Feature1 and Feature2 space (Figure 4), we can see that the centroids in the purple and yellow clusters are in close proximity. This may indicate that k=3 is not the correct number of clusters for this dataset, or that these two centroids are further apart in different dimensions, and could be better represented by a 3D plot. Due to the inability of *KMeans* to visualize in 3D space, and based on what we know biologically about the existence of two evolutionary clades, subsequent analyses were performed with k=2 clusters despite the slightly lower silhouette score.

The process of evaluating whether the KMeans algorithm is clustering data randomly or as expected involves a detailed examination of the distribution of biological groups (Clades and Families) within the identified clusters. This analysis aims to provide insights into the algorithm's performance and its ability to segregate data based on inherent patterns. The percent representation for each Clade within their respective clusters (Figure 5) serves as a valuable metric for assessing the clustering outcome. It is essential to note that, due to the non-deterministic nature of *KMeans*, these percentages may slightly vary with each run. As a result, these percentages should be interpreted with caution and not treated as absolute, but rather as indicative of the general trends in the clustering behavior.

Nonetheless, the results are promising: Cluster 0 is composed of 83.8% Complex corals and 16.2% Robust corals, while Cluster 1 is characterized by 61.7% Robust corals and 38.3% Complex corals. To delve deeper into the family-level distinctions within the clusters (Figure 6), it becomes apparent that Pocilloporidae is exclusively represented in Cluster 0, while Agariciidae is solely present in Cluster 1. These clade-and-family-specific clustering results reinforce the idea that the algorithm is not randomly grouping specimens but is, in fact, capturing meaningful biological distinctions.

Conclusions and Future Directions:

These results underscore previous findings that phylogenetic signals exist in X/Ca ratios of cultured carbonates (Ulrich et al., 2021), and if the findings of this project prove consistent across larger datasets with both cultured and non-cultured specimens, X/Ca ratios may emerge as a link between carbonate skeletal composition and phylogeny. This suggests that biominerals carry a distinctive chemical signature reflective of their evolutionary context, potentially enabling us to position them accurately in the evolutionary timeline. For the next leg of this project, I plan to perform a PCA in order to reduce the dimensions represented in the dataset, then re-perform *KMeans* and assess the results. Visualization in 3D space may also be a valuable way to assess this data. Additionally, I plan to eventually try out several different classifiers, such as *AdaBoost, RandomForest,* and *SupportVectorClassifier* in order to see if geochemistry of carbonates could be used as a predictive tool.

Figures:

