ragp: Pipeline for mining of plant hydroxyproline-rich glycoproteins with implementation in R

Running title: HRGP filtering pipeline using the ragp R package

Milan B. Dragićević*, Danijela M. Paunović, Milica D. Bogdanović, Slađana I. Todorović and Ana D. Simonović

Institute for Biological Research "Siniša Stanković", Department of Plant Physiology, Bul. Despota Stefana 142, University of Belgrade, 11000 Belgrade, Serbia

*To whom correspondence should be addressed: mdragicevic@ibiss.bg.ac.rs, tel +381643864319

Supplementary data submitted:

Supplement 1. 21-mer train set (train_21.csv file)

Supplement 2. 21-mer test set (test_21.csv file)

Supplement 3: 15-mer train set (train_15.csv file)

Supplement 4: 15-mer test set (test_15.csv file)

Keywords: arabinogalactan, glycoprotein annotation, HRGP, hydroxyproline-prediction, machine learning

Abstract

Hydroxyproline-rich glycoproteins (HRGPs) are one of the most complex families of macromolecules found in plants, due to the diversity of glycans decorating the protein backbone, as well as the heterogeneity of the protein backbones. While this diversity is responsible for a wide array of physiological functions associated with HRGPs, it hinders attempts for homology based identification. Current approaches, based on identifying sequences with characteristic motifs and biased amino acid composition, are limited to prototypical sequences. ragp is an R package for mining and analysis of HRGPs, with emphasis on arabinogalactan proteins. The ragp filtering pipeline exploits one of the HRGPs key features, the presence of hydroxyprolines which represent glycosylation sites. Main package features include prediction of proline hydroxylation sites, amino acid motif and bias analyses, efficient communication with web servers for prediction of N-terminal signal peptides, glycosylphosphatidylinositol modification sites and disordered regions and the ability to annotate sequences through hmmscan and subsequent GO enrichment, based on predicted Pfam domains. As such, ragp extends R's rich ecosystem for high-throughput sequence data analyses. The ragp R package is available under the MIT Open Source license and is freely available to download from GitHub at: https://github.com/missuse/ragp.

Introduction

Hydroxyproline-rich glycoproteins (HRGPs) comprise a superfamily of diverse plant cell wall O-glycosylated proteins, ubiquitous in the plant kingdom, with a vast array of functions associated (Ellis et al. 2010, Showalter et al. 2010). HRGPs constitute almost 10% of the cell walls dry weight; embedded into cellulose/hemicellulose and pectic polysaccharides networks, they are sometimes referred to as the third network of the plant cell wall (Nguema-Ona et al. 2014). The protein backbones of HRGPs feature different Pro-rich motifs that govern hydroxylation of Pro to hydroxyproline (Hyp, O) as sites of subsequent O-glycosylation (Johnson et al. 2017). Based on the type and degree of glycosylation, HRGPs have been divided into three multigene families: highly glycosylated arabinogalactan proteins (AGPs), moderately glycosylated extensins (EXTs) and non-, weakly or highly glycosylated proline rich proteins (PRPs) (Hijazi et al. 2014, Showalter et al. 2010). AGPs are involved in cell proliferation and expansion, reproductive development, embryonic patterning, growth of roots, root hairs and pollen tubes, secondary wall deposition, xylem differentiation, programmed cell death, hormone responses, abscission, abiotic and biotic stress responses as well as morphogenesis *in vitro*, including somatic embryogenesis (Ellis et al. 2010, Seifert and Roberts 2007, Simonović et al. 2015, Tan et al. 2012). Most but not all AGPs are attached to the plasma membrane by glycosylphosphatidylinositol (GPI) membrane anchor, thus connecting plasma membrane with cytoskeleton and cell wall elements, and providing a mechanism by which AGPs may be involved in signaling and defining cell shape and polarization (Ellis et al. 2010, Seifert and Roberts 2007). EXTs are structural proteins able to form covalent scaffolds – intra and inter-molecular network - within the cell wall (Hijazi et al. 2014, Nguema-Ona et al. 2014). Little is known about the function of PRPs; it seems that they are implicated in defense against biotic and abiotic stresses, particularly in active growing cells (Battaglia et al. 2007, Hijazi et al. 2014).

Pro hydroxylation of HRGPs occurs in the secretory pathway by the action of proline hydroxylases, to be followed by O-glycosylation and glycan processing (Nguema-Ona et al. 2014). According to Hyp

contiguity hypothesis, the type of glycosylation can be predicted based on Hyp-containing motifs found in HRGPs (Ellis et al. 2010, Johnson et al. 2017, Tan et al. 2012). Namely, blocks of contiguous Hyp residues, commonly preceded by Ser (SO_{3-5}) are glycosylated with oligoarabinosides linked to Hyp, and a single galactose residue linked to Ser; these are typical extensin glycomodules, but may occur in some AGPs as well. AGPs' protein backbones are primarily decorated with branched type II arabino-3,6-galactans which are O-linked to noncontiguous Hyp residues found as clustered dipeptides (AG-II glycomodules): AO, SO, TO, VO, GO, OA and some others. The extent of glycosylation of PRPs has not been extensively studied (Johnson et al. 2017). PRPs typically contain shorter stretches of contiguous P, glycosylated with arabinose oligosaccharides. There are also hybrid HRGPs that feature glycomodules from different HRGP classes, chimeric HRGPs composed of HRGP modules and non-HRGP protein domains, as well as very small proteins such are arabinogalactan (AG) peptides (Schultz et al. 2002). Thus, the HRGP superfamily is often considered as a spectrum or a continuum of different glycoproteins with varying degree and types of glycosylation (Ellis et al. 2010, Johnson et al. 2017, Showalter et al. 2010). Recently, Johnson et al. (2017) classified HRGPs not into 3, but into 23 descriptive sub-classes. Considering the variety of physiological and structural roles of HRGPs and their abundance in all plants, mining for HRGPs in plant genomes and transcriptomes, as a prerequisite for their further study, is an important goal in plant glycobiology. However, this goal is rather challenging, as it does not rely on simple homology search, for several reasons. Namely, HRGPs have characteristic features of intrinsically disordered proteins (proteins without well-defined three dimensional structure): they are rich in most disorder-promoting amino acid – proline (due to its rigid conformation), they contain repeated sequence motifs (especially EXTs and PRPs) and are extensively glycosylated (Johnson et al. 2017). The lack of a stable structure and hydrophobic core lessens the sequence constraints imposed on these proteins, so that they can rapidly mutate and evolve, thus hampering the efforts for homology based identification (Johnson et al. 2017). AGPs have particularly low sequence similarity, and their diagnostic

dipeptide glycomodules are scattered throughout the protein backbone, so mining for AGP sequences using homology searches typically identifies only a few closely related family members (Schultz et al. 2002, Showalter et al. 2010, Tan et al. 2012), usually those with conserved domains (Simonović et al. 2015). Therefore, the search for HRGPs is commonly based on (combination of) their key features:

1) The presence of N-terminal signal peptide (N-sp) is a common feature of HRGPs, since they are synthetized and post-translationally modified in the secretory pathway.
2) The amino acid composition of HRGPs is biased towards disorder-promoting residues, particularly Pro, as well as residues that comprise glycomodules. Thus, classical AGPs that are rich in Pro (P), Ala (A), Ser (S) and Thr (T) can be identified as sequences that have amino acid composition with more than 50% PAST or, for AG peptides, more than 35% PAST (Schultz et al. 2002, Showalter et al. 2010). Likewise, PRPs are characterized by amino acid composition with greater than 45% PVKCYT (Showalter et al. 2010) or PVKY (Johnson et al. 2017), while EXTs bias is defined as >45% PSKY (Johnson et al. 2017).
3) The presence of a glycosylphosphatidylinositol anchor signal peptide (GPI-sp), a hydrophobic C-terminal region that signals the addition of GPI, is a feature of many but not all AGPs and AG peptides and some other HRGPs (Ellis et al. 2010, Showalter et al. 2010, Simonović et al. 2016).
4) Motifs useful for mining HRGPs are virtually limited to extensin $SO_{3\text{-}5}$ glycomodules, but some EXT may in addition have Y-based cross linking motifs (Johnson et al. 2017, Showalter et al. 2010). AG-II glycomodules, being scattered dipeptides, were not often used for sequence mining, but are used by specific approaches (Ma et al. 2017). Finally, several known PRPs have PPVX(K/T) and KKPCPP motifs (Showalter et al. 2010).
5) Conserved domains that are exclusively found in HRGPs are not known, except for arabinogalactan peptide domain (PF06376, formerly DUF1070) that we have recently identified (Simonović et al. 2016). This domain is found at the C-terminus of some AG peptides and most of it represents GPI-sp,

which is cleaved during the processing. However, non-exclusive HRGP domains present in chimeric HRGPs, and particularly in chimeric AGPs, such as fasciclin (PF02469), ns-LTP-like (PF00234), plastocyanin-like (PF02298) and several others are useful when performing homology based HRGP mining.

Current approaches to mining HRGPs are based on some or all of the abovementioned features, and are commonly organized as pipelines for filtering protein sequences. First of such pipelines was BIO OHIO (Showalter et al. 2010), relying on all of the above HRGPs' features, combined with expression analysis of HRGPs and enzymes involved in their synthesis. Motif and amino acid bias (MAAB) bioinformatics pipeline (Johnson et al. 2017) was recently developed not only for mining, but also for classification of HRGPs into 23 descriptive sub-classes. Finally, Ma et al. (2017) proposed a decision-based approach (Python script "Finding-AGP") dependent on several values derived from the total and partial amino acid composition and protein length, which enabled filtering potential chimeric AGPs with as low as three AG glycomodules. Even though developed for comprehensive search, these pipelines may miss short sequences which do not exhibit pronounced amino acid bias and that contain few characteristic motifs, such as chimeric AGPs and AG peptides. Domain search can capture chimeric HRGPs only if they contain domains already known to associate with HRGPs, but not novel chimeric combinations. More importantly, none of these pipelines adopts the key HRGPs' feature – the presence of hydroxylated proline. The protein sequence code for Pro hydroxylation has been studied for over two decades, by comparing amino acid context of Pro with its actual hydroxylation or glycosylation status in different proteins, by analyzing proline hydroxylase crystal structure and its active site, by using recombinant hydroxylase substrates (sporamin) and by other approaches (Ellis et al. 2010). These efforts provided general rules for the Pro hydroxylation code, but many ambiguities remain.

Hereby we present a pipeline for mining HRGPs implemented in the R package ragp, which is distinguished from other pipelines described in the literature (Johnson et al. 2017, Ma et al. 2017,

Showalter et al. 2010) by the assumption that only particular/specific prolines in a protein can be hydroxylated as a necessary prerequisite for glycosylation. Inference on the positions of these prolines is accomplished by a machine learning model trained on plant sequences with experimentally determined hydroxyprolines from the UniProtKB/Swiss-Prot data base. The Pro hydroxylation prediction is combined with previously described standard HRGP mining tools, but also with domain annotation with GO enrichment, HRGP classification via MAAB and disordered region prediction. In order to place the mentioned sequence features in a visual context, ragp also provides resources for schematic plotting of protein features. Overall, ragp is a freely available, comprehensive, fast and customizable pipeline for HRGPs filtering, annotation, classification and graphical presentation.

