DOI 10.5281/zenodo.5512429

Original paper

Arabinogalactan protein mining and diversity - the case of *Centaurium erythraea*

Milan DRAGIĆEVIĆ*, Ana SIMONOVIĆ

Department for Plant Physiology, Institute for Biological Research "Siniša Stanković", National Institute of Republic of Serbia, University of Belgrade, Belgrade, Serbia

Accepted: 24 August 2021 / Published online: 20 September 2021

Summary. *Centaurium erythraea* (common centaury) is a medicinal plant with extraordinary developmental plasticity *in vitro* that is used as a model organism for studying *in vitro* morphogenesis in our lab. Several experimental lines of evidence have identified arabinogalactan proteins (AGPs) as one of the key players involved in centaury morphogenesis; however, the role of specific genes has yet to be determined. AGPs are ubiquitous plant cell surface glycoproteins associated with various physiological functions. AGP sequences are characterized by the presence of non-continuous hydroxyproline residues, which serve as O-glycosylation anchor sites for branched arabinogalactans. Due to a biased amino acid composition rich in disorder-promoting amino acids, AGP sequences lack a stable structure and consequently have lessened evolutionary constraints. Therefore, homology-based approaches to AGP sequence mining have limited success. We have recently developed a bioinformatics pipeline for AGP sequence mining, ragp, which exploits their key feature – the presence of hydroxyprolines. This pipeline combines estimation of proline hydroxylation based on local sequence context by a machine learning model with a flexible motif search. After applying this pipeline to the centaury transcriptome, AGP regions were found to associate with a variety of conserved domains. Here we introduce a streamlined way to train models for prediction of Pro hydroxylation, analyze important protein sequence features determining Pro hydroxylation status, present some of the AGP types found in centaury and discuss model limitations and future prospects.

Keywords: arabinogalactan proteins, centaury, hydroxyproline, hydroxyproline rich glycoproteins, machine learning, ragp, sequence mining.

INTRODUCTION

Arabinogalactan proteins (AGPs) are ubiquitous plant cell surface glycoproteins belonging to the hydroxyproline rich glycoprotein (HRGP) superfamily (Ellis et al. 2010). HRGPs are involved in a variety of physiological functions during plant growth and development, signaling, environmental sensing and response to external stimuli such as infection and wounding (Deepak et al. 2010; Kieliszewski et al. 2010). The HRGP superfamily members are characterized by the presence of 4-hydroxyproline (Hyp, O) produced via hydroxylation of proline residues by prolyl 4-hydroxylases during protein maturation. Unlike many other post-translational modifications, proline hydroxylation is irreversible, and common in animals, plants and microbes (Gorres and Raines 2010). In plants, these hydroxyprolines serve as O-glycosylation sites, while the local sequence context determines the type of glycosylation. Based on glycosylation type, HRGPs are usually divided into: (1) extensins, characterized by the presence of continuous O stretches, usually preceded by serine (e.g., SOOO), which are glycosylated by short and linear oligosaccharides made from L-arabinose (Ara); (2) arabinogalactan proteins, characterized by clustered non-continuous dipeptide motifs of O with A, S, T, G and V (e.g. OA, AO, OT, TO etc.), which are glycosylated by branched type II arabino-3,6-galactan polysaccharides, and (3) proline-rich proteins, characterized by OOV[QK], OOVX[KT], and KKOCOO motifs which are less extensively glycosylated

E-mail: mdragicevic@ibiss.bg.ac.rs

with short Ara residues (Johnson et al. 2017). Hybrid and chimeric HRGPs have also been identified; the former contain mixed characteristics of the mentioned HRGP classes, whereas the latter contain HRGP features associated with specific protein domains. For instance, AGP regions have been shown to associate with fasciclin-like, plastocyanin-like (early nodulin-like) and other domains (Ellis et al. 2010).

Due to a biased amino acid composition rich in O as well as polar/charged residues, which are disorder-promoting, HRGP proteins/regions lack a well-defined three-dimensional structure and are intrinsically disordered proteins (IDP). This property of HRGPs has resulted in a larger sequence space and accelerated evolution (Johnson et al. 2017). As a consequence, HRGP sequences are usually not identified by homology search (e.g., BLAST or diamond), but by their biased amino composition and the presence of described motifs, combined with a predicted N-terminal secretory signal (N-sp), because HRGPs are extracellular proteins (Showalter et al. 2010, Johnson et al. 2017). These approaches are confined to the identification of prototypical HRGPs, while chimeric sequences, which lack a biased amino acid composition, are usually identified by the presence of protein domains known to associate with HRGPs, and specifically AGPs. However, this is limited to sequences containing domains already known to associate with AGP regions.

