Learning Comprehensible Relational Features to Distinguish Subfossil Decapod Crustacean Dactyls

Mark Goadrich and Jeffrey Agnew

Department of Mathematics and Computer Science
Department of Geology

Centenary College of Louisiana

Abstract. Our research explores the application of Inductive Logic Programming to a new domain involving decapod crustacean claws. We find that we can distinguish dactyl shapes by automatically extracting relational features that describe their underlying spatial structure. We first use medial axis techniques to find the shock graph of each dactyl outline, which is then converted into a first-order logic representation capturing the connections, distances and angles between the nodes in this graph. We then use Aleph to find relational classification rules based on the shock graph representations. These relational rules provide a concise and human-understandable way to describe the morphological differences between closely related decapods, and can be seen as a first step to creating automatically learned quantitative taxonomic keys.

1 Introduction

Because decapod crustacean claws are potentially affected by numerous selective agents, they are excellent candidates for evolutionary studies of morphology. Despite being commonly found in shell-rich fossil assemblages, decapod dactyls (i.e., claw movable fingers) are usually ignored because of the assumption that they can be identified only to high taxonomic levels.

However, outline-based and geometric morphometric methods have successfully discriminated the dactyls of sibling species and hybrids of the stone crab *Menippe* [2], closely related species of *Panopeus* [3] and other xanthoid genera including *Cataleptodius*, *Dyspanopeus* and *Eriphia* [1]. Principal component analyses of elliptic Fourier descriptors [6] also have been used to quantify ontogenetic shape trajectories and wear in dactyls [1]. Although these techniques allow statistical tests of differences in dactyl morphologies, dactyl shapes must still be described qualitatively.

Our research introduces a new method for distinguishing dactyl shapes by automatically extracting relational features that describe their underlying spatial structure. Using Aleph [9], an Inductive Logic Programming (ILP) algorithm, we learn general rules that capture informative biological relationships, and find that we can limit overfitting by restricting ourselves to a simple representation of the data.

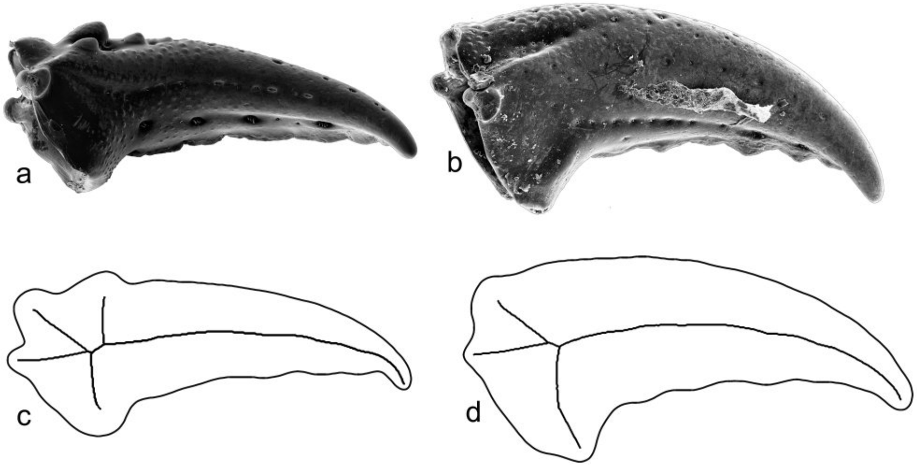

Fig. 1. Scanning Electron Microscope images of *Eriphia gonagra* (a) and *Menippe mercenaria* (b) dactyls, and their corresponding shock graphs, (c) and (d).

2 Dataset Formulation

Our dataset for this study consists of 38 dactyl images, 12 belonging to *Eriphia gonagra* and 26 belonging to *Menippe mercenaria*. Figure 1 (a) and (b) show representative left minor dactyls of these two species. We first use medial axis techniques, used for shape recognition algorithms in computer vision, to find the shock graph of each dactyl outline. Next, these shock graphs are converted into a first-order logic representation capturing the connections, distances and angles between the nodes in each graph.

2.1 Shock Graphs

We begin with the dataset from Agnew [1], where each dactyl image was scaled and aligned using the SHAPE software [5]. To create relational features for each dactyl and expose the underlying skeleton of the images, we chose to convert each image into a shock graph [4] using the flux skeleton implementation of ShapeMatcher [7]. A shock graph is created from a 2D image by first converting the image into an outline. This outline is then thinned along the normal vector according to the calculated flux at each point. Where these normal vectors meet, edges, end points and branch points can be found when looking at the image pixels.

