A NOVEL MACHINE LEARNING-BASED APPROACH FOR PREDICTING PORK TENDERNESS USING TRYPTIC PEPTIDES FROM DISTINCT PROTEIN FRACTIONS

Logan G. Johnson[1], Elisabeth Huff-Lonergan[1] and Steven M. Lonergan[1*]

[1]Department of Animal Science, Iowa State University, United States of America

*Corresponding author email: slonerga@iastate.edu

I. INTRODUCTION

Liquid chromatography-tandem mass spectrometry (LC-MS/MS) has improved the depth and scale of proteome analysis, where a summation of unique peptides is often used to calculate the abundance of individual proteins. Postmortem proteolytic and metabolic changes in pork alter protein solubility and yield a distinct proteome from living muscle. These changes drive development of pork quality [1], specifically meat tenderness, as pork chops with different tenderness values have varying degrees of proteolysis, resulting in different proteomes [2]. The extent to which tryptic peptides can help characterize the molecular phenotype and predict fresh pork quality is not known. The objective was to determine the utility of individual peptide abundances and machine learning models to predict pork tenderness. It was hypothesized that the unique peptide abundances would predict pork tenderness.

II. MATERIALS AND METHODS

Fresh pork loins (N = 120) were collected from a commercial harvest facility at 1 d postmortem. Quality attributes were assessed at 1 d postmortem and after approximately 2 weeks of aging [3]. Chops were ranked based on instrumental star probe (SP) values, a measure of meat tenderness, and divided into 4 distinct categories (A, $\bar{x}$ = 4.23 kg, 3.43–4.55 kg; B, $\bar{x}$ = 4.79 kg, 4.66–5.00 kg; C, $\bar{x}$ = 5.43 kg, 5.20–5.64 kg; D, $\bar{x}$ = 6.21 kg, 5.70–7.41 kg; n = 25 per category). Aged chops were homogenized in liquid nitrogen, and proteins soluble in a low-ionic strength buffer (50 mM Tris-HCl [pH 8.5] and 1 mM ethylenediaminetetraacetic acid; Sarcoplasmic) were extracted. Separately, the insoluble protein from the sarcoplasmic extract was collected and washed with Standard Salt Solution (100 mM potassium chloride, 20 mM potassium phosphate, 2 mM magnesium chloride, 2 mM ethylenebis[oxyethylenenitroilo] tetraacetic acid, and 1 mM sodium azide) and Tris Wash Buffer (5 mM Tris-HCl [pH 8.0]), and solubilized in a buffer containing 8.3 M urea, 2 M thiourea, and 1% dithiothreitol (Myofibrillar) [4]. Protein concentration was determined on each extract. The two protein fractions for each sample were analyzed in separate LC-MS/MS experiments but were prepared following a similar protocol. Samples were processed and analyzed with LC-MS/MS as described [3]. Within a run, peptide ratios were calculated, comparing each sample to the master control abundance for that run. Only peptides identified as unique and used to calculate protein abundance were utilized. All analyses were conducted in R (v. 4.2.2) and RStudio using the *caret* package [5]. Peptide data were $\log_2$ transformed and median normalized. Only peptides identified in at least half of the samples (n > 50) were retained. Missing values were imputed using a bagged tree model using the *missForest* package. Data were split into training (80%) and testing (20%) subsets with equal distribution of predicted variables represented between the subsets. Highly correlated predictors were removed (Pearson's $|r| > 0.90$). Recursive feature elimination was conducted using the rfe function from the *caret* package and assessed with 10-fold cross-validation repeated 5 times. Machine learning models were employed to predict SP category (classification) and value (regression). Models were trained using 10-fold cross-validation repeated 5 times, and the best resulting model was used to predict the testing set. The predictive metrics of each model were assessed using model accuracy, Cohen's kappa coefficient (Kappa), predicted root mean square error (RMSE), and mean absolute error (MAE).