In an attempt to expand the AGP sequence search space beyond sequences with a biased composition, Ma et al. (2017) developed a bioinformatics pipeline which combines identification of clustered noncontinuous dipeptides characteristic for AGPs with several other sequence features, which led to the discovery that AGP-like regions are commonly associated with a wide range of protein domains, such as the receptor-kinase domain, X8 domain, leucine rich repeat motifs and glycosyl hydrolase-like among others. To facilitate HRGP sequence mining, and to make it available to a broader audience, we have developed the ragp R package (Dragićević et al. 2020, https://github.com/missuse/ragp, https://missuse.github.io/ragp/). In addition to the usual HRGP mining toolbox, such as prediction of N-sp, motif and amino acid bias classification (Johnson et al. 2017), as well as a search for clustered AG motifs, ragp incorporates an additional filtering layer – machine learning (ML) prediction of proline hydroxylation probability, thus exploiting the key HRGP feature. Only sequences predicted to contain several O are further analyzed. In addition, searches for clustered AG motifs are able to use information provided by this model, so instead of searching for AG motifs containing P, ragp allows users to search for AG motifs containing predicted O.

Centaurium erythraea Rafn (common or European centaury) is a medicinal plant, characterized by a vigorous regeneration potential and extraordinary developmental plasticity *in vitro*. Centaury has been used for decades as a model organism for studying morphogenetic processes *in vitro* at the Department for Plant Physiology (Simonović et al. 2021). Studies based on interaction of AGPs with β-D-glucosyl Yariv reagent, which selectively binds AGPs causing their precipitation and inactivation (Trifunović et al. 2014; Simonović et al. 2015), or with fluorescent antibodies that bind AGP glycan epitopes (Filipović et al. 2021), have linked AGPs with somatic embryogenesis and shoot organogenesis in centaury. However, the roles of specific genes have yet to be found. The first step is identification of sequences of potential interest.

In this paper we: (1) present the diversity of chimeric AGP sequences identified in centaury; (2) introduce a more efficient and streamlined ML pipeline for building models for proline hydroxylation prediction, which can accommodate reproducible rapid retraining on new training data; (3) dissect the model predictions, thus providing insight into what local protein sequence features are important for proline hydroxylation in plants, and (4) provide suggestions for future directions.