Results

Hydroxyproline prediction model performance

The central element of ragp workflow is the prediction of hydroxyproline positions in plant proteins. In order to train a robust model, four machine learning (ML) algorithms were compared in terms of their prediction performance: k-nearest neighbors (knn), random forest (rf, Breiman 2001), support vector machines with radial basis function kernel (svm) and gradient boosting by xgboost (xgb, Chen and Guestrin 2016). These algorithms were trained on plant protein sequences with experimentally determined hydroxyprolines, or more specifically a classification task was trained on local 21-mer sequences centered on the target prolines/hydroxyprolines. Since the feature set constructed to describe these local 21-mer sequences contained 1294 unique features split across 16 feature groups (Table I), several approaches to feature selection were attempted: filter selection using information gain ratio (IGr, Quinlan 1986), filter selection using minimum redundancy maximum relevance (mRMR) criterion (Peng et al. 2005), as well as wrapper selection via sequential forward search (sfs, Kohavi and John 1997) which operated not on the level of individual features but over the 16 feature sets.

To reduce bias during model selection we used nested cross validation (CV) where the outer loop was used to estimate model performance and the inner loop was used to tune hyper parameters by model based optimization (MBO). The performance of the models was scored using mean area under the receiver operating characteristic (ROC) curve (AUC) on the hold out instances of the outer nested CV loop.

The performance of rf and xgb algorithms was competitive and noticeably higher compared to knn and svm (Figure 1) regardless of the feature selection approach used. Svm offered the poorest performance in terms of AUC in all cases, and especially when mRMR filter selection and no feature selection were applied. In case of knn, rf and xgb the performance improvement when using any type of feature selection was modest compared to fitting the models on all of the 1294 features. For all four algorithms the peak performance was achieved when sequential forward selection was used, with xgb performing slightly better (mean ± sd: 0.982 ± 0.033 AUC) compared to rf (0.976 ± 0.039 AUC), followed by knn (0.972 ± 0.029) and svm (0.965 ± 0.043 AUC). Therefore, the algorithm used to build the Hyp prediction model incorporated in ragp was xgb along with sequential forward feature selection over 16 feature sets. The feature sets selected by sfs for constructing the model were F1, F3, F4 and F10 (Table I) resulting in 308 unique features on which the model was trained. To further evaluate this model and specifically to tune the decision threshold, an additional nested CV was performed using F1, F3, F4 and F10 feature sets without feature selection. Using this resampling setup the xgb model obtained a mean AUC score of 0.978 in the outer loop. AUC is a useful metric to evaluate models, especially in the case of imbalanced classification problems since it is independent of the decision threshold. For the model to be used in production, it is paramount to anticipate how the decision threshold affects its performance. For binary scoring classifiers the decision threshold controls how predicted posterior probabilities are converted into class labels. With an aim to optimize the decision threshold, the relationship between several model evaluation metrics and the threshold value was examined based on the hold out

predictions in the mentioned nested CV. The performance measures considered can be defined in terms of components of the confusion matrix: true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN):

- Sensitivity or the true positive rate:

- Specificity or the true negative rate:

- Accuracy (ACC)

- Balanced Accuracy (BACC)

- Matthews correlation coefficient (MCC, Matthews 1975):

$$\frac{}{\sqrt{(TP \quad FP)\, x\, (TP \quad FN)\, x\, (TN \quad FP)\, x\, (TN \quad FN)}}$$

- Cohen's kappa (kappa, Cohen 1960):

where

There is no single optimal decision threshold, and different performance measures favor different cutoff values (Figure 2). In terms of ACC, MCC and kappa the model performs quite stable with a plateau ranging from 0.3 to 0.7 probability cutoff, while BACC is maximized at in the range 0.2 - 0.4. We chose to pick the default threshold for the model based on BACC and it is set at 0.224 in ragp. Based on the performed nested CV, the mean sensitivity at this threshold was 0.951 at 0.925 specificity, while the median sensitivity was 1 at 0.955 specificity. Another notable threshold is the 0.95 specificity cutoff at 0.364, at which the mean sensitivity was 0.916. It should be mentioned that ragp users are free to change the threshold in order to meet their own stringency criteria.

The final model was tuned using two times repeated 3-fold CV and MBO of hyper parameters for 100 iterations on the whole train set. The proposed hyper parameters: nrounds = 758, min_child_weight = 1.028, max_depth = 15, eta = 0.005, gamma = 0.527, colsample_bytree = 0.801, subsample = 0.825, alpha = 0.707, lambda = 1.61 and colsample_bylevel = 0.986, were used to create a model using the whole train set (Supplement 1) which was evaluated using independent data (test set sequences as described in Data preparation, Supplement 2) which was not used in any way during the model building. Using the default thresholds as set in ragp package (0.224) the model obtained 0.938 sensitivity and 0.971 specificity (with an AUC of 0.986) when applied on the test set sequences. The limitation of this model is that it is unable to predict hydroxylation for P which are within 10 N- or C- terminal amino acids since 21-mers were used for feature creation. For the current application, the hydroxylation of N-terminal P is of little interest, since HRGPs contain N-sp usually longer than 10 amino acids. Thus, there are no experimentally verified hydroxyprolines on the N-terminal side of secreted proteins. However prediction of P hydroxylation on the C-terminal side would be beneficial. In order to accomplish this, we examined the impact of k-mer length on model performance by constructing the same type of feature sets as for 21-mers (F1, F3, F4 and F10) using the appropriate shorter k-mer lengths (Figure 3). Models constructed using k-mer lengths of 19, 17, and 15 amino acids provide similar performance to the model

trained on 21-mers, while a notable decline in performance was observed with 13-mers (Figure 3). Therefore we chose to utilize the model constructed using 15-mers as the supplementary model in ragp package which is used to predict C-terminal hydroxyprolines. As for the 21-mer trained model, the dependence of several model evaluation metrics and the threshold value was examined based on the hold out predictions in nested CV (Figure 4). Using this type of evaluation, the mean sensitivity at the BACC maximizing decision threshold (0.22) was 0.946 at 0.936 specificity. Another notable threshold is the 0.95 specificity cutoff at 0.333 at which the mean sensitivity was 0.925. It could appear that the model trained using 15-mers performs slightly better compared to the model trained using 21-mers, however we are reluctant to make such a claim because the 15-mer model was trained on a smaller data set (35 k-mers were removed compared to the 21- mer train set due to removal of duplicated sequences and homolog reduction) in which several “hard to predict” instances were removed due to similarity.

The final 15-mer model was tuned using two times repeated 3-fold CV and MBO of hyper parameters for 100 iterations and the proposed hyper parameters: nrounds = 379, min_child_weight = 1.357, max_depth = 5, eta = 0.031, gamma = 1.882, colsample_bytree = 0.749, subsample = 0.589, alpha = 0.457, lambda = 2.329 and colsample_bylevel = 0.98, were used to create a model on the whole 15-mer train set (Supplement 3) which was evaluated on the 15-mer test set (Supplement 4) obtaining 0.962 sensitivity and 0.952 specificity (with an AUC of 0.986) at the default decision threshold of 0.22.

The performance on the test set sequences for both 21-mer and 15- mer models was compared to established hydroxyproline prediction servers RF-Hydroxysite (Ismail et al. 2016) PredHydroxy (Shi et al. 2015), iHyd-PseCp (Qiu et al. 2016) and iHyd-PseAAC (Xu et al. 2014) on the corresponding 21-mer and 15-mer test sets (Figure 5 and Table II). When evaluated on the test set sequences PredHydroxy, iHyd-PseAAC and iHyd-PseCp tend to sacrifice sensitivity (0.4 - 0.47) for high specificity (0.84 – 0.99) in prediction. This is likely caused by the under-representation of plant sequences in sets used for the

training of the corresponding machine learning algorithms leading to performance that is mostly driven by proline hydroxylation patterns present in animal protein sequences, which differ significantly to their plant counterparts. One notable exception is RF-Hydroxysite which scored an AUC of 0.955 (Table II, Figure 5) on the 21-mer test set obtaining 0.969 sensitivity and 0.828 specificity at the default stringency (0.6) and maximal k-mer window size of 17. Additionally specificity rose to 0.874 without sacrificing sensitivity when the probability threshold was set to 0.675 which is optimal for the test set used. However the RF-Hydroxysite web server was not designed for high throughput analyses, allowing the analyses of one protein sequence at a time. Based on the test set sequences the models incorporated in ragp ***predict_hyp*** function offer slightly lower sensitivity compared to RF-Hydroxysite web server and much higher specificity at the default stringency, with the benefit of high throughput analysis.

HRGP sequence mining from 62 Phytozome proteomes

The ragp workflow was performed on 62 plant proteomes obtained from the Phytozome V12 database. The annotation data is available for download at Zenodo (doi: 10.5281/zenodo.2605302, url: https://zenodo.org/record/2605302) under the creative commons attribution 4.0 international license. The first filtering step in the ragp workflow is N-sp prediction using three web servers: SignalP4.1, TargetP1.1 (Emanuelsson *et al*, 2007) and Phobius (Käll *et al*. 2007). Based on the predictions, it is apparent that SignalP4.1 is the most conservative algorithm, while TargetP1.1 at the default settings is the most relaxed (Figure 6). Due to the relatively high discrepancy between the predictions of these three algorithms, we used a majority vote to determine if the sequence should pass the N-sp filtering step. Such an approach filtered 266135 out of 2797062 analyzed protein sequences. These potentially secreted protein sequences were subjected to Hyp prediction in the second filtering step using ***predict_hyp*** ragp function with the default model thresholds, and 69982 protein sequences (26.3% of secreted proteins) had three or more predicted Hyp. MAAB classification using the ***maab*** function from

ragp package was performed on all 266135 potentially secreted proteins and the output was examined in order to check how many hydroxyprolines are present in sequences with MAAB classes (Figure 7). A total of 3075 sequences were classified as MAAB classes 1 – 23 (prototypical HRGPs). The green alga *Chlamydomonas reinhardtii* had the highest number (136) of these sequences, while the green alga *Ostreococcus lucimarinus* did not contain any sequences classified into MAAB classes 1 – 23. All MAAB 1 - 23 classified sequences from 51 organisms contained at least 3 predicted hydroxyprolines, while in the remaining 10 organisms only a few (22 in total) MAAB classified sequences were not predicted to contain at least 3 Hyp. It should be noted that no sequences were classified as MAAB classes 13, 14 and 17.

Apart from finding prototypical HRGP sequences by MAAB classification we also performed a scan for AGP motifs in the 266135 sequences predicted to be secreted using the ***scan_ag*** ragp function. Only predicted Hyp positions were considered when searching for AGP motifs. 36732 protein sequences from 62 plant species were found to contain at least one AGP motif span which was defined by having at least three dipeptides (AO, TO, SO, GO, VO, OA, OT, OS, OG and OV) separated by a maximal of 10 amino acids between any two dipeptides. To identify potential hybrid AGPs, hmmer3 software was used with Pfam 32 data base on these protein sequences (Figure 8). The most frequent identified domains in AGP motif containing sequences are the protein kinase (PK) and protein tyrosine kinase (PTK) domains which are jointly identified and overlapping in the same sequences (Figure 8A), they are followed by leucine rich repeat domains (LRR_8, LRRNT_2 and LRR_4) which are often found with PK/PTK domains in the same sequences. Domains such as plant lipid transfer proteins (Tryp_alpha_amyl and LTP_2), plastocyanin-like (Cu_bind_like) and fasciclin are known to be a part of chimeric AGPs, while others such as X8 and Glycoside hydrolase family 17 (Glyco_hydro_17) have recently been proposed to be affiliated with AGPs based on sequence analyses (Ma et al., 2017). It is noteworthy that the most frequently identified potential chimeric AGP domain - the PK/PTK domain has eluded experimental evidence for

linkage with AGPs in the literature. Structure of several of the mentioned PTKs (Figure 9), visualized using the ragp function ***plot_prot***, suggests that the Hyp containing arabinogalactan motifs are on the extracellular side while the kinase domains are on the intracellular side. These PTKs often contain leucine rich repeat domains on the extracellular side. Domains subtilisin-like (Peptidase_S8), peptidase inhibitor I9 (Inhibitor_I9), fibronectin type-III (fn3_6) and protease associated (PA) are found on the same protein sequences which belong to the subtilase family of serine proteases. The majority of subtilase family sequences would have been removed by slightly increasing the filtering stringency (increasing the number of AGP motifs to four instead of three); these sequences have usually only several predicted Hyp (Figure 8B) and thus might be false positives.