III. RESULTS AND DISCUSSION

Predictor variable summary information after each filtering step is outlined in Table 1. Model predictive abilities are highlighted in Table 2. The top 4 predictors of importance for star probe category included titin isoform X6, AMP deaminase, and fumarate hydratase peptides from the sarcoplasmic fraction and an elongation factor 1-alpha peptide from the myofibrillar fraction. The top 4 predictors of importance for star probe value were three titin isoform X6 peptides from the sarcoplasmic fraction and one elongation factor 1-alpha peptide from the myofibrillar fraction. Models with the highest classification accuracy were bagged classification and regression trees and random forest with 55% and 50% accuracy, respectively. Models with the best regression performance included a support vector machine- polynomial kernel and stochastic gradient boosting with the lowest RMSE and MAE values.

Table 1 Summary information on predictor variables

Predictor Variables	Total	Filter *n* > 50[1]	Filter |r| > 0.90[2]
Sarcoplasmic Fraction	5,119	2,178	1,545
Myofibrillar Fraction	2,777	1,547	434

[1] Number of predictors after removing those in fewer than 50 samples
[2] Number of predictors after removing highly correlated predictors

Table 2 Predictive metrics of machine learning models for classification and regression analyses

Model	R Package	Classification		Regression	
		Accuracy	Kappa	RMSE	MAE
Bagged Classification & Regression Trees	*ipred*	0.55	0.400	0.968	0.863
Multivariate Adaptive Regression Spline	*earth*	0.35	0.133	0.953	0.819
Random Forest	*randomForest*	0.50	0.333	0.929	0.840
Stochastic Gradient Boosting	*gbm*	0.45	0.267	0.887	0.713
Support Vector Machine- Linear Kernel	*kernlab*	0.35	0.133	0.891	0.782
Support Vector Machine- Radial Kernel	*kernlab*	0.45	0.267	0.913	0.809
Support Vector Machine- Polynomial Kernel	*kernlab*	0.25	0.000	0.835	0.720
Boosted Logistic Regression	*caTools*	0.40	0.187	–	–
Linear Discrimination Analysis	*MASS*	0.40	0.200	–	–
Model Averaged Neural Network	*nnet*	0.45	0.267	–	–
Naive Bayes	*naivebayes*	0.40	0.200	–	–

IV. CONCLUSION

The fractionation of proteins based on solubility highlights the complexity of the aged meat proteome compared to the proteome of muscle. Individual peptide data were predictive of SP category and value using classification and regression approaches. Future work with a greater number of samples and more refined and tuned models will help validate these observations and identify predictors associated with pork tenderness.

ACKNOWLEDGEMENTS

Funding from the USDA-NIFA project 2019 6701729181, IAHEES project IOW04121, and the National Science Foundation under Grant No. DGE-1828942 is acknowledged.

REFERENCES

1. Melody, J. L., Lonergan, S. M., Rowe, L. J., Huiatt, T. W., Mayes, M. S., & Huff-Lonergan, E. (2004). Early postmortem biochemical factors influence tenderness and water-holding capacity of three porcine muscles. Journal of Animal Science 82: 1195–1205.
2. Carlson, K. B., Prusa, K. J., Fedler, C. A., Steadham, E. M., Huff-Lonergan, E., & Lonergan, S. M. (2017). Proteomic features linked to tenderness of aged pork loins. Journal of Animal Science 95: 2533–2546.
3. Johnson, L. G., Zhai, C., Reever, L. M., Prusa, K. J., Nair, M. N., Huff-Lonergan, E., & Lonergan, S. M. (2023). Characterizing the sarcoplasmic proteome of aged pork chops classified by purge loss. Journal of Animal Science 101: 1–12.
4. Carlson, K. B., Prusa, K. J., Fedler, C. A., Steadham, E. M., Outhouse, A. C., King, D. A., Huff-Lonergan, E., & Lonergan, S. M. (2017). Postmortem protein degradation is a key contributor to fresh pork loin tenderness. Journal of Animal Science 95: 1574–1586.
5. Kuhn, M., & Johnson, K. (2013). Applied predictive modeling. New York, NY: Springer New York.