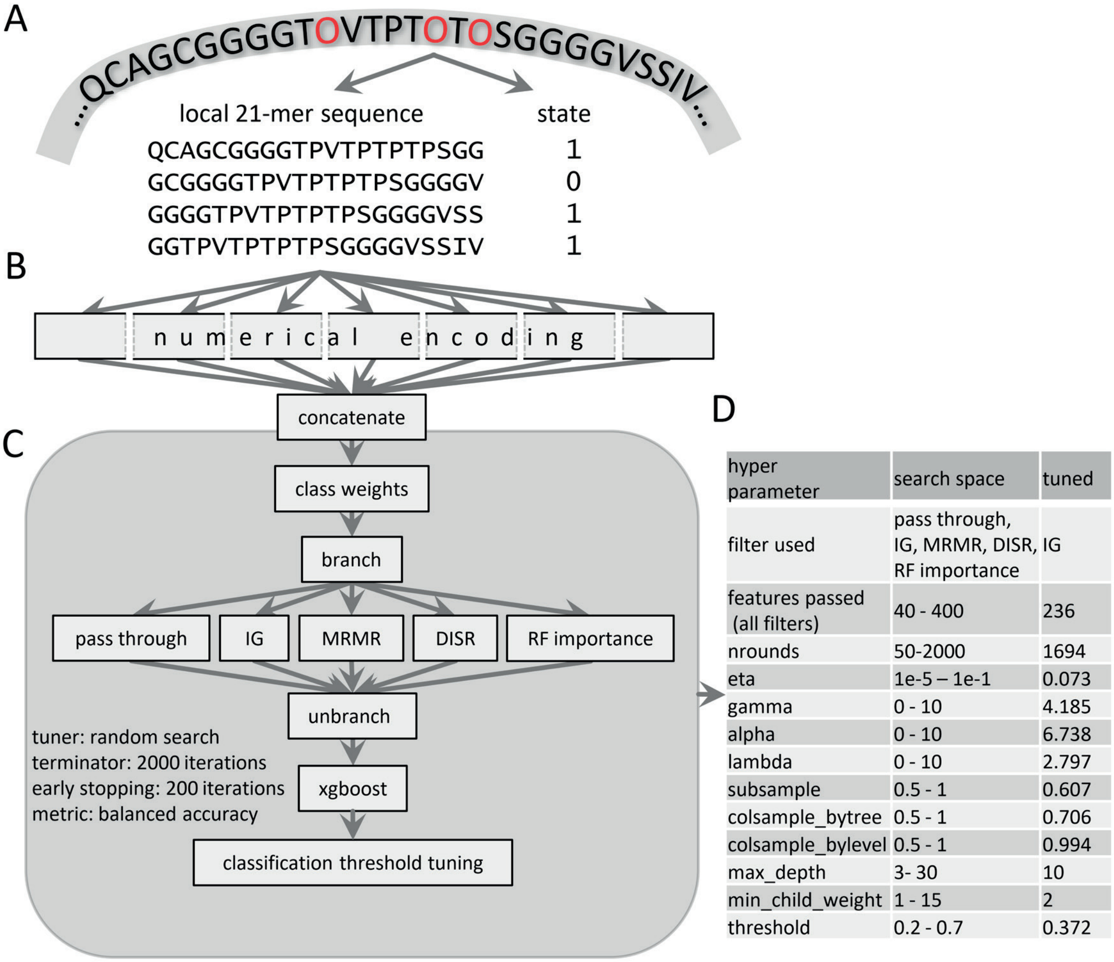
MATERIALS AND METHODS

AGP mining and annotation

AGP sequence mining was performed as recommended in Dragićević et al. (2020) with one modification: the number of amino acids separating AG motifs was set to four instead of the recommended ten, in order to perform a slightly more stringent scan. As a starting point, the complete predicted protein sequences from the *de novo* assembled centaury transcriptome (Ćuković et al. 2020) were used. Domain annotation was performed using hmmer 3.3.2 (Eddy 2011) with Pfam34 database (http://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam34.0/). Transmembrane region annotation was performed using Phobius (Käll et al. 2007), while glycosylphosphatidylinositol (GPI) anchor sites were annotated using PredGPI (Pierleoni et al. 2008) *via* the ragp interface (Dragićević et al. 2020).

Machine learning pipeline creation and tuning

The same training/test sets and protein sequence features were utilized as in Dragićević et al. (2020) in order to achieve performance comparability (Fig. 1A, B). The ML pipeline was created using the mlr3 set of packages (Lang et al. 2019). The pipeline consisted of: inferring class weights; filter layer; learner layer and classification threshold tuning (Fig. 1C). Class weights were set due to disbalance in the class frequencies and were equal to the ratio of the class-



hyper parameter	search space	tuned
filter used	pass through, IG, MRMR, DISR, RF importance	IG
features passed (all filters)	40 - 400	236
nrounds	50-2000	1694
eta	1e-5 – 1e-1	0.073
gamma	0 - 10	4.185
alpha	0 - 10	6.738
lambda	0 - 10	2.797
subsample	0.5 - 1	0.607
colsample_bytree	0.5 - 1	0.706
colsample_bylevel	0.5 - 1	0.994
max_depth	3- 30	10
min_child_weight	1 - 15	2
threshold	0.2 - 0.7	0.372

Fig. 1. Schematic of a streamlined ML pipeline for training the hydroxyproline prediction model. **A,** identification of protein sequence regions with sufficient experimental evidence based on Swiss-Prot and literature data; extraction of local 21-mer sequence with 10 amino acids on each side of the target prolines as well as hydroxylation state of the target prolines; **B,** encoding the local protein sequences into various numeric representations; **C,** machine learning pipeline (see Materials and Methods for clarification); **D,** hyper parameter search space and optimized values.

es. The filter layer contained the following branches: pass through - does not filter or transform the feature set, IG - information gain univariate feature filter (Quinlan 1986), MRMR - minimum redundancy maximum relevance univariate feature filter (Peng et al. 2005), DISR - double input symmetrical relevance univariate feature filter (Meyer et al. 2008), RF importance - feature filter based on feature importance of a random forest learner (Breiman 2001) fit with default parameters and 500 trees. In each tuning instance, only one path through the graph was evaluated, so selection of the best path was performed during tuning. Each of the filters was set to pass between 40 and 400 features/dimensions – treated as a tunable hyper-parameter. The learner layer consisted of an xgboost algorithm (Chen and Guestrin 2016), with multiple tuned hyper-parameters (Fig. 1D). The ML algorithm output probabilities were converted to classes based on a classification threshold that was treated as a tunable hyper-parameter. The whole pipeline was tuned via random search for 2000 iterations, with early termination if performance did not improve for 200 iterations, using five-fold cross validation. The metric used for optimization was balanced accuracy (the mean of sensitivity and specificity). During the five-fold cross validation, all k-mers from the same protein were present in the same fold (block resampling), and the ratio of the classes (Hyp/Pro) was kept similar in each resampling instance (stratification).