Shock graphs have been used in computer vision as a technique for object recognition; when combined with algorithms for graph similarity, they can help identify when an object has been rotated or distorted over time and space. Sample shock graphs for each species can be found in Figure 1 (c) and (d); note that the top bump in *Eriphia gonagra* creates an edge not seen in *Menippe*

Predicate Type	Predicate Name
Head	eriphia(+example).
Basic	hasNode(+example, -node).
	hasEdge(+example, -edge, -node, -node).
	angle(+edge, +edge, -float).
	distance(+node, +node, -float).
	(+float)>(+float).
	(+float)<(+float).
	(+float)=<(+float).
	(+float)>=(+float).
Acute	obtuse(+float).
	acute(+float).
Full	interiorNode(+node).
	between0and20(+float).
	. . .
	between280and300(+float).

Table 1. Background knowledge and modes generated from shock graph representation.

mercenaria graphs. Also note the difference in length and angle of the bottom bump in *M. mercenaria*. We believe this shape representation can be used to qualitatively understand the phenotypic variations present between these two species.

2.2 First-order representation and Aleph

Aleph [9] is a top-down ILP covering algorithm, written completely in Prolog. As input, Aleph takes background knowledge in the form of either intensional or extensional facts, a list of modes declaring how literals can be chained together, and a designation of one literal as the head predicate to be learned. We chose our head predicate to be the smaller class of *Eriphia gonagra* dactyls, and investigate three levels of background knowledge, shown in Table 1.

First, our **basic** extentional facts are based on the shock graph, such that we create two predicates, `hasNode` and `hasEdge`, to connect the nodes and edges with each example. We also calculate the `angle` between each adjacent edge, the `distance` between any two nodes in the graph, and include the predicates of >, <, >= and =< to compare these angles and distances.

The next level of background knowledge, **acute**, includes intensional definitions for `acute`, floating-point numbers less than 90, and `obtuse`, floating-point numbers greater than 90. Finally, the **full** background knowledge level includes a predicate for designating nodes as being adjacent to two other nodes with `interiorNode`, and `between` predicates to generalize floating-point numbers to into bins of size 20, ranging from 0 to a maximum of 300 because of the maximum image size.

```
eriphia(A) :-
    hasEdge(A,B,C,D), hasEdge(A,E,D,F), hasEdge(A,G,F,H), distance(D,H,I),
    distance(F,C,J), J<I, hasEdge(A,K,F,L), distance(L,H,M),
    J<M, distance(L,C,N), J<N, hasEdge(A,O,P,Q),
    distance(Q,H,R), M<R, distance(L,Q,S), J<S.
```

Fig. 2. Sample rule learned from fold 0, which covered 9 positive and 0 negative training examples, and 3 positive and 0 negative testing examples.

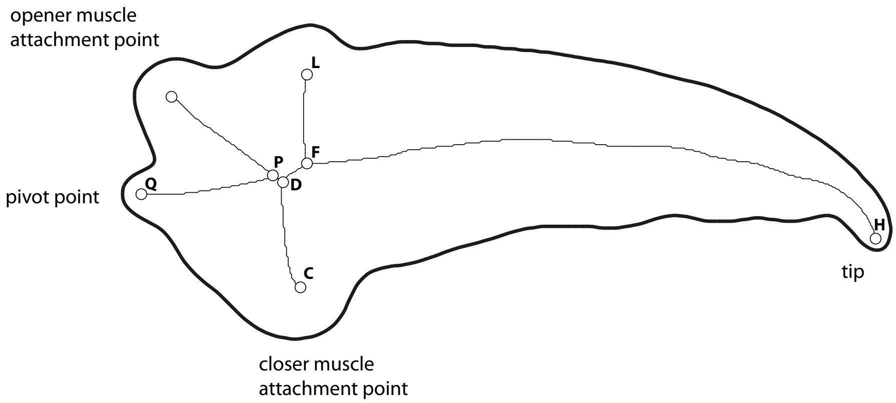

Fig. 3. One possible match of the the nodes C, D, F, H, L, P and Q from the rule in Figure 2 when applied to the first *Eriphia gonagra* example.

3 Experimental Results

We divided the data of 38 examples into five folds of roughly equal size, distributing the positive and negative examples separately to ensure a distribution in each subfold comparable to the complete dataset. In Aleph, we used the **induce** method of exploring and removing seed examples, with the heuristic search method and m-estimate evaluation function, setting m to 20. Other parameter settings changed were to have a minimum accuracy of 0.2, a search depth of 10, a variable-chaining length of 20, a maximum clause length of 20, and a maximum search nodes explored of 20,000.

Figure 2 shows a sample rule learned from fold 0 using only the basic background knowledge. This rule captures all of the positive *Eriphia gonagra* examples and none of the negative examples, in both the training set and testing set. It includes a sequence of connected nodes, C to D to F to H, where the distance between nodes C and F, called J, is less than other calculated distances in this rule. A corresponding distance J is learned in almost all folds, and when this rule is applied to the positive examples, as seen in Figure 3, node C frequently corresponds to the closer muscle insertion point and H to the tip point.