Discussion

Predicting hydroxyproline positions

The key innovation of the ragp workflow in HRGP mining and analysis is the incorporation of a Hyp prediction ML model. During model selection and training we set several goals: the model should be able to generalize, model evaluation should illustrate the true model performance and the predictions should be rapid so high throughput workflows would be possible.

Four ML algorithms were evaluated knn, svm, rf and xgb, with svm and rf already previously utilized to build Hyp prediction models (Shi *et al*. 2015, Qiu *et al*. 2016, Ismail *et al*. 2016). One of the first decisions when considering building a model based on protein sequences is how to numerically encode the sequence information. Protein sequences can be numerically encoded in a variety of ways and many of the resulting features are mutually highly correlated or provide no information to the ML algorithm about the task at hand. For the current task we chose a relatively broad set of sixteen sequence descriptor sets (Table I) with a total of 1294 unique features; however they represent only a small portion of all possibilities. Since it is likely many of these features represent no value to the model, and

in a best case scenario would just serve to prolong the computation time needed for predictions, several feature selection methods were applied: two filter methods, one using information gain ratio, and the other using minimum redundancy maximum relevance, as well as a wrapper method sequential forward search. For clarity, the difference between these types of feature selection methods will be mentioned. Filter methods operate by assigning an importance value to each feature based on some metric (IGr and mRMR were used here) which is external to the ML algorithm being trained. Based on these values the features can be ranked and a feature subset of the top ranking features can be selected. Wrapper methods select features based on the ML algorithm performance during resampling, in short by using sfs in the current setup first the performance of the ML algorithm was assessed using individual feature sets and the feature set providing the highest performance was chosen, then the algorithm performance was evaluated using the selected feature set and each of the remaining ones; this proceeded until the performance started to decline with the addition of feature sets. Sequential forward search was performed over feature sets (Table I) and not over individual features for two main reasons: 1. computation cost is greatly reduced when 16 feature sets are used as opposed to 1294 features; 2. two of the utilized ML algorithms – rf and xgb perform internal feature selection during model fitting, therefore a fine grained wrapper selection would most likely not benefit these two algorithms.

In order to reduce bias when evaluating model performance special consideration was taken when constructing the resampling procedure:

1. When modeling data derived from biological sequences a usual step is removal of homolog sequences to reduce the overestimation of prediction accuracy in cross validation. This overestimation is caused by the fact that highly homologous sequences (k-mers in this case) can be found both in the training and the hold out instances during cross validation. However, removal of homologs inevitably leads to loss of information, especially in the current application since many k-

mers from the same protein sequence are highly overlapping. To ameliorate these issues we used two complementary approaches:

1.1. We performed removal of homolog k-mers based on Levensteins distance in a stepwise manner so that no two k-mers share more than 90% homology (in the 21-mer set, and slightly less for shorted k-mers). Further homolog removal (< 90 % homology) greatly reduced the training set and increased class imbalance so we decided against it.

1.2. Protein blocked k-fold cross-validation where all k-mers from the same protein are either used for model building or hold out predictions during cross validation was used for tuning and evaluation of model performance. In our opinion this produces a more unbiased estimation of performance, compared to non-blocked resampling, closely resembling the model use case scenario.

2. In order to obtain truthful performance estimates for a learner, all parts of model building should be included in the resampling (Cawley and Talbot 2010, Varma and Simon 2006). In this study nested cross validation was performed to estimate model performance. The inner loop was used to tune the algorithm hyper parameters while the outer loop was used to estimate the performance. Such an approach could introduce some positive bias in the case of wrapper selection since the outer resampling loop is not used solely for the purpose of estimating performance but is also used to drive the feature selection. In case of wrapper selection a completely unbiased evaluation approach would require three nested resampling loops: inner loop for hyper parameter tuning, middle loop for feature selection and outer loop for model evaluation. Due to computation cost this was not performed in the current study.

All of the feature selection methods compared resulted in a slight accuracy gain when contrasted to models without feature selection for all tested algorithms. The highest performing model based on the resampling performance was constructed using sfs coupled with xgb algorithm (Figure 1). This model

was thoroughly evaluated by additional nested cross validation using previously selected features (Figure 2) and by using a test set (Figure 5), and the results show it obtains state-of-the-art performance for the task at hand. This isn't to say the model cannot be improved by using another algorithm, set of features, set of hyper parameters or by stacking several models. However in our opinion the biggest improvement in model generalization would be achieved if more sequences and thus a higher diversity of sequences are used for training. As the pool of plant sequences with experimentally determined Hyp positions increases the model generalization power will as well.

ragp workflow: hydroxyproline aware HRGP filtering and analysis

The key feature of HRGPs is the presence of hydroxyprolines which represent glycosylation sites (Showalter et al. 2010). While many HRGP sequences can be mined based on biased amino acid composition, or the presence of certain amino acid motifs, there exists an abundance of chimeric proteins comprised from specific domains and HRGP motifs which are much harder to identify based on the mentioned features. This is especially true for AGPs since they do not have well defined motif contexts – it is not known how many AG glycomodules are required and how far apart in the sequence are they allowed to exist for the motif to be glycosylated. Mining for such sequences would be straightforward if the positions of hydroxyprolines were known in advance, which is unfortunately not possible due to the high disproportion in cost and effort between nucleic acid and protein sequencing. A possible solution is to utilize currently available knowledge to predict hydroxyproline positions in proteins sequences deduced from next generation sequencing experiments. ragp workflow exploits this idea by incorporating a machine learning model to infer probable Hyp sites in protein sequences. The ragp workflow (Figure 10) consists of two layers: the filtering layer in which hydroxyproline containing secreted proteins are filtered and the analysis layer in which the filtered sequences are analyzed for AGP motifs, classified into MAAB classes (Johnson et al. 2017), domains, disordered regions and potential GPI

attachment sites are predicted. The mentioned ML model is built into the ragp function ***predict_hyp*** and its predictions are utilized by the ragp workflow in two ways:

- Proteins which contain less than a certain number of predicted Hyp are simply filtered out and not considered further. In our proposed workflow we have set the minimal number of predicted Hyp to three, however depending on the goal this can be a larger number or even dependent on sequence length.
- Hyp predictions are used in a hydroxyproline aware scan using ***scan_ag*** ragp function. This function offers users to construct a customizable AGP motif scan by picking the minimal number of motifs needed for a match, the maximum number of amino acids between motifs for a match, the type of amino acids dipeptides for a motif match as well se the ability to mask extensin motifs.

The workflow is explained in greater detail at: https://missuse.github.io/ragp/ with tutorials on HRGP filtering, analysis and protein sequence visualization. Additionally a user friendly web application is available at: https://ragp.shinyapps.io/Rapp/ which offers a graphical interface to ***predict_hyp***, ***scan_ag*** and ***maab*** ragp functions.

While ***predict_hyp***, ***scan_ag*** and ***maab*** represent core ragp functions, the package offers several additional benefits for HRGP sequence analysis. Because HRGPs are secreted proteins, ragp incorporates a fast interface to web servers commonly used for predicting N-sps based on primary protein sequence which is available via the functions ***get_signalp*** (interface to SignalP4.1, http://www.cbs.dtu.dk/services/SignalP-4.1/), ***get_targetp*** (interface to TargetP1.1, http://www.cbs.dtu.dk/services/TargetP/) and ***get_phobius*** (interface to Phobius, http://phobius.sbc.su.se/). For instance by using ***get_signalp*** we were able to obtain N-sp predictions for more than 2.7 million protein sequences in less than one day. Many HRGPs are anchored to the membrane by a GPI anchor and ragp offers fast GPI prediction via the functions ***get_big_pi*** and ***get_pred_gpi*** which represent an interface to big-PI Plant Predictor

(http://mendel.imp.ac.at/gpi/plant_server.html, Eisenhaber et al. 2003) and PredGPI (http://gpcr.biocomp.unibo.it/predgpi/pred.htm, Pierleoni et al. 2008) web servers. After identification of certain sequences of interest, insights into the positions of disordered regions, which in many instances overlap with HRGP motifs, is available via ***get_espritz*** function which communicates with Espritz server (http://protein.bio.unipd.it/espritz/, Walsh et al. 2012). Finally ragp offers domain and GO annotation via the functions ***get_hmm*** and ***pfam2go*** which communicate with hmmscan (https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan) and map Pfam accessions to GO terms (http://current.geneontology.org/ontology/external2go/pfam2go). In combination these functions are available for a wide array of protein sequence annotation tasks and can be utilized for protein function prediction.

It should be noted that we designed ragp prioritizing ease of use: the package functions have a consistent input and accept a range of input objects - from basic R data structures to FASTA files. The default values for additional function arguments were carefully chosen, and in most use cases can be left as is. The function outputs are also basic R data structures: data frames or lists, which can be manipulated by many popular R packages, according to user preference.

The ragp workflow was performed on predicted protein sequences from 62 plant proteomes to gain insights in both the variability of HRGP sequences in the plant kingdom as well as to gain impressions of the workflow itself. The annotation data is available for download at https://zenodo.org/record/2605302. The hydroxyproline aware workflow identifies 99.29 % of prototypical HRGP sequences (MAAB classes 1 – 23) compared to performing MAAB classification without Hyp prediction, capturing all prototypical HRGP sequences in 51 out of the 62 analyzed plant proteomes. On the other hand a great number of non-prototypical HRGP sequences are implied by the workflow, for instance around 30000 potential chimeric AGPs were found in the 62 analyzed plant proteomes suggesting that O-glycosylation with arabinogalactan chains is very common in plants. The

fact that the most frequently identified domain in these chimeric AGPs is the receptor tyrosine kinase domain suggests that the heavily glycosylated extracellular regions of these proteins serve to perceive signals in the cell wall which are in turn sent as information to the cell by the action of the tyrosine kinase domain. This is coherent with the proposed role of chimeric AGPs as extracellular sensors (Showalter and Basu 2016). However based on the detected chimeric AGP PTK sequences, and the sheer amount of these sequences is plant proteomes, it seems AGPs not only interact with extracellular regions of receptor kinases as discussed in Showalter and Basu (2016) they are also a constitutive part of these proteins.

Conclusion

ragp represents the first implementation of a HRGP mining workflow in the R statistical language. It implements common strategies for finding and classifying HRGP sequences along with an additional step in which proline hydroxylation is estimated, which leads to increased specificity of the filtered sequences. Since R is one of the leading bioinformatics platforms, the filtered sequences can be further analyzed by many specialized packages using the same environment.