Model interpretation

Model agnostic interpretation of model behavior was performed using the DALEX package (Biecek 2018). Permutation feature importance was calculated based on reduction in performance measured as AUC; when the corresponding feature was permuted, this was repeated 10 times. Partial dependence plots (Friedman 2001) were employed in order to display a marginalized version of the learned function in a lower dimensional space.

RESULTS AND DISCUSSION

Diversity of centaury AGPs

The ragp pipeline identified 501 unique complete protein sequences that contain at least three AG motifs containing O, no more than four amino acids apart in the centaury transcriptome. Of these, 278 protein sequences had at least one Pfam domain associated with it, while 134 had multiple Pfam domains. AGP regions are associated with an array of different protein architectures, a subset of which is presented in Fig. 2. AGP regions were detected prior to the transmembrane helix in multiple centaury receptor-like kinases. AGP spans were also found in cell wall hydrolytic enzymes, such as cysteine proteases and glycosyl hydrolases, associated with domains of unknown function, DUF1191 and DUF1682, with L-ascorbate oxidase-like proteins, as well as with dirigent-like domains. As expected, AGP regions are usually located outside of domains (conserved regions), except in the case of arabinogalactan peptide and COBRA-like domains (Fig. 2). Arabinogalactan peptide domain is a signature domain of a class of AG peptides, and is the only domain that we know of which is exclusively found associated with AGP regions (Simonović et al. 2016). Experimental confirmation is needed for COBRA-like proteins; however, they are attached to the membrane by GPI anchors (Roudier et al. 2002) like many AGPs (Ellis et al. 2010), and contain multiple N-glycosylation motifs (Asn–X–Ser/Thr), so it would not be surprising if they were also glycosylated with branched arabinogalactans. This analysis indicates that glycosylation with branched arabinogalactan chains is common in cell wall proteins, many of which have central roles in cell wall metabolism, ranging from signal transduction, lignan and lignin biosynthesis, pectate hydrolysis, cellulose crystallization and others.

An updated pipeline for prediction of proline hydroxylation

The development of the mlr3 (Lang et al. 2019) R package for efficient, object-oriented programming on the building blocks of machine learning has enabled unprecedented freedom in ML pipeline construction in R. Using the described framework, we constructed an updated ML pipeline for hydroxyproline probability prediction which can be easily and efficiently trained on an existing or expanded protein