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Basic	89.4	90.0	75.0	81.8
Acute	86.8	88.8	66.6	76.2
Full	81.6	72.7	66.6	69.6
All Pos	31.6	31.6	100.0	48.0
All Neg	68.4	-	0.0	-

Table 2. Pooled results from five-fold cross-validation experiments, comparing Basic, Acute and Full background knowledge.

We believe natural selection could be acting on the distances in this learned relationship between these areas of the shock graph. Because *Menippe* feeds almost exclusively on hard-shelled prey and *Eriphia* is more of an opportunistic generalist, *Menippe* should have claws with stronger biting forces than *Eriphia* [11]. Our learned rule discusses the length and angle of the closer muscle insertion point in relation to the tip. This relationship is directly related to the mechanical advantage of the claw, such that a shorter length in *E. gonagra* will result in weaker closing strength.

We compare the results of using Aleph and each of the three levels of background knowledge (basic, acute and full) with two baseline algorithms, one which classifies all examples as positive, and another which classifies all examples as negative. The true positive, false positive, true negative and false negative results across the five testsets are pooled to find the overall accuracy, precision, recall and F1 score for each algorithm. These results are reported in Table 2, and we can see Aleph clearly outperforms the baseline algorithms.

When comparing the different levels of background knowledge, we find that simpler is better. The heuristic search employed by Aleph incorporates the additional background knowledge predicates into our learned rules, however, these rules have a lower testset performance and tend to overfit, scoring lower than the basic background knowledge across all evaluation metrics.

4 Conclusions and Future Work

This research demonstrates the feasibility of learning relational features to distinguish between decapod dactyl shapes. By combining techniques from computer vision and ILP, we can learn general rules that are informative to both biologists and paleobiologists, and find that we can limit overfitting by restricting ourselves to a simple representation of the data.

Recent work by Macrini *et al.* [8] extends shock graphs to bone graphs to decrease their brittle dependency on noise variations of the initial shape. We plan to replace shock graphs with bone graphs as the basis for learning, and expect to see increases in our performance as well as more general features.

Suard *et al.* [10] have investigated kernel methods applied to shock graphs. They propositionalize many features of the graphs to create their kernels for the purpose of shape retrieval and image clustering, as opposed to our research of learning explanatory and discriminatory patterns using the relational graph descriptions. Although our findings point to less background knowledge instead of more, we plan to investigate some of their features and hopefully increase the understandability of our rules without sacrificing their generalization.

Our current dataset is quite small, with test folds having only 2 or 3 positive examples. We plan to further investigate this approach with a larger dataset consisting of 970 major and minor dactyls from nine xanthoid crab species. This dataset will allow us to evaluate whether this method can be used to distinguish dactyls of several closely related species. Also, because many of the dactyls of these species change shape with growth, we can quantify those allometric transformations and identify dactyl sizes where species level differences emerge.

5 Acknowledgements

We would like to thank the authors of the software packages Aleph, SHAPE and ShapeMatcher for the availability of their code.

References

1. J. G. Agnew. *Dactyls Reveal Evolutionary Patterns in Decapod Crustaceans*. PhD thesis, Louisiana State University - Baton Rouge, 2008.
2. J. G. Agnew and L. C. Anderson. Phenotypic differences among sibling species, hybrids and fossils of the stone crab *menippe*. In *Geological Society of America Annual Meeting*, volume 34, page 399, 2002.
3. J. G. Agnew and L. C. Anderson. Inferring diet of crabs using wear patterns on claws. In *Geological Society of America Annual Meeting*, volume 38, page 442, 2006.
4. P. Dimitrov, C. Phillips, and K. Siddiqi. Robust and efficient skeletal graphs. In *Computer Vision and Pattern Recognition*, 2000.
5. H. Iwata and Y. Ukai. SHAPE: A computer program package for quantitative evaluation of biological shapes based on elliptic fourier descriptors. *The Journal of Heredity*, 93(5):384–5, 2002.
6. F. Kuhl and C. R. Giardina. Elliptic fourier features of a closed contour. *Computer Graphics Image Processing*, 18:236–258, 1982.
7. D. Macrini. Indexing and matching for view-based 3-d object recognition using shock graphs. Master's thesis, University of Toronto, July 2003.
8. D. Macrini, K. Siddiqi, and S. Dickinson. From skeletons to bone graphs: Medial abstraction for object recognition. In *Computer Vision and Pattern Recognition*, 2008.
9. A. Srinivasan. The Aleph Manual Version 5. *http://web.comlab.ox.ac.uk/oucl/research/areas/machlearn/Aleph/*, 2003.
10. F. Suard, A. Rakatomamonjy, and A. Bensrhair. Mining shock graphs with kernels. Technical Report AR-06-01, University of Rouen, Dec 2006.
11. A. B. Williams. *Shrimps, lobsters, and crabs of the Atlantic coast of the eatern United States, Maine to Florida*. Smithsonian Institution Press, Washington, D.C., 1984.