Materials and methods

Predicting proline hydroxylation

Data preparation. With the aim to train a machine learning algorithm to predict the probability of proline hydroxylation in plant proteins, we initially acquired 40 plant protein sequences with experimentally determined hydroxyprolines from the manually curated UniProtKB/Swiss-Prot data base (UniProt release 2017_07, www.uniprot.org) (The UniProt Consortium 2017). After removal of non-sequenced regions the ratio Hyp/Pro in this set was close to two. Since we trust this ratio does not reflect the true ratio of these amino acids in secreted plant proteins, an additional set of 269 plant

sequences from the UniProtKB/Swiss-Prot data base with experimentally determined N-sp and without reported hydroxyprolines was included. This Hyp-negative group included only secreted proteins, in order to train the model on sequences that closely resemble ones on which the model would be used. The literature describing the sequencing for each protein was thoroughly checked and cross referenced to the UniProtKB/Swiss-Prot annotation and non-sequenced and ambiguous regions (which had discrepancies between different sources) were removed in order to minimize introduction of false labels that could occur if some sequences contained Hyp which escaped detection, or if they alternatively had falsely labeled Hyp. After this local 21-mer sequences - ±10 amino acids around the target Hyp/Pro were extracted and duplicated 21-mers were removed. Redundancy removal (homolog reduction) was performed based on Levensteins distance (count of the minimal number of amino acid substitutions, insertions or deletions required to turn one k-mer sequence to another) calculated using the R package stringdist (van der Loo 2014). Redundancy removal was performed separately for the Hyp positive and Hyp negative group of 21-mers, and proceeded in a stepwise manner by eliminating sequences that differed from others in exactly 1 position based on Levensteins distance. Elimination was performed one 21-mer sequence at a time by removing the sequence which had the maximum number of homologs and re-evaluation after each 21-mer sequence was removed similarly as in Schwartz et al. (2009). This resulted in a set of 225 protein sequences with 1093 21-mers which shared at most 90% sequence identity around the target Hyp/Pro. These 225 protein sequences were then split at random to 80% train set (181 unique protein sequences with 150 hydroxylated sites and 737 non hydroxylated sites) and 20% test set (44 unique protein sequences not present in the train set with 32 hydroxylated sites and 174 non hydroxylated sites). The 21-mer train and test sets are provided as Supplement 1 and Supplement 2. In order to evaluate the impact of k-mer length on model performance, the 21-mer sequences were reduced to 19, 17, 15 and 13-mers centered on the target Hyp/Pro and for each group homolog reduction was performed as described above for the 21-mer set. The 15-mer train and test sets which

were used to create and evaluate the supplementary model in ragp package are provided as Supplement 3 and Supplement 4.

Feature engineering. A huge number of various numerical representations can be used to encode protein sequences yielding a very high dimensional modeling task. For the current task, 16 feature sets were investigated. The first feature set (F1) was constructed by one-to-one mapping of six common uncorrelated amino acid physicochemical properties (Normalized Average Hydrophobicity - CIDH920105, Average Flexibility Indices - BHAR880101, Free Energy of Solution in Water [kcal/mole] - CHAM820102, Residue Volume - BIGC670101, Steric Parameter - CHAM810101 and Relative Mutability - DAYM780201) to amino acids surrounding the target Hyp/Pro resulting in a feature set of 120 dimensions (6 properties • 20 amino acids, 10 on each proline side in 21-mers). The second feature set (F2) was constructed by one to one mapping of five multidimensional patterns obtained by multiple dimension scaling of amino acid physicochemical attributes (Atchley et al. 2005) to amino acids in 21-mers resulting in a feature set of 100 dimensions (5 properties • 20 amino acids). Features F3 – F8 represent auto-correlation descriptors: normalized Moreau-Broto auto-correlation descriptors calculated from the physicochemical properties used for F1 (F3) and from the multidimensional patterns used for F2 (F4); Moran auto-correlation descriptors (F5 calculated from attributes used for F1 and F6 calculated from attributes used for F2); Geary auto-correlation descriptors (F7 calculated from attributes used for F1 and F8 calculated from attributes used for F2). The dimensions of these feature sets are equal to no. attributes • lag. Features F9 – F10 represent sequence-order descriptors calculated based on two physicochemical distance matrices: Schneider-Wrede (Schneider and Wrede 1994) and Grantham physicochemical distance matrix (Grantham 1974): sequence-order-coupling number (Chou 2000) (F9) and Quasisequence- order descriptors (Chou 2000) (F10) with weighting factor set to 0.1. The remaining feature sets consisted of: Conjoint Triad Descriptor (F11) (Shen et al. 2007), Pseudo-Amino

Acid Composition (Chou 2001) with lambda 20 (F12), Amphiphilic Pseudo-Amino Acid Composition (Chou 2005) with lambda 20 (F13), Composition (F14), Transition (F15) and Distribution (F16) descriptors (Dubchak et al. 1995). The feature sets are summarized in Table I. Features F3 – 16 were constructed utilizing the R package protr (Xiao et al. 2015) while F1 and F2 were constructed using code developed for ragp package.

Model training. The performance of several machine learning algorithms to correctly predict Pro hydroxylation was evaluated: random forests (Breiman 2001) as implemented in the R package ranger (Wright et. al. 2017), gradient boosted trees as implemented in the R package xgboost (Chen at al. 2018), support vector machines with radial basis function kernel as implemented in the R package e1071 (Meyer et al. 2018) and k-nearest neighbors as implemented in the R package kknn (Schliep and Hechenbichler 2016). The training of these algorithms was performed using the mlr package (Bischl et al. 2016) framework. To explore the hyper parameter space of the mentioned algorithms, model-based optimization, also known as Bayesian optimization, was performed utilizing the R package mlrMBO (Bischl et al. 2017). For rf three hyper parameters were optimized: number of threes grown in the range 50 - 2000; sampled data fraction for each tree 0.1 - 1; the number of features randomly selected for each split (mtry) 1 - 20; sampling was performed with replacement. For svm two hyper parameters were optimized: cost of constraints violation in the range 10^{-6} - 10^{6}; RBF kernel width (gamma) in the range 10^{-6} - 10^{6}; additionally for svm the data were scaled to zero mean and unit variance in each training instance of model evaluation. For xgb ten hyper parameters were optimized: number of trees grown (nrounds) 50 – 2000; step size shrinkage (eta) 0.005 - 0.2; maximum tree depth (max_depth) 3 - 15; subsample ratio of columns when constructing each tree (colsample_bytree) 0.3 – 1, subsample ratio of columns for each split (colsample_bylevel) 0.3 – 1; subsample ratio of the training instance (subsample) 0.3 – 1; regularization parameters lambda, alpha and gamma 0 – 3; minimum sum of instance weight

(hessian) needed in a child (min_child_weight) 1 – 10. For knn two hyper parameters were optimized: Number of neighbors considered (k) 1-50; and the parameter defining the Minkowski distance 0.5 – 8, additionally the data were scaled to zero mean and unit variance in each training instance of the knn model evaluation. For every algorithm except knn which does not support class weights, the balance of positive and negative class weights was set to the ratio of negative to positive cases in the train set - 737/150 for 21-mers. To obtain solidly performing hyper parameter combinations MBO was performed for 100 iterations, using mean AUC of the hold out folds in cross validation as selection metric. Kriging was utilized as secondary learner used to propose new hyper parameter combinations during MBO (default in mlrMBO package).

Model evaluation and feature selection. In order to optimize the algorithm performance using a highly dimensional feature space (1294 unique features for 21-mer sequences) we compared three types of feature selection; two filter selection approaches based on information gain ratio (Quinlan 1986) and minimum redundancy maximum relevance (Peng et al. 2005), as well as wrapper selection using sequential forward search which operated on the 16 described feature sets (F1 – F16) and not on individual features. In order to select the optimal number of the top ranking features for the filter feature selection methods the absolute number of features selected was tuned jointly with other algorithm hyper parameters in the range 20 - 700. Model evaluation was performed using nested cross validation. The inner cross validation loop which consisted of two times repeated 3-fold cross validation was used for hyper parameter tuning, while the outer cross validation loop used for evaluation of performance consisted of 10-fold cross validation. In all cases protein blocked cross validation was used, where all k-mers from the same protein are either used for model building or hold out predictions during cross validation. Model performance was scored using the AUC metric. Following nested cross validation the highest performing modeling pipeline was additionally evaluated using nested CV with

three times repeated 10-fold CV outer loop and two times repeated 3-fold CV inner loop for hyper parameter tuning via MBO and the predictions on the hold out instances were used to evaluate the impact of decision threshold on several performance metrics: sensitivity, specificity, accuracy, balanced accuracy, Cohen's kappa (Cohen 1960) and Matthews correlation coefficient (Matthews 1975). Finally the chosen algorithm was trained on the whole train set using hyper parameters obtained by MBO for 100 iterations using two times repeated 3-fold CV, and this model was evaluated on the test set sequences using the prechosen decision threshold.

Evaluation of k-mer length. To investigate the impact of reducing the size of k-mers on model performance we truncated the 21-mer sequences to 19 (9 on each target P side), 17, 15 and 13 amino acids. After obtaining training sets for each k-mer size stepwise homolog removal was performed as described above for the 21-mer data set. The features and ML algorithm which produced the highest performing model when trained on 21-mers were used to create models based on the mentioned k-mer sizes. These models were evaluated using protein blocked nested cross validation as described previously. The shortest well performing k-mer size was chosen to train a supplementary model for prediction of C-terminal hydroxyprolines. Threshold tuning, final model training and evaluation on the test set sequences were performed as for the 21-mer model.

Communication with external prediction web servers

N-terminal signal peptide and glycosylphosphatidylinositol lipid anchoring prediction. N-sp prediction in ragp is achieved by efficient communication with TargetP, SignalP (Emanuelsson et al. 2007) and Phobius (Käll et al. 2007) web servers using the functions ***get_targetp***, ***get_signalp*** and ***get_phobius***. Phobius is also used to predict locations of transmembrane regions. Prediction of GPI lipid anchoring positions (omega sites) is achieved by querying big Pi plant predictor (Eisenhaber et al. 2003) and

PredGPI (Pierleoni et al. 2008) web servers thorough the functions ***get_big_pi*** and ***get_pred_gpi***. These functions utilize the R package httr (Wickham 2018) to send data to the corresponding servers via the POST method, and the R package xml2 (Wickham et al. 2018) to parse the server response into R data structures.

Domain and disorder prediction. Domain annotation in ragp is achieved via functions ***get_hmm*** which queries hmmscan web server (Finn et al. 2011) and ***pfam2go*** which maps PFAM annotations to GO terms (using http://geneontology.org/external2go/pfam2go mapping). To identify potential disordered regions in proteins ragp contains ***get_espritz*** function which queries ESpritz (Walsh et al. 2012) web server. These functions also utilize httr and xml2 R packages.

HRGP classification and regex based filtering

Classification of prototypical HRGPs in ragp is achieved using the motif and amino acid bias classification scheme (Johnson et al. 2017) through the function ***maab***. The MAAB classification relies on knowledge if the protein is bound to the membrane by a GPI anchor. The ***maab*** function offers the ability to internally query Big Pi Plant Predictor or PredGPI in order to remove ambiguities in classes. For non-prototypical AGPs filtering in ragp is based on finding localized clusters of AG glycomodules using regular expressions, which can incorporate knowledge on hydroxyproline positions in sequences. This is achieved using the function ***scan_ag*** which constructs regular expressions based on user input and allows for customization of the type and number of AG glycomodules, the number of amino acids between AG glycomodules, as well as masking of extensin motifs.

Protein feature visualization

Visualization of protein features acquired via ragp can be achieved using the function ***plot_prot***. This function takes as input protein sequences and outputs a protein structure diagram. Domains, N-terminal signal peptides, transmembrane regions, extracellular and intracellular protein regions, GPI attachment sites, AG glycomodule spans, hydroxyproline positions and disordered regions can be shown. The output is a ggplot2 (Wickham 2016) object which can be further manipulated to customize the theme, colors and plot annotations.