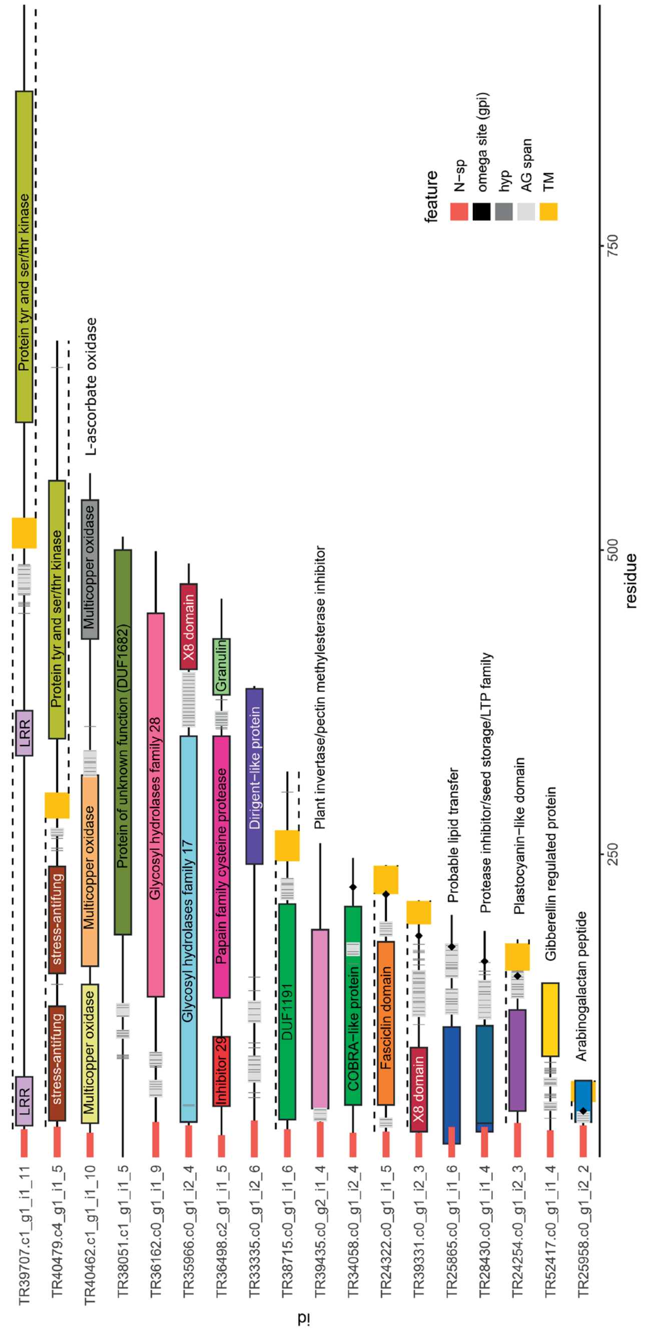


Fig. 2. Diagram of protein architectures of selected ragp-mined chimeric AGPs from centaury transcriptome. N-sp – N-terminal secretion signal peptide as predicted by SignalP5 (Armenteros et al. 2019); TM – transmembrane region as predicted by Phobius (Käll et al. 2007); hyp – predicted hydroxyprolines via ragp (Dragićević et al. 2020); AG span – ragp hyp-aware scan of AGP motifs, where three or more motifs no more than four amino acids apart were considered, while motifs which were part of continuous stretches of three or more hyp were not considered; omega site – GPI anchor sites as predicted by PredGPI (Pierleoni et al. 2008). Domains were annotated using hmmscan 3.3.2 (Eddy 2011) using Pfam 34 data base of protein domains.

sequence training set (Fig. 1). This ML pipeline achieves a balanced accuracy of 0.983 (sensitivity: 1 and specificity: 0.966, Fig. 3A) when using the same training/test sets and local sequence numerical representations as in Dragićević et al. (2020), which is comparable to the performance of the model incorporated in the ragp package, with the advantage of reducing the training time to several hours on a desktop PC. The main difference compared to the model incorporated in ragp is abandonment of computationally demanding wrapper feature selection which applies model tuning for every feature subset, unification of the filter methods so that they compete within the same ML graph, and jointly tuning the classification threshold with other hyper parameters (Fig. 1). The pipeline can be easily extended with the addition of new blocks, such as data transformations, additional filters and learners. The 2000 iterations of random search serve to inspect a large hierarchical feature space including different learner and filter hyper-parameters, as well as to probe paths through the directed graph, while the early termination feature serves to reduce training time if good hyper parameter combinations are found early during the search. The same or similar pipeline can be used to predict practically any feature based on local protein sequence given an appropriate training set.

Local sequence features which govern proline hydroxylation in plants

While the primary interest of predictive modeling is to generate accurate predictions, a secondary interest is to understand why the model works. In the current application, it might be possible to gain insights into what local sequence features determine proline hydroxylation status. In an attempt to understand the reasoning behind the obtained predictions, a model agnostic approach was used where the variables were ranked by importance based on reduction of model performance when each variable is permuted (Fig. 3B). After obtaining variable importance, the distribution of the top five features was examined per target class (Fig. 3C), while partial dependence plots showed the marginalized dependence between the target class and the top five important features (Fig. 3D). The relationship between the kernel density (Fig. 3C) and partial dependence (Fig. 3D) of the top five features is evident.

The feature with the highest importance, "Grantham. Xr.P" (Fig. 3B) belongs to the quasi sequence order descriptor (Chou 2000) derived from the amino acid distance matrix proposed by Grantham (1974). It is calculated based on the sequence order coupling $\tau_d = \sum_{i=1}^{N-d} (d_{i,i+d})^2$ where *d=1, 2,…maxlag* and $d_{i,i+d}$ is the distance between the two amino acids at position *i* and *i+d*; for each amino acid type, a quasi-sequence-order descriptor can be defined as $X_r = \frac{f_r}{\sum_{r=1}^{20} f_r + \omega \sum_{d=1}^{maxlag} \tau_d}$ where f_r is the normalized occurrence for amino acid type *r* and *ω* is a weighting factor (*ω* = 0.1). Increase in the variable "Grantham.Xr.P" is interpreted by the model as in increase in the probability of Pro hydroxylation (Fig. 3D), thus it can be concluded that an increase in Pro frequency (increase in f_r where r = Pro), as well as higher content of amino acids which are close to Pro in the Grantham distance matrix (A, T and G) in k-mers (lower $\tau_d = \sum_{i=1}^{N-d} (d_{i,i+d})^2$) favors Pro hydroxylation. Two of the five most important features are associated with the amino acid preceding the target Pro (features CHAM820102_-1 and F4_-1, Fig. 2B), leading to the conclusion that this position is critical for Pro hydroxylation. The "CHAM820102" attribute represents the free energy of solution in water (Charton M and Charton BI 1982) and values lower than 0 (amino acids P, R, G, S, A) are associated with an increase in hydroxylation probability (Fig. 3D). While Feature "F4_-1" represents one to one mapping of Atchley factor 4 (Atchley et al. 2005) to the amino acid preceding the target Pro. Values above -0.17 of this feature, which are associated with the amino acids N, E, I, P, R, S, T, G, V, L and A, positively impact the probability of proline hydroxylation (Fig. 3D).