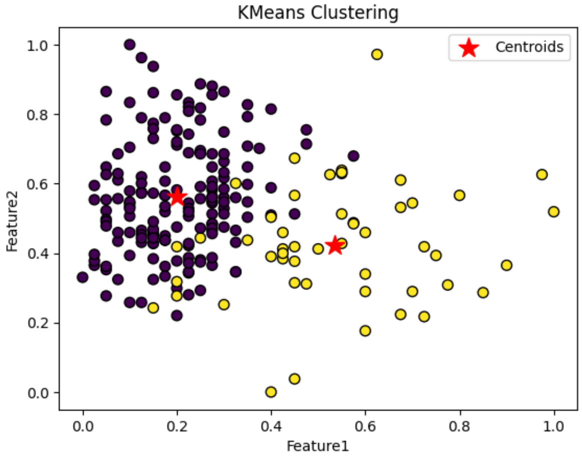


Figure 1: *KMeans* clustering with n=2 clusters

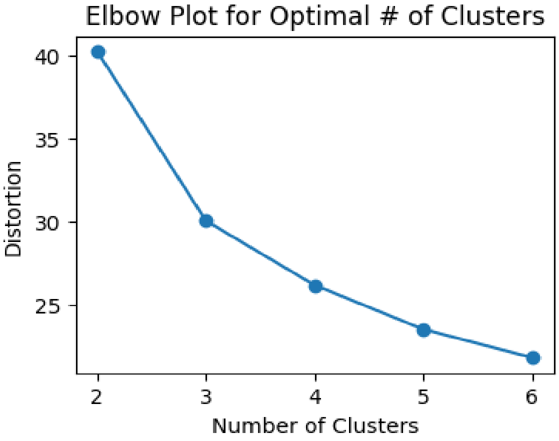


Figure 2: Elbow Plot showing distortion compared to number of clusters

Number of Clusters	Silhouette Score
2	0.25
3	0.30
4	0.19
5	0.23
6	0.19

Table 2: Number of clusters and Silhouette score

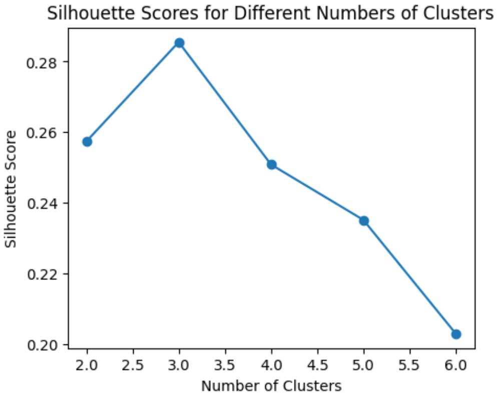


Figure 3: Silhouette scores across numbers of clusters

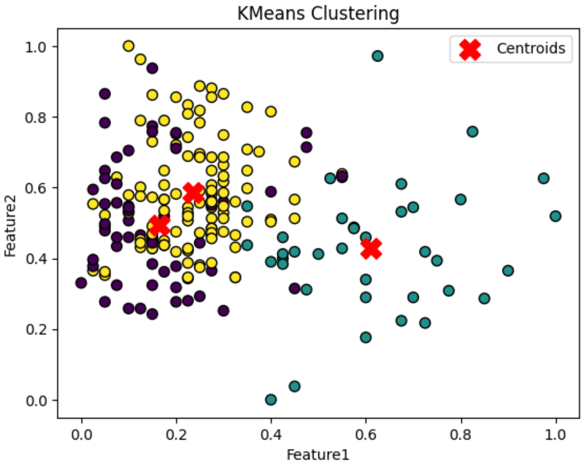


Figure 4: KMeans clustering with n=3

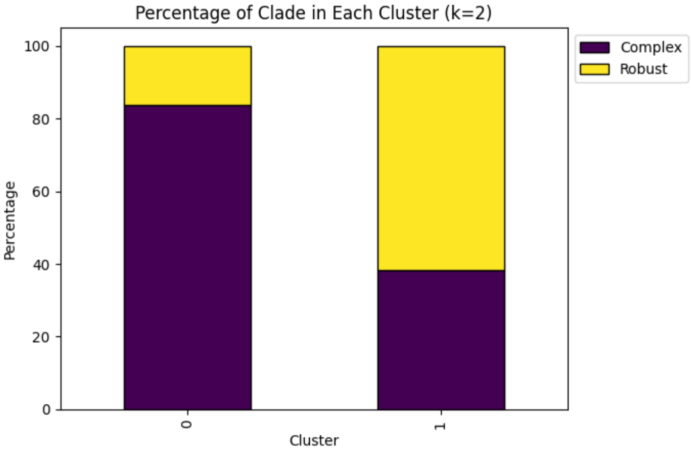


Figure 5: Percent representation of Complex or Robust clades in clusters

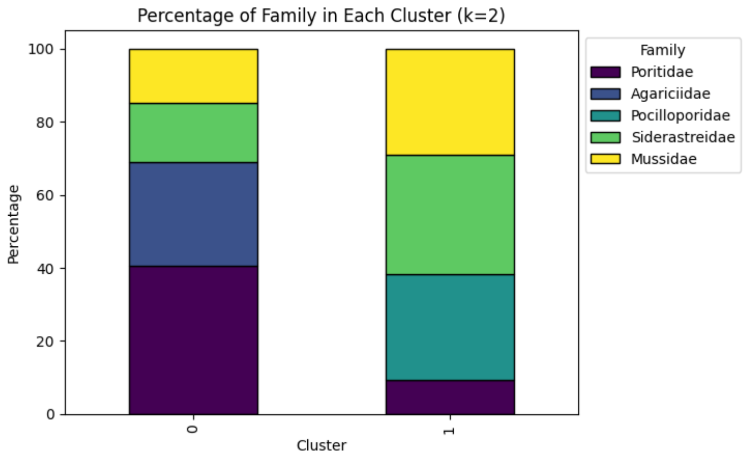


Figure 6: Percent representation of phylogenetic Families in each Cluster

Sources:

Blamart, D., Rollion-Bard, C., Meibom, A., Cuif, J. P., Juillet-Leclerc, A. and Dauphin, Y.: Correlation of boron isotopic composition with ultrastructure in the deep-sea coral *Lophelia pertusa* : Implications for biomineralization and paleo-pH, Geochem. Geophys. Geosyst., 8(12), 2007.

Cohen, A. L. and Gaetani, G. A.: Ion partitioning and the geochemistry of coral skeletons: Solving the mystery of the vital effect, in Ion Partitioning in Ambient-Temperature Aqueous Systems, edited by G. Ferraris, M. Prieto, and H. Stoll, pp. 377–397, Mineralogical Society of Great Britain & Ireland, London, 2011.

Edgar, K. M., Hull, P. M. and Ezard, T. H. G.: Evolutionary history biases inferences of ecology and environment from δ13C but not δ18O values., Nat. Commun., 8(1), 1106, 2017.

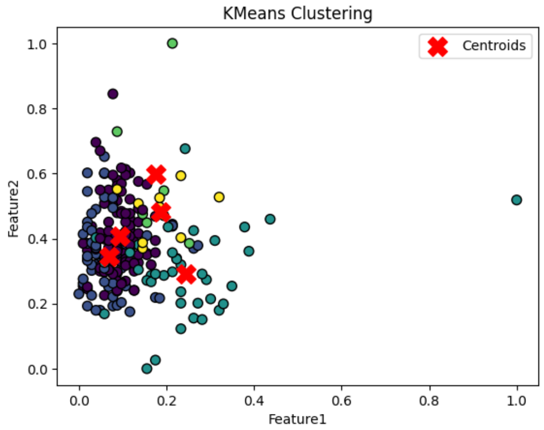
Gilbert, P. U. P. A., Bergmann, K. D., Boekelheide, N., Tambutté, S., Mass, T., Marin, F., Adkins, J. F., Erez, J., Gilbert, B., Knutson, V., Cantine, M., Hernández, J. O. and Knoll, A. H.: Biomineralization: Integrating mechanism and evolutionary history., Sci. Adv., 8(10), eabl9653, 2022.

Gold, D. A. and Vermeij, G. J.: Deep resilience: An evolutionary perspective on calcification in an age of ocean acidification., Front. Physiol., 14, 1092321, 2023.

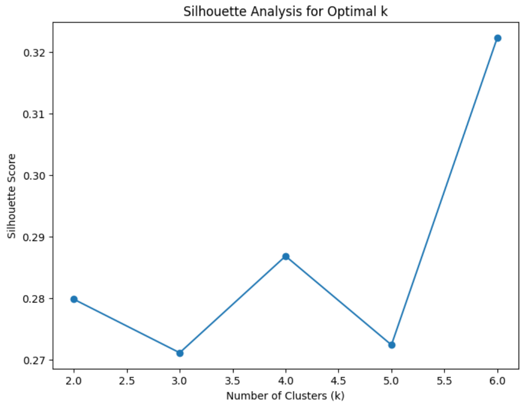
Porter, S. M.: Seawater chemistry and early carbonate biomineralization, Science, 316(5829), 1302, 2007**.**

Ulrich, R. N., Guillermic, M., Campbell, J., Hakim, A., Han, R., Singh, S., Stewart, J. D., Román-Palacios, C., Carroll, H. M., De Corte, I., Gilmore, R. E., Doss, W., Tripati, A., Ries, J. B. and Eagle, R. A.: Patterns of element incorporation in calcium carbonate biominerals recapitulate phylogeny for a diverse range of marine calcifiers., Front. Earth Sci., 9, 2021.