The source code for all ragp functions is avalible at https://github.com/missuse/ragp/tree/master/R

Annotation of 62 plant proteomes

ragp workflow was performed on 62 plant proteomes obtained from Phytozome V12 database (https://phytozome.jgi.doe.gov/pz/portal.html - PhytozomeV12_unrestricted). The filtering and analysis part of the workflow were performed as described in the discussion section. Briefly, sequences predicted to contain an N-sp by a majority vote between Phobius, SignalP 4.1 and TargetP 1.1 servers were filtered. To obtain N-sp predictions from these servers ragp functions ***get_phobius***, ***get_signalp*** and ***get_targetp*** with default settings were used. Hyp positions in these sequences were predicted using ***predict_hyp*** ragp function. MAAB classification was performed (using ***maab*** ragp function) on all N-sp containing sequences and the number of predicted Hyp per sequence was augmented to this classification so that comparison between MAAB classification with and without Hyp prediction can be performed. Arabinogalactan motifs were identified using ***scan_ag*** ragp function. Criterion for a motif was the presence of at least three dipeptides AO, SO, TO, GO, VO, OA, OS, OT, OG and OV which are no more than 10 amino acids apart, where O are the predicted Hyp in the sequences. For the motif count extensin motifs (spans of at least three P or O) were masked. Domains present in AGP motif containing sequences were detected using hmmer3 3.1b2 (http://hmmer.org/download.html) using Pfam 32 hidden Markov model database (ftp://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam32.0/).

Annotations are available at Zenodo platform for research sharing (doi: 10.5281/zenodo.2605302, url: https://zenodo.org/record/2605302).

Funding

This work was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia [Project numbers TR31019, OI173024]

Acknowledgments

The authors would like to thank Bob Rudis (Rapid7, Boston, MA, USA), Aleksandar Radisavljević (Endava, Belgrade, Serbia) and Thomas Shafee (La Trobe University, Melbourne, Australia) for helpful suggestions during ragp development.

Abbreviations

AGP - arabinogalactan protein

AUC – area under the receiver operating characteristic curve

EXT - extensine

F1 – F16 – Feature sets used for model training (details in Table I)

GPI – Glycosylphosphatidylinositol

GPI-sp - glycosylphosphatidylinositol anchor signal peptide

HRGP – Hydroxyproline-rich glycoprotein

IGr - information gain ratio criterion for feature selection (Quinlan 1986)

knn – k-nearest neighbors machine learning algorithm

MAAB – motif and amino acid bias (an approach to HRGP classification (Johnson et al. 2017))

mRMR – Minimum redundancy maximum relevance criterion for feature selection (Peng et al. 2005)

N-sp – N terminal secretory signal sequence

PRP – proline rich protein

rf – random forest machine learning algorithm (Breiman 2001)

ROC - receiver operating characteristic curve

sfs – sequential forward selection approach to feature selection

svm – support vector machine machine learning algorithm

xgb – xgboost gradient boosting machine learning algorithm (Chen and Guestrin 2016)

References

Atchley WR, Zhao J, Fernandes AD, Drüke T. 2005. Solving the protein sequence metric problem. *Proc Natl Acad Sci USA*. 102:6395-6400.

Battaglia M, Solorzano RM, Hernandez M, Cuellar-Ortiz S, Garcia-Gomez B, Marquez J, Covarrubias AA. 2007. Proline-rich cell wall proteins accumulate in growing regions and phloem tissue in response to water deficit in common bean seedlings. *Planta*. 225:1121-1133.

Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, Casalicchio G, Jones ZM. 2016. mlr: Machine Learning in R. *J Mach Learn Res*. 17:1-5.

Bischl B, Richter J, Bossek J, Horn D, Thomas J, Lang M. 2017. mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions. *http://arxivorg/abs/170303373*.

Breiman L. 2001. Random Forests. *Mach Learn*. 45:5-32.

Cawley GC, Talbot NLC. 2010. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J Mach Learn Res*. 11:2079-2107.

Chen T, Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System. *CoRR* abs/1603.02754:1603.02754.

Chen T, He T, Benesty M, Khotilovich V, Tang Y, Cho H, Chen K, Mitchell R, Cano I, Zhou T, Li M, Xie J, Lin M, Geng Y, Li Y. 2018. xgboost: Extreme Gradient Boosting. R package version 0.71.2. https://CRAN.R-project.org/package=xgb

Chou KC. 2000. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochem Biophys Res Commun*. 278:477-483.

Chou KC. 2001. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins*. 43:246-255.

Chou KC. 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*. 21:10-19.

Cohen J. 1960. A Coefficient of Agreement for Nominal Scales. *Educ Psychol Meas*. 20:37-46.

Dubchak I, Muchnik I, Holbrook SR, Kim SH. 1995. Prediction of protein folding class using global description of amino acid sequence. *Proc Natl Acad Sci USA*. 92:8700-8704.

Eisenhaber B, Wildpaner M, Schultz CJ, Borner GHH, Dupree P, Eisenhaber F. 2003. Glycosylphosphatidylinositol Lipid Anchoring of Plant Proteins. Sensitive Prediction from Sequence- and Genome-Wide Studies for Arabidopsis and Rice. *Plant Physiol*. 133:1691-1701.

Ellis M, Egelund J, Schultz CJ, Bacic A. 2010. Arabinogalactan-Proteins: Key Regulators at the Cell Surface? *Plant Physiol*. 153:403-419.

Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc*. 2:953.

Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res*. 39 (Web Server issue):10.1093/nar/gkr1367.

Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science*. 185:862-864.

Hijazi M, Velasquez SM, Jamet E, Estevez JM, Albenne C. 2014. An update on post-translational modifications of hydroxyproline-rich glycoproteins: toward a model highlighting their contribution to plant cell wall architecture. *Front Plant Sci*. 5:395.

Ismail HD, Newman RH, Kc DB. 2016. RF-Hydroxysite: a random forest based predictor for hydroxylation sites. *Mol BioSyst*. 12:2427-2435.

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 Physiol*. 174:886-903.

Käll L, Krogh A, Sonnhammer ELL. 2007. Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res.* 35:W429-W432.

Kawashima S, Kanehisa M. 2000. AAindex: Amino Acid index database. *Nucleic Acids Res* 28:374-374.

Kohavi R, John GH. 1997. Wrappers for feature subset selection. *Artif Intell.* 97:273-324.

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. *Front Plant Sci.* 8:66.

Matthews BW. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure.* 405:442-451.

Meyer D, Dimitriadou R, Hornik K, Weingessel A, Leisch F. 2018. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7-0. https://CRAN.R-project.org/package=e1071

Nguema-Ona E, Vicré-Gibouin M, Gotté M, Plancot B, Lerouge P, Bardor M, Driouich A. 2014. Cell wall O-glycoproteins and N-glycoproteins: aspects of biosynthesis and function. *Front Plant Sci.* 5:499.

Peng H, Long F, Ding C. 2005. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27:1226-1238.

Pierleoni A, Martelli PL, Casadio R. 2008. PredGPI: a GPI-anchor predictor. *BMC Bioinformatics.* 9:392.

Qiu WR, Sun BQ, Xiao X, Xu ZC, Chou KC. 2016. iHyd-PseCp: Identify hydroxyproline and hydroxylysine in proteins by incorporating sequence-coupled effects into general PseAAC. *Oncotarget.* 7:44310-44321.

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

Schliep K, Hechenbichler K. 2016. kknn: Weighted k-Nearest Neighbors. R package version 1.3.1. https://CRAN.R-project.org/package=kknn

Schneider G, Wrede P. 1994. The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site. *Biophys J.* 66:335-344.

Schultz CJ, Rumsewicz MP, Johnson KL, Jones BJ, Gaspar YM, Bacic A. 2002. Using genomic resources to guide research directions. The arabinogalactan protein gene family as a test case. *Plant Physiol.* 129:1448-1463.

Schwartz D, Chou MF, Church GM. 2009. Predicting Protein Post-translational Modifications Using Meta-analysis of Proteome Scale Data Sets. *Mol Cell Proteomics* 8:365-379.

Seifert GJ, Roberts K. 2007. The biology of arabinogalactan proteins. *Annu Rev Plant Biol* 58:137-161.

Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H. 2007. Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci USA.* 104:4337-4341.

Shi SP, Chen X, Xu HD, Qiu JD. 2015. PredHydroxy: computational prediction of protein hydroxylation site locations based on the primary structure. *Mol BioSyst*. 11:819-825.

Showalter AM, Basu D. 2016. Extensin and Arabinogalactan-Protein Biosynthesis: Glycosyltransferases, Research Challenges, and Biosensors. Front Plant Sci 7:814-814.

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 Physiol.* 153:485-513.

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. *Arch Biol Sci.* 68:737-746.

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 Tiss Org Cult.* 121:721-739.

Tan L, Showalter AM, Egelund J, Hernandez-Sanchez A, Doblin MS, Bacic A. 2012. Arabinogalactan-proteins and the research challenges for these enigmatic plant cell surface proteoglycans. *Front Plant Sci.* 3:140.

The UniProt Consortium. 2017. UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 45:D158-D169.

van der Loo MPJ. 2014. The stringdist package for approximate string matching. *The R Journal.* 6:111-122.

Varma S, Simon R. 2006. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics.* 7:91-91.

Walsh I, Martin AJM, Di Domenico T, Tosatto SCE. 2012. ESpritz: accurate and fast prediction of protein disorder. *Bioinformatics.* 28:503-509.

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis.* New York (NY): Springer-Verlag.

Wickham H. 2018. httr: Tools for Working with URLs and HTTP. R package version 1.4.0. https://CRAN.R-project.org/package=httr

Wickham H, Hester J, Ooms J. 2018. xml2: Parse XML. R package version 1.2.0. https://CRAN.R-project.org/package=xml2

Wright MN, Ziegler A.2017. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *J Stat Softw*. 77:1-17.

Xiao N, Cao DS, Zhu MF, Xu QS. 2015. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics.* 31:1857-1859.

Xu Y, Wen X, Shao XJ, Deng NY, Chou KC. 2014. iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. *Int J Mol Sci.* 15:7594-7610.

Figure 1. Comparison of hydroxyproline prediction performance in nested cross validation. Area under the receiver operating characteristic curve (AUC) on the 10 hold out instances in nested cross validation (mean ± standard deviation) was used to estimate the performance of four learners: knn – k-nearest neighbors, rf – random forest, svm – support vector machine, xgb – xgboost gradient boosting. Nested cross validation was performed with 10-fold outer loop to estimate performance and two times repeated 3-fold inner loop to optimize hyper parameters by model based optimization for 100 iterations. Different feature selection methods used are shown in plot panels.