Larger than 0.3 values of the normalized Moreau-Broto amino acid autocorrelation of "CHAM820102" with lag = 1 and 3 respectively, are associated with an increase in the probability of Pro hydroxylation (Fig. 2D). These parameters are calculated by $ATS_{(d)} = \frac{\sum_{i=1}^{N-d} P_i P_{i+d}}{N-d}$ where lag *d = 1, 3*, while P_i and P_{i+1} represent the centralized and standardized "CHAM820102" attribute of the amino acids at position *i* and *i+d*, while N is the number of amino acids in the k-mer. It is not simple to draw conclusions from this relationship, therefore we performed a simulation experiment where we generated 1 million random 21-mer protein sequences with P in the middle. Approximately 5.6% of these sequences produced a value of "MB_CHAM820102.lag1" higher than 0.3, while 7.4% had a value of "MB_CHAM820102.lag3" higher than 0.3, while ~2.4% of these sequences satisfied both requirements. The sequences that satisfied both requirements had a higher occurrence of P (1.49 per sequence on average, excluding the target middle P which is present in every sequence), R (1.24 per sequence on average), G (1.12) and S (1.11), while the most common dipeptides were PP, RP, PR, GP, PS, SP, PG, PA and AP, the latter six are AG motifs. Based on the top five features it seems that the model has managed to learn local sequence features which were already associated with HRGP/AGPs such as local stretches of P characteristic of extensins, or AG motifs. While it is slightly

