Computational Identification of Actin and Myosin Proteins in Rotifers Through Integrative Machine Learning

Victor E. Rodriguez-Torres[1], Elizabeth J. Walsh[1,2] and Jonathon E. Mohl[1,3]

[1]Bioinformatics Program, The University of Texas at El Paso, 500 W. University Ave., El Paso, 79968, Texas, USA.
[2]Department of Biological Sciences, The University of Texas at El Paso, 500 W. University Ave., El Paso, 79968, Texas, USA.
[3]Department of Mathematical Sciences, The University of Texas at El Paso, 500 W. University Ave., El Paso, 79968, Texas, USA.

vrodrigueztor@miners.utep.edu, ewalsh@utep.edu, jemohl@utep.edu

35th Joint UTEP/NMSU Workshop on Mathematics, Computer Science and Computational Sciences
New Mexico State University, Las Cruces, New Mexico
Saturday, April 11, 2026

Abstract

Actin and myosin are fundamental cytoskeletal proteins that regulate cellular motility, contraction, and structural organization. Characterizing these proteins in rotifers is essential for understanding their cellular biology, development processes, and evolutionary diversification. In this study, we developed an integrative machine learning pipeline to identify actin- and myosin-related sequences in rotifer proteomes. The framework combines Logistic Regression (LR), Multi-Layer Perceptron (MLP), and eXtreme Gradient Boosting (XGBoost), with feature reduction performed using Principal Component Analysis (PCA). Model performance was assessed using five-fold cross-validation, where LR produced the highest F1-score, followed by MLP and XGBoost. The principal components contributing most to classification included descriptors related to hydrophobicity, polarizability, solvent accessibility, charge, and structural properties. To evaluate the accuracy of the pipeline, the trained models were applied to annotated rotifer protein sequences from UniProt which identified 346 actin-like and 503 myosin-like sequences. Predictions were highly consistent across classifiers, supporting the robustness of the framework. Phylogenetic analyses of predicted sequences, together with representatives of established myosin classes, revealed prominent clusters corresponding to myosin classes VI and VII. This integrative approach facilitates large-scale annotation of cytoskeletal proteins in non-model organisms and provides new insights into rotifer functional genomics. Although the models achieved strong predictive performance, experimental validation through gene expression, structural characterization, or functional assays will be necessary to confirm the biological roles of the predicted proteins.

Keywords: bioinformatics pipeline, cytoskeletal proteins, functional genome annotation, machine learning, protein annotation, Rotifera