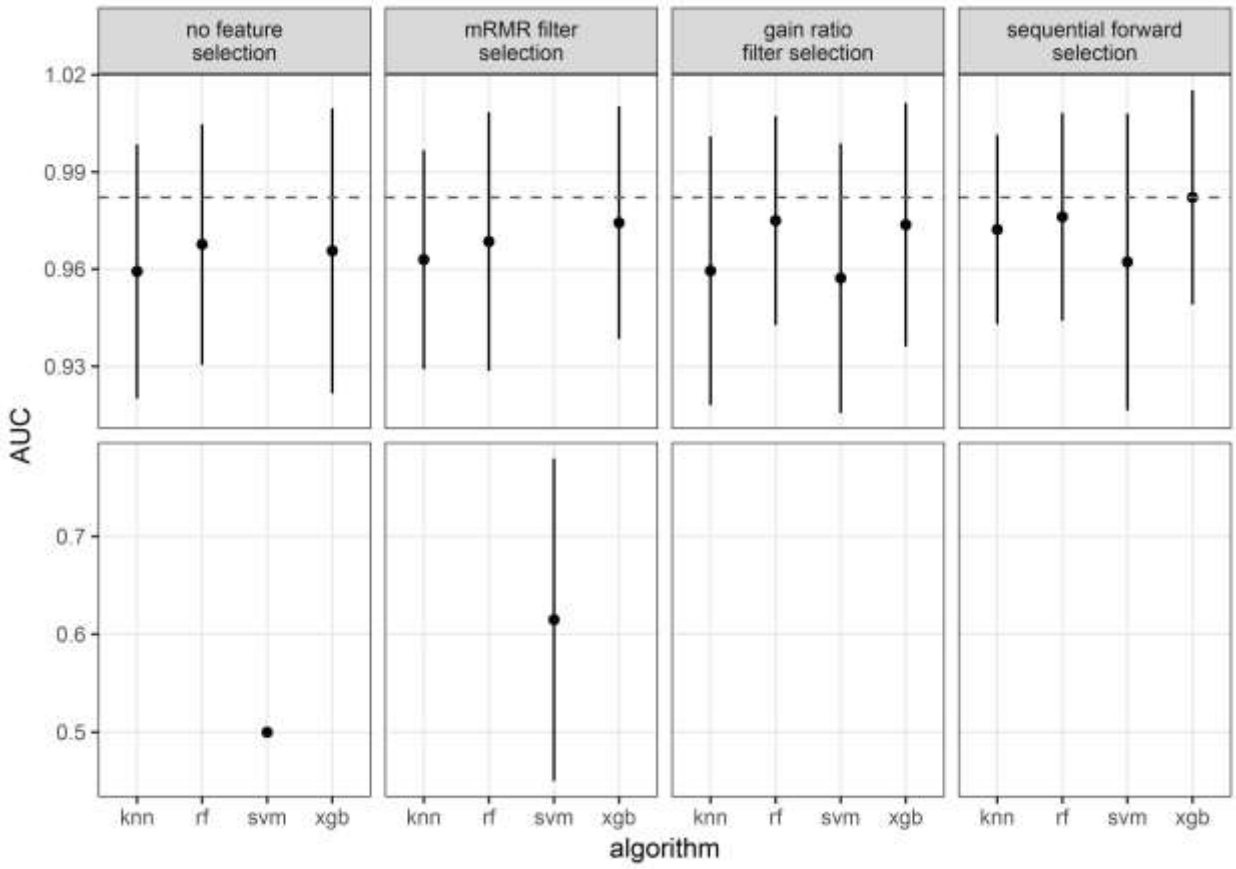

Figure 2. **The effect of decision threshold on 21-mer model performance in nested cross validation**. The model was trained using F1, F3, F4 and F10 feature sets (Table I) constructed using 21-mers. The mean, median, 25% and 75% quantiles for each metric are shown. The metrics were calculated based on the hold out predictions in the nested CV outer loop (three times repeated 10-fold CV). The inner loop, which was used for hyper parameter tuning by model based optimization, consisted of two times repeated 3-fold CV. Horizontal dashed lines correspond to decision thresholds 0.224 which maximizes balanced accuracy and 0.364 which represents the 0.95 specificity cutoff (based on the mean).

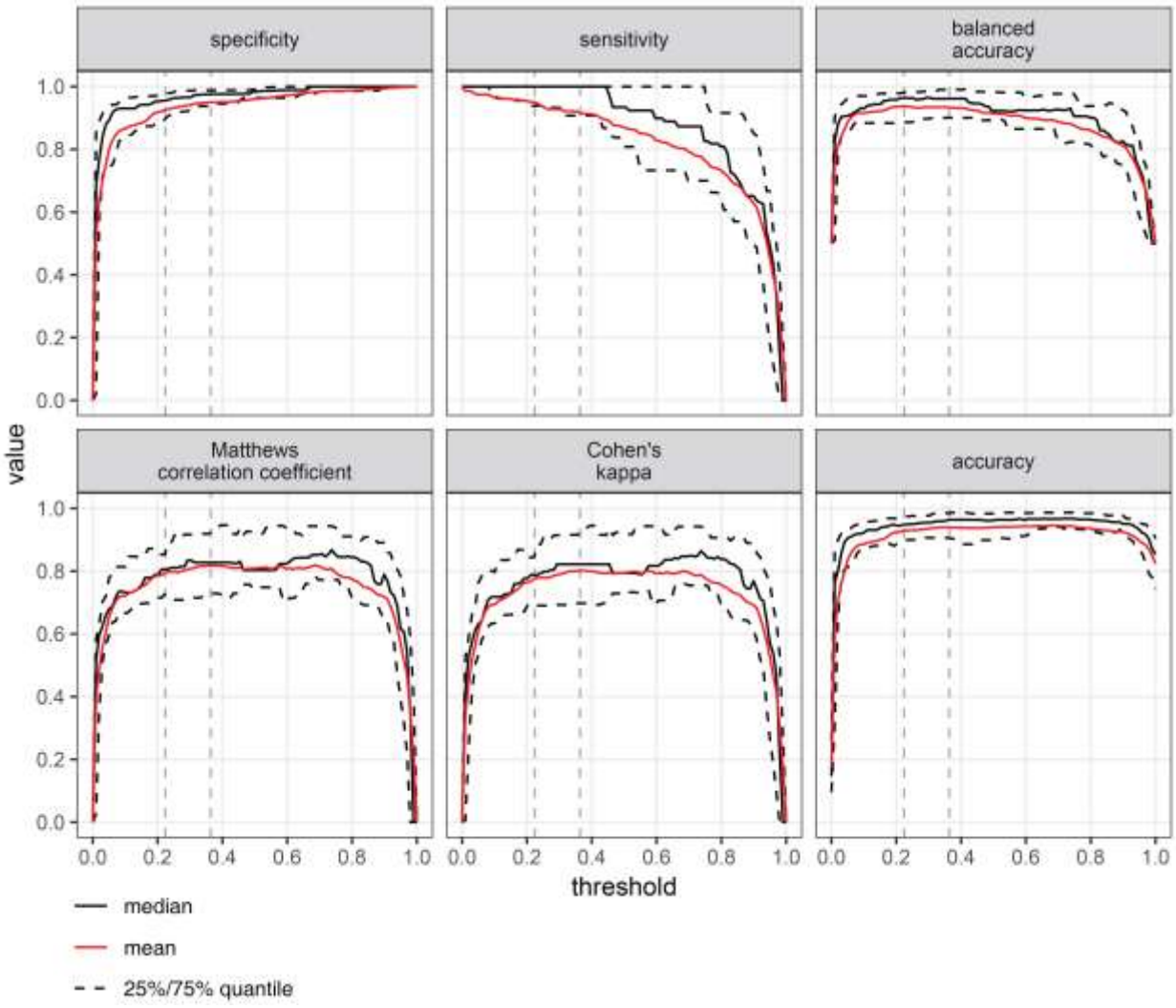

Figure 3. Estimation of performance of Hyp prediction models trained using varied length k-mers. Box plots of AUC for the holdout samples in the outer loop of nested cross validation are shown. Nested cross validation was performed with three times repeated 10-fold outer loop to estimate performance and two times repeated 3-fold inner loop to optimize hyper parameters by model based optimization for 100 iterations. The models were trained using F1, F3, F4 and F10 feature sets (Table I) constructed using 21, 19, 17, 15 and 13-mers.

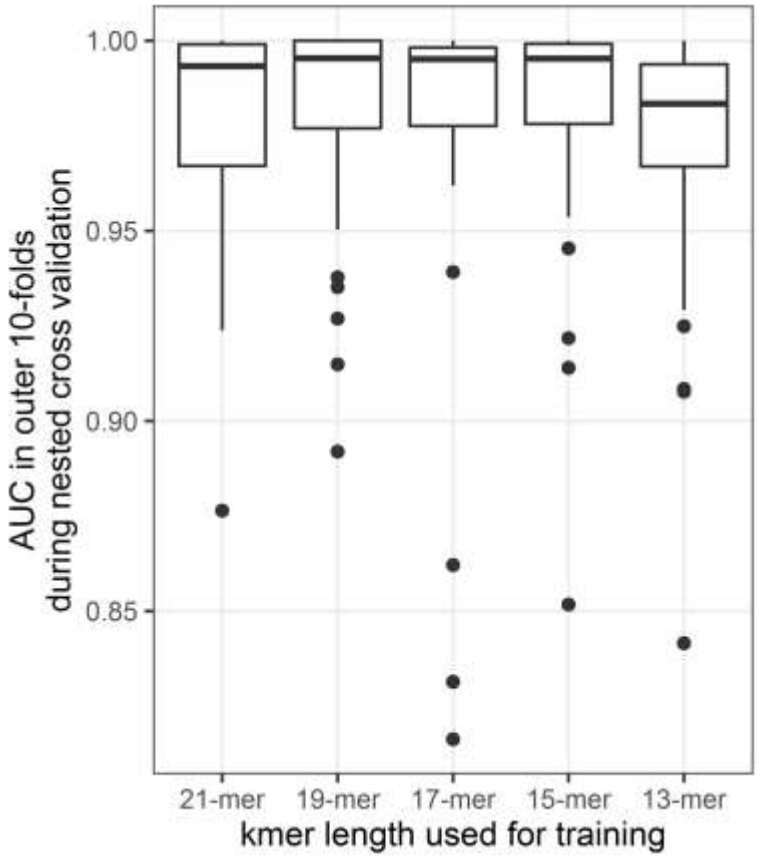

Figure 4: **The effect of decision threshold on 15-mer model performance in nested cross validation**. The model was trained using F1, F3, F4 and F10 feature sets (Table I) built using 15-mers. The mean, median, 25% and 75% quantiles for each metric are shown. The metrics were calculated based on the hold out predictions in the nested CV outer loop (three times repeated 10-fold CV). The inner loop, which was used for hyper parameter tuning by model based optimization, consisted of two times repeated 3-fold CV. Horizontal dashed lines correspond to decision thresholds 0.220 which maximizes balanced accuracy and 0.333 which represents the 0.95 specificity cutoff (based on the mean).