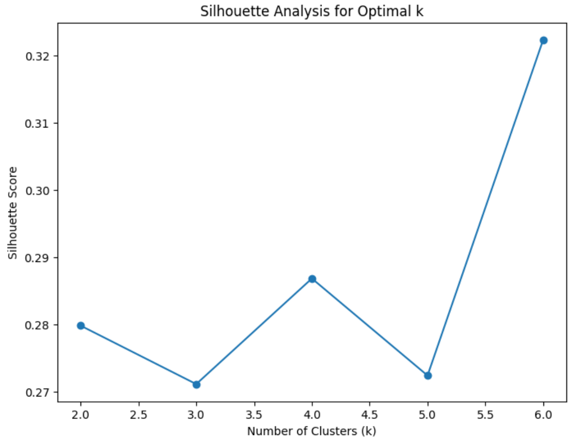
Supplemental Figures:



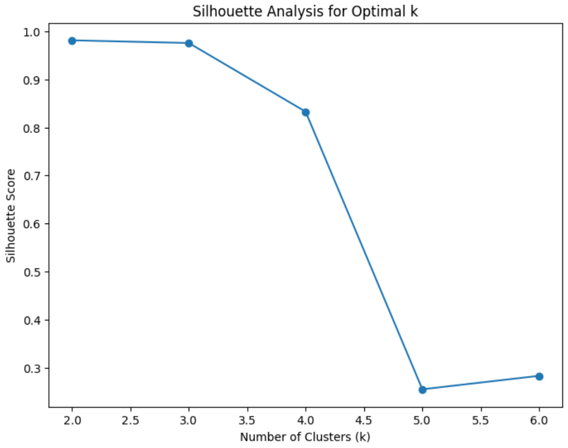
S1: KMeans clustering (*MinMaxScaler)* before removing outliers



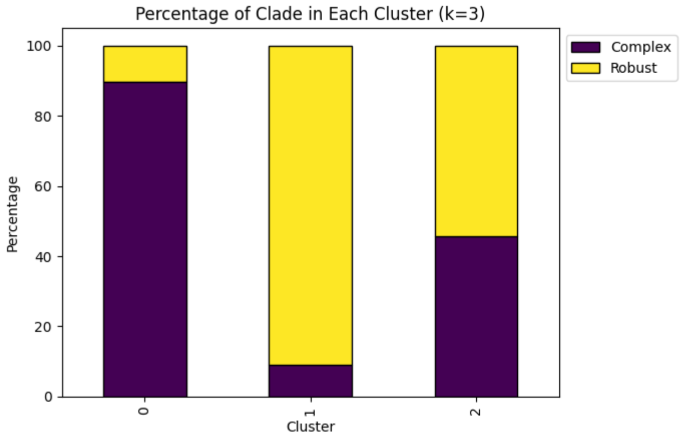
S2: Silhouette score analysis (*MinMaxScaler)* before removal of outliers.



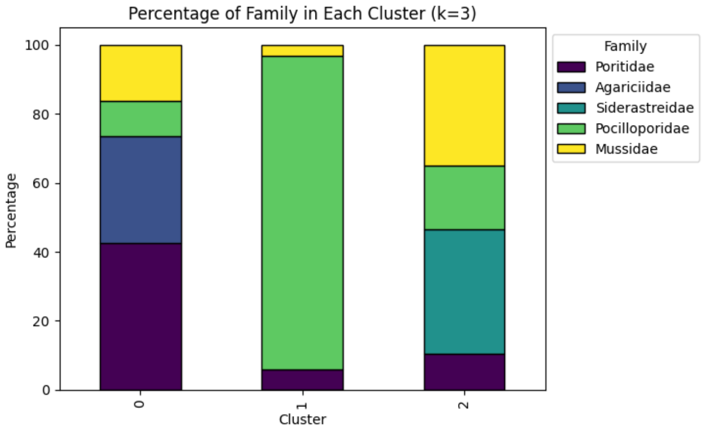
S3: Silhouette score analysis (*StandardScaler)* before removing outliers



S4: Silhouette score analysis (*RobustScaler)* before removal of outliers



S5: Percentage representation of evolutionary clades by cluster



S6: Percent representation of phylogenetic Families by cluster