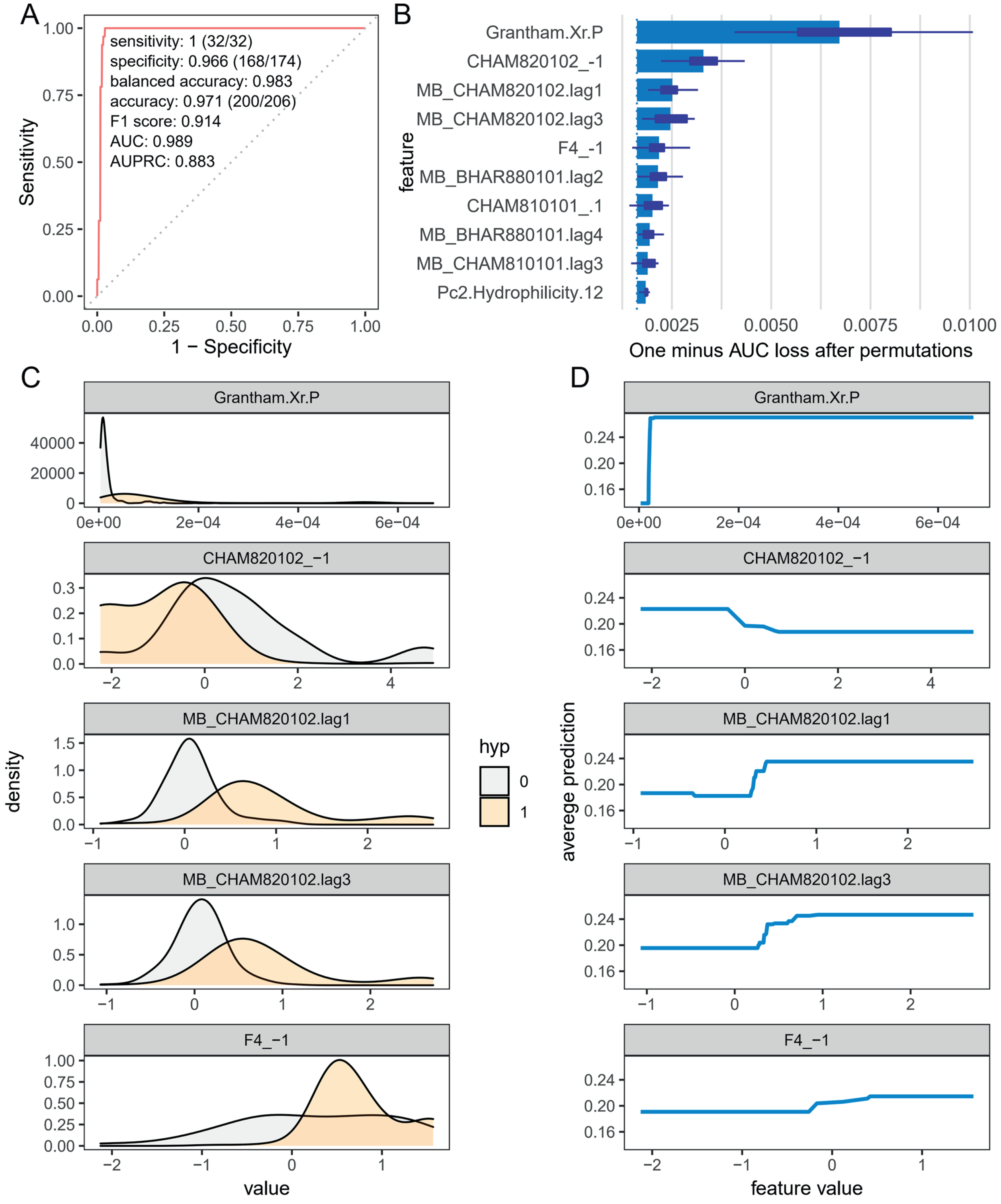


Fig. 3. Performance and interpretation of the model. **A**, receiver operating characteristic curve along with several performance metrics of hydroxyproline prediction pipeline based on the validation protein sequences; **B**, permutation feature importance of the top 10 features, calculated based on reduction in performance measured as AUC, when the corresponding feature was permuted. Bar length represents the mean loss of performance while boxplots show the uncertainty associated with permutations (10 permutations used); **C,** kernel density plots of the top five features grouped by hydroxylation status (indicated in the color legend); **D**, partial dependence plots of the top five features.

anti-climactic to conclude that the model has learned something we already knew, it is also reassuring both regarding our presumptions and confidence in the model predictions. In addition, we should mention the model uses around 230 features to predict P hydroxylation and not just the top five most important considered here.

Model limitations and future prospects

The current approach to ML hydroxyproline prediction has several limitations. Due to use of symmetrical 21-mers, the hydroxylation status of P in the ten N- or C-terminal amino acids cannot be estimated. In ragp this is compensated by training supplementary models on shorter k-mer spans sacrificing some accuracy for the ability to predict N- and C-terminal prolines. The other limitation which is not easily compensated for is the relatively limited set of labeled protein sequences, containing 225 protein sequences with 1093 non-redundant 21-mers (182 hydroxyprolines and 911 prolines), which limits optimism for model generalization power. The growth of this set depends on experimental determination of hydroxyproline positions in plant protein sequences, which is expensive and time consuming. A potential solution to increase how well the trained models generalize, which does not rely on new labeled data, is the use of protein language models (Elnaggar et al. 2020), which were pretrained on a large corpus of protein sequences in a self-supervised fashion by randomly masking a portion of the amino acids in the input and training the model to predicted the masked residues. These models can be used for protein feature extraction or be fine-tuned on downstream tasks. We are currently exploring this strategy for hydroxyproline/proline classification.

ACKNOWLEDGEMENTS

This research was funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia (Grant № 451-03-9/2021-14/200007).

REFERENCES

Armenteros JJA, Tsirigos K, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nature Biotechnology. 37:420–423.

Atchley WR, Zhao J, Fernandes AD, Drüke T. 2005. Solving the protein sequence metric problem. Proceedings of the National Academy of Sciences USA. 102:6395–6400.

Biecek P. 2018. DALEX: Explainers for Complex Predictive Models in R. Journal of Machine Learning Research 19(84):1–5.

Breiman L. 2001. Random Forests. Machine Learning. 45:5–32.

Charton M, Charton BI. 1982. The structural dependence of amino acid hydrophobicity parameters. Journal of Theoretical Biology. 99(4):629–644.

Chen T, Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System. arXiv.1603.02754.

Chou KC. 2000. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochemical and Biophysical Research Communication. 278:477–483.

Ćuković K, Dragićević M, Bogdanović M, Paunović D, Giurato G, Filipović B, Subotić A, Todorović S, Simonović A. 2020. Plant regeneration in leaf culture of *Centaurium erythraea* Rafn. Part 3: *de novo* transcriptome assembly and validation of housekeeping genes for studies of in vitro morphogenesis. Plant Cell, Tissue and Organ Culture. 141:417–433.

Deepak S, Shailasree, S, Kini RK, Muck A, Mithöfer A, Shetty SH. 2010. Hydroxyproline-rich glycoproteins and plant defence. Journal of Phytopathology. 158:585–593.