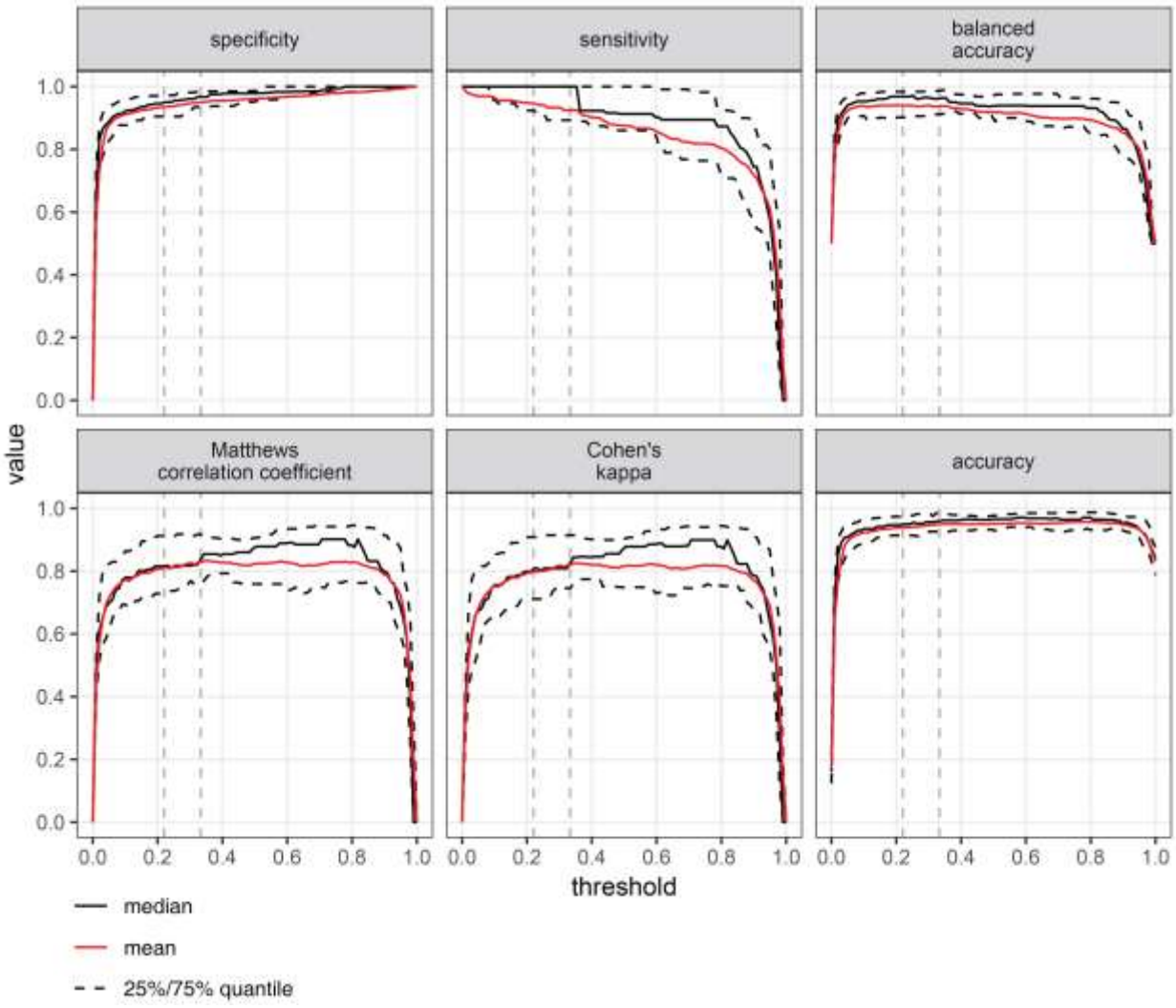

Figure 5. Comparison of several Hyp prediction algorithms with ragp models. Receiver operating characteristic curves obtained by predicting proline hydroxylation on the 21-mer and 15-mer test set sequences using RF hydroxysite, PredHydroxy and ragp models. Area under the curve (AUC) for each model is indicated in the color legend.

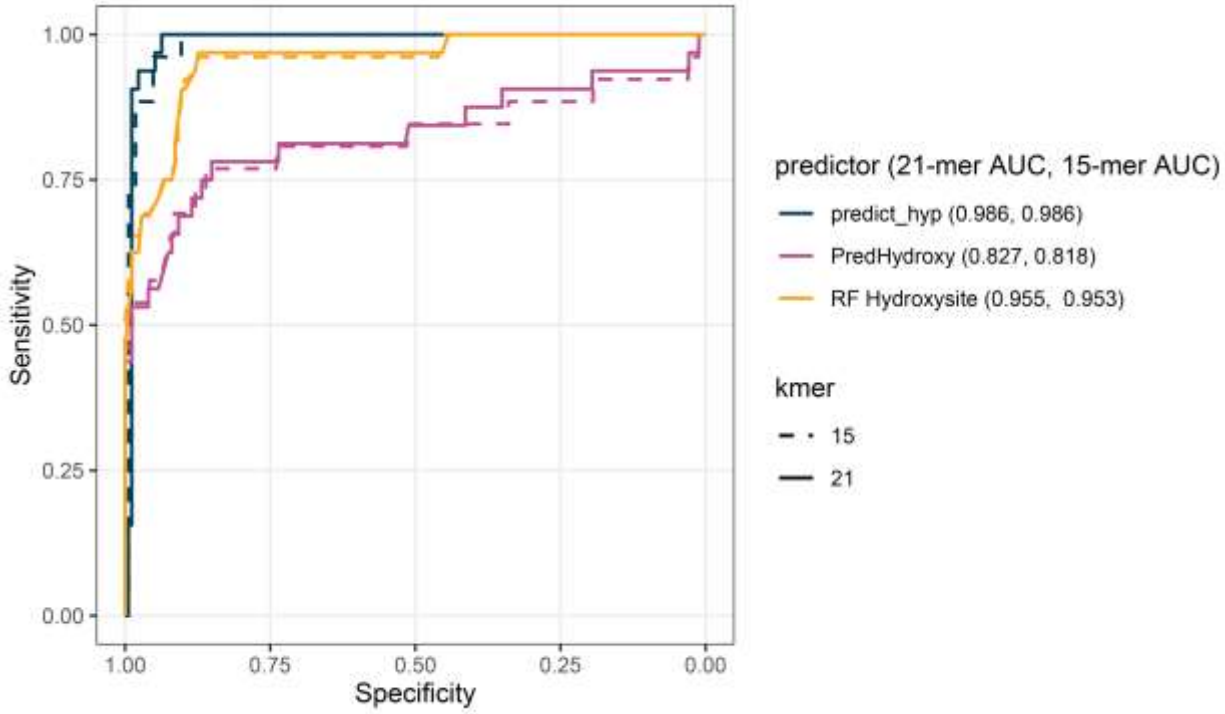

Figure 6. Concordance of N-terminal signal peptide predictions. Euler diagram of N-terminal signal peptide (N-sp) predictions by Phobius (http://phobius.sbc.su.se/), SignalP 4.1 (http://www.cbs.dtu.dk/services/SignalP-4.1/) and TargetP 1.1 (http://www.cbs.dtu.dk/services/TargetP/). A total of 2797062 protein sequences from 62 plant species (Phytozome V12, https://phytozome.jgi.doe.gov/pz/portal.html) was used. 266135 protein sequences were predicted to contain an N-sp by at least two web servers.

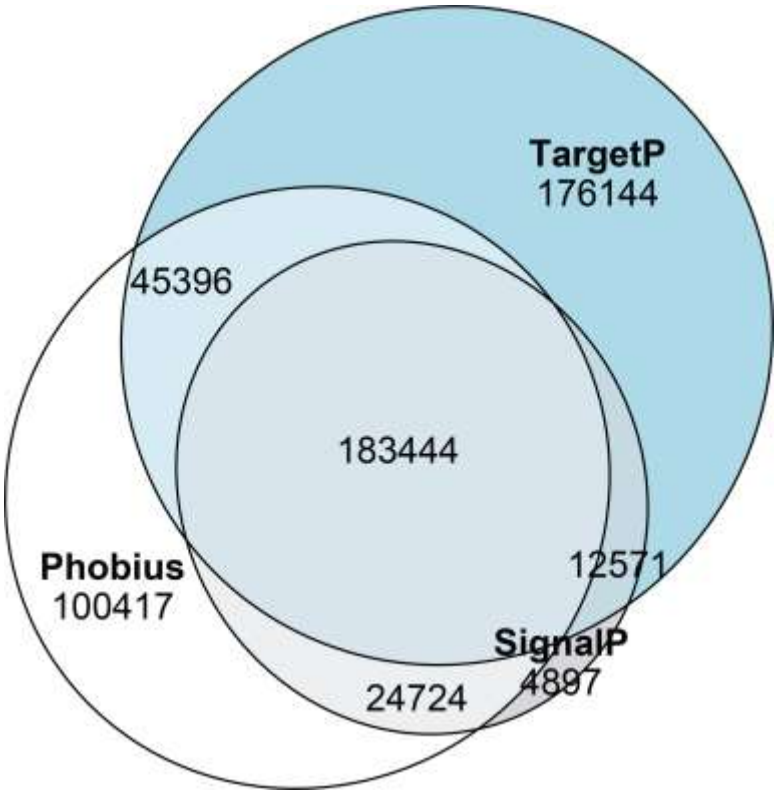

Figure 7. MAAB classified HRGP sequence content in 62 Phytozome proteomes. Number of sequences classified as prototypical HRGPs by motif and amino acid bias classification (MAAB classes 1 - 23) in the 62 analyzed plant proteomes (*O. lucimarinus* did not contain any sequences classified into MAAB classes 1 - 23). Sequences were grouped (fill legend) based on the number of predicted hydroxyprolines.

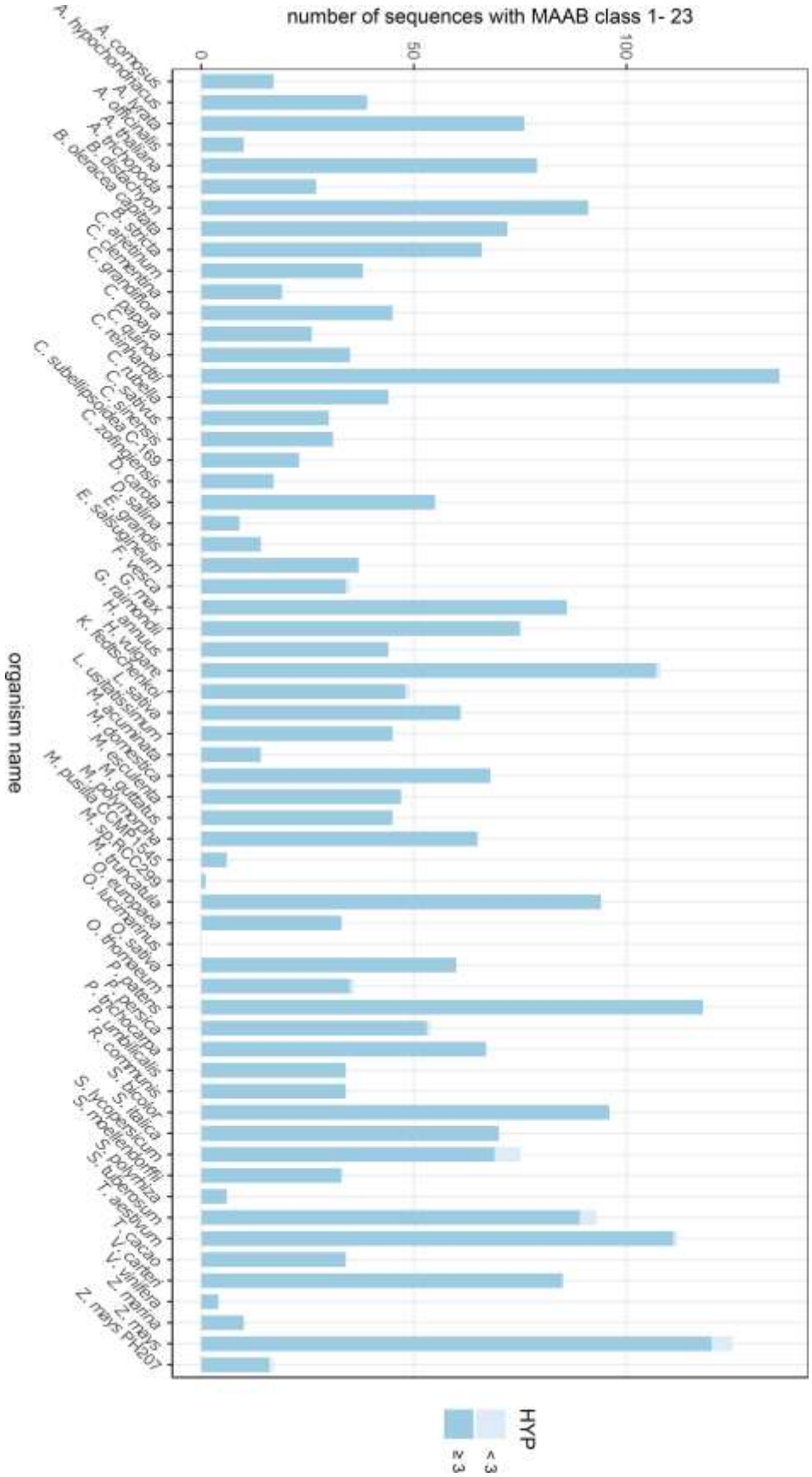

Figure 8. Domain and hydroxyproline distribution in arabinogalactan motif containing sequences. A - top 20 domains by occurrence identified in sequences with AGP motifs. Domain identification was performed using hmmer3 software and Pfam 32 database. Domains with independent E-value < 0.01 were considered and each domain was counted once per sequence. B - Box plots of the number of predicted hydroxyprolines in arabinogalactan motifs in sequences containing at least one of the top 20 domains by occurrence.

