DINOSAUR CLASSIFICATION: AN EXPLORATION

EMILY SCHEMANSKE

UMSI

ABSTRACT

Finding dinosaur body fossils is rare, and classification of new discoveries is a complex task. The existing fossil record contains many suspiciously similar genera. Representing each genus as a vector allows for quick pairwise comparisons and uncovers possibly problematic records. Geospatial data can further support (or reject) the possibility that two separately classified genera may actually be the same. While not conclusive, the methods presented here may enhance current classification techniques.

INTRODUCTION

Roughly 300 different genera of dinosaurs have been named, but there are some valid questions surrounding whether each is truly unique. In particular, it can be difficult to determine whether a sample is from a juvenile or adult. A 2010 analysis concluded that *Triceratops* and *Torosaurus* actually represent growth stages of a single genus, although this continues to be debated to this day. Here, data mining techniques are used to explore the similarities in *Triceratops*, *Torosaurus*, and others to identify questionable classifications.

MATERIALS & METHODS

Three tables collected from PaleobioDB were combined to conduct this research:

- **measurements** (fossil dimensions)
- **specimens** (categorical fossil data)
- **paleodb** (geospatial fossil data)

Genera were represented in matrix form. Each row represented a single genus, and each column the mean measurement of a particular bone for that genus. They were then compared using (1) cosine similarity and (2) euclidean distance:

$$\cos(X,Y) = \frac{\sum_i^p x_i \cdot y_i}{\sqrt{\sum_i^p x_i^2} \cdot \sqrt{\sum_i^n y_i^2}} \quad (1)$$

$$d(X,Y) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2} \quad (2)$$

Euclidean distance proved to be more informative in this analysis.

NOTES

- Due to sparsity, traditional machine learning approaches, such as decision tree classifiers and k-Means clustering, struggled to find and make use of meaningful patterns.
- Data was imputed where possible, however, many of the missing values could not be filled in a reliable way and had to be dropped.

RESULTS

Since *Torosaurus* and *Triceratops* represent a possible case of misclassification, this study focused on the family *Ceratopsidae* that contains them both. Each genus in that family was represented as a vector of body fossil measurements. Pairwise comparisons were made using measurements common to both genera (several were missing).

Cosine similarity scores were uninformative, with $\mu = 0.997$ and $\sigma = 0.003$. Euclidean distances were more varied, with the smallest value for the pair of *Centrosaurus* and *Chasmosaurus*. The score for *Torosaurus*/*Triceratops* ranked sixth out of 46 comparisons and is included in Table 1 for reference.

Genera	Euclidean Dist
Centrosaurus, Chasmosaurus	64.78
Monoclonius, Styracosaurus	95.51
Torosaurus, Triceratops	112.53

Table 1: Euclidean distances of pairs of dinosaur genera

Fossil measurements are not the only quantities to be considered in the process of fossil identification. The latitude, longitude, and stratigraphic units in which they are discovered are of great importance. It so happens that the majority of *Centrosaurus* and *Chasmosaurus* body fossils lie in the "Dinosaur Park" formation and in the same areas of Alberta and Montana as shown in Figure 1.

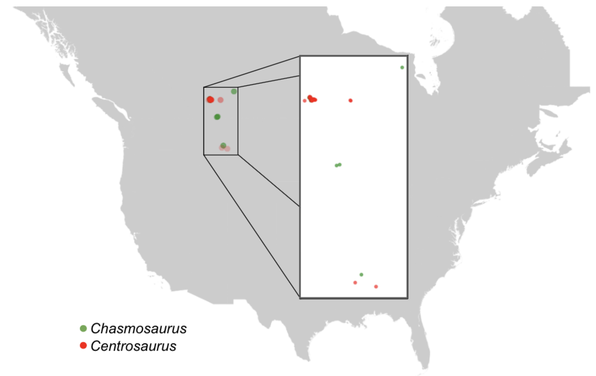

Figure 1: *Chasmosaurus*, *Centrosaurus* fossil distribution

CONCLUSION

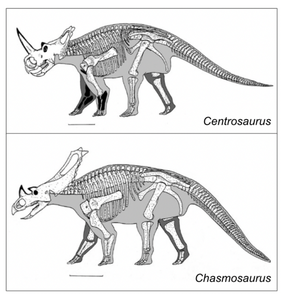

Figure 2: Artist's renderings

The two major findings in this work are that:

1. Pairwise Euclidean distances between dinosaur genera call certain classifications into question, and
2. Spatio-temporal features of specimen records support Euclidean distance metrics.

In other words, fossil pairs with lower (closer) Euclidean distances tend to be found in the same geographic regions and strata. In some way this suggests that it may be a valid metric for finding pairs of dinosaur genera in the fossil record who might actually be the same.

Dinosaur fossil data is relatively rare, so while data mining techniques might not stand on their own as a classification tool, they may support paleobiologists in classifying newly discovered fossils or identifying problematic records.

REFERENCES

[1] The Paleobiology Database. Pbdb service, 2022. Data retrieved from Paleobiology Database on 28 November, 2022, using the 'specimens' and 'measurements' tables, and the following parameters: taxa = Saurischia and Ornithischia. https://paleobiodb.org/classic/displayDownloadGenerator.

FUTURE RESEARCH

Aside from the dinosaur body fossils examined here, the Paleobiology Database contains a wealth of information about body and trace fossils from the kingdoms *Animalia*, *Plantae*, and *Fungi*. The methods for calculating similarities and differences in this subset of the data can be applied to broader categories to test for scale and validity. In addition, given the spatio-temporal nature of the fossil record, a time-series approach could be considered. A more robust dataset may facilitate a more comprehensive representation of each genus, hence more valid comparisons.

CONTACT INFORMATION

GitHub Repository https://github.com/landise/paleobio_classification

Email landise@umich.edu