Dragićević M, Paunović D, Bogdanović M, Todorović S, Simonović A. 2020. ragp: Pipeline for mining of plant hydroxyproline-rich glycoproteins with implementation in R. Glycobiology 30(1):19–35.

Eddy SR. 2011. Accelerated Profile HMM Searches. PLOS Computational Biology. 7:e1002195

Ellis M, Egelund J, Schultz CJ, Bacic A. 2010. Arabinogalactan-proteins: key regulators at the cell surface? Plant Physiology. 153:403–419.

Elnaggar A, Heinzinger M, Dallago C, Rihawi G, Wang Yu, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M et al. 2020. ProtTrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. arXiv. 2007.06225

Filipović B, Trifunović-Momčilov M, Simonović A. Jevremović S, Milošević S, Subotić A. 2021. Immunolocalization of some arabinogalactan protein epitopes during indirect somatic embryogenesis and shoot organogenesis in leaf culture of centaury (*Centaurium erythraea* Rafn). In Vitro Cellular and Developmental Biology – Plant. 57:470–480.

Friedman JH. 2001. Greedy function approximation: A gradient boosting machine. The Annals of Statistics. 29(5):1189–1232.

Gorres KL, Raines RT. 2010. Prolyl 4-hydroxylase. Critical Reviews in Biochemistry and Molecular Biology. 45(2):106–124.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. Science. 185(4154):862–864.

Johnson KL, Cassin AM, Lonsdale A, Bacic A, Doblin MS, Schultz CJ. 2017. A motif and amino acid bias bioinformatics pipeline to identify hydroxyproline-rich glycoproteins. Plant Physiology. 174:886–903.

Käll L, Krogh A, Sonnhammer EL. 2007. Advantages of combined transmembrane topology and signal peptide prediction — the Phobius web server. Nucleic Acids Research. 35:429–432.

Kieliszewski MJ, Lamport DTA, Tan L, Cannon MC. 2010. Hydroxyproline-rich glycoproteins: form and function. Annual Plant Reviews. 41:321–342.

Lang M, Binder M, Richter J, Schratz P, Pfisterer F, Coors S, Au Q, Casalicchio G, Kotthoff L, Bischl B. 2019. mlr3: A modern object-oriented machine learning framework in R. Journal of Open Source Software. 4(44):1903.

Ma Y, Yan C, Li H, Wu W, Liu Y, Wang Y, Chen Q, Ma H. 2017. Bioinformatics prediction and evolution analysis of arabinogalactan proteins in the plant kingdom. Frontiers in Plant Science. 8:66.

Meyer PE, Schretter C, Bontempi G. 2008. Information-theoretic feature selection in microarray data using variable complementarity. IEEE Journal of Selected Topics in Signal Processing. 2(3):261–274.

Peng H, Long F, Ding C. 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence. 27:1226–1238.

Pierleoni A, Martelli PL, Casadio R. 2008. PredGPI: a GPI-anchor predictor. BMC bioinformatics 9(1):1–11.

Quinlan JR. 1986. Induction of decision trees. Machine Learning. 1:81–106.

Roudier F, Schindelman G, DeSalle R, Benfey PN. 2002. The COBRA family of putative GPI-anchored proteins in *Arabidopsis*. A new fellowship in expansion. Plant Physiology. 130(2):538–48.

Showalter AM, Keppler B, Lichtenberg J, Gu D, Welch LR. 2010. A bioinformatics approach to the identification, classification, and analysis of hydroxyproline-rich glycoproteins. Plant Physiology. 153:485–513.

Simonović A, Filipović B, Trifunović M, Malkov S, Milinković V, Jevremović S, Subotić A. 2015. Plant regeneration in leaf culture of *Centaurium erythraea* Rafn. Part 2: the role of arabinogalactan proteins. Plant Cell, Tissue and Organ Culture. 121:721–739.

Simonović A, Dragićević M, Bogdanović M, Trifunović-Momčilov M, Subotić A, Todorović S. 2016. DUF1070 as a signature domain of a subclass of arabinogalactan peptides. Archives of Biological Sciences 68:737–746.

Simonović AD, Trifunović-Momčilov M, Filipović BK, Marković MP, Bogdanović MD, Subotić AR. 2021. Somatic embryogenesis in *Centaurium erythraea* Rafn—current status and perspectives: a review. Plants. 10(1):70.

Trifunović M, Tadić V, Petrić M, Jontulović D, Jevremović S, Subotić, A. 2014. Quantification of arabinogalactan proteins during in vitro morphogenesis induced by β-d-glucosyl Yariv reagent in *Centaurium erythraea* root culture. Acta Physiologiae Plantarum. 36:1187–1195.