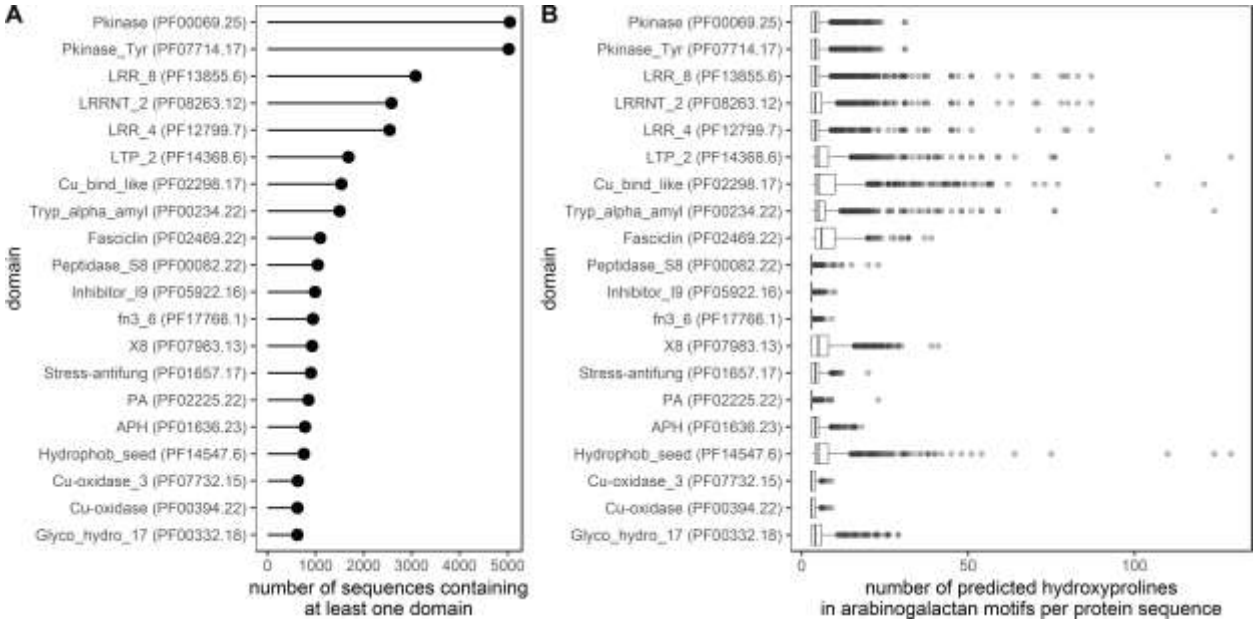

Figure 9. **Schematic structure of several Arabidopsis protein tyrosine kinases detected to contain arabinogalactan motifs**. Protein structure was visualized using the ragp function ***plot_prot*.** The diagram contains the following elements: trans-membrane regions (TM) are shown in yellow, extra-cellular regions are indicated by the dashed line above the sequences, while intracellular regions are indicated by the dashed line bellow the sequence (as predicted by Phobius – ***get_phobius*** ragp function). Signal peptides (as predicted by SignalP – ***get_signalp*** ragp function) are indicated by the thick red line on the N-terminal side. Hydroxyprolines (as predicted by ***predict_hyp*** ragp function) are indicated by vertical dark gray lines. AG glycomodul spans (as predicted by ***scan_ag*** ragp function) are indicated by the light grey background. Domains (as predicted by hmmscan – ***get_hmm*** ragp function) are indicated by rectangles with an appropriate fill as indicated in the legend.

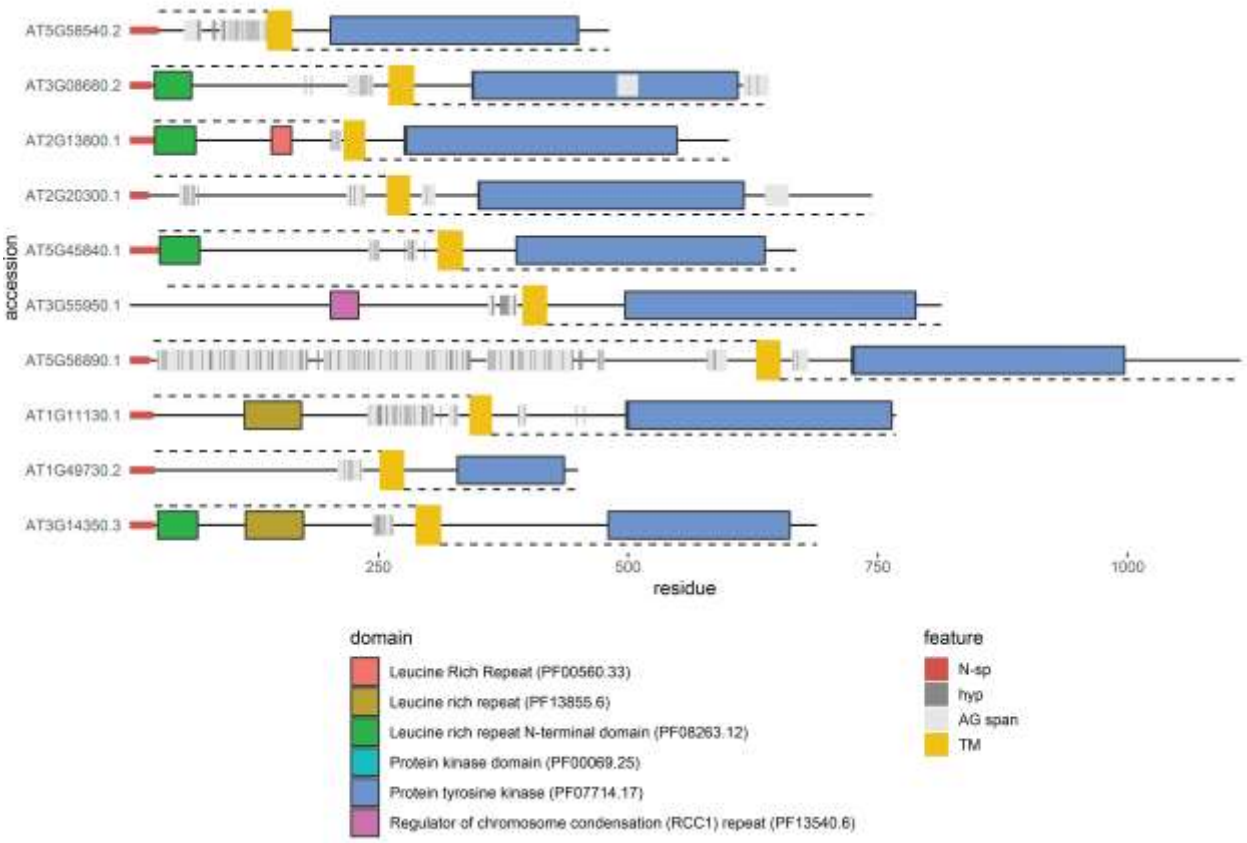

Figure 10. ragp workflow. The ragp workflow consists of the filtering and analysis layers. In the filtering layer Hyp containing sequences are identified. The first step in the filtering layer is prediction of secretory signals (N-sp). Since several prediction algorithms are utilized, the decision which sequences are secreted is reached by a majority vote. The second step of the filtering layer is prediction of proline hydroxylation and filtering sequences containing at least three potential hydroxyprolines. The analysis layer consists of identification of AGPs by searching for localized clusters of characteristic arabinogalactan motifs (AG glycomodules), performing motif and amino acid bias classification, domain annotation using Pfam data base, gene ontology enrichment, glycosylphosphatidylinositol attachment site prediction and disordered region prediction. ragp functions which are available for these tasks are boxed grey in the diagram.

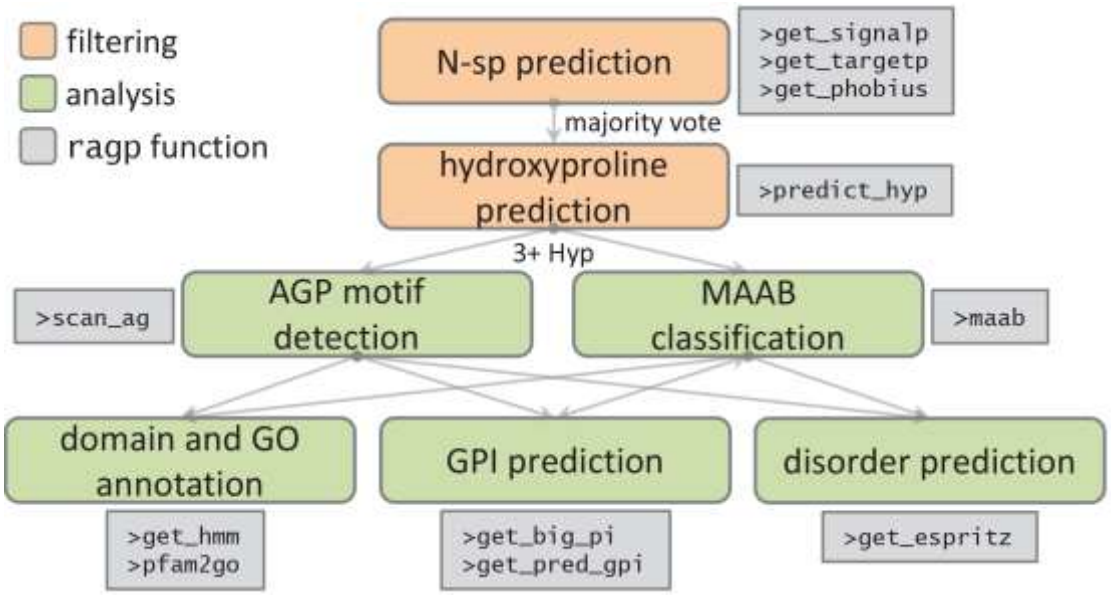

Table I. Feature sets used for prediction of proline hydroxylation

Feature set	description	lag	dimension	amino acid attributes	reference
F1	one to one - attribute to position	-	120	CIDH920105, BHAR880101, CHAM820102, BIGC670101, CHAM810101, DAYM780201	Kawashima and Kanehisa (2000)
F2	one to one - attribute to position	-	100	Factor I - Factor V	Atchley et al. (2005)
F3	Moreau-Broto autocorrelation descriptor	12	72	CIDH920105, BHAR880101, CHAM820102, BIGC670101, CHAM810101, DAYM780201	Kawashima and Kanehisa (2000)
F4	Moreau-Broto autocorrelation descriptor	12	60	Factor I - Factor V	Atchley et al. (2005)
F5	Moran autocorrelation descriptor	12	72	CIDH920105, BHAR880101, CHAM820102, BIGC670101, CHAM810101, DAYM780201	Kawashima and Kanehisa (2000)
F6	Moran autocorrelation descriptor	12	60	Factor I - Factor V	Atchley et al. (2005)
F7	Geary autocorrelation descriptor	12	72	CIDH920105, BHAR880101, CHAM820102, BIGC670101, CHAM810101, DAYM780201	Kawashima and Kanehisa (2000)
F8	Geary autocorrelation descriptor	12	60	Factor I - Factor V	Atchley et al. (2005)
F9	Sequence-order-coupling number	12	24	Schneider-Wrede and Grantham physicochemical distance matrix	Chou (2000); Schneider and Wrede (1994); Grantham (1974)
F10	Quasi-sequence-order descriptor	12	64	Schneider-Wrede and Grantham physicochemical distance matrix	Chou (2000); Schneider and Wrede (1994); Grantham (1974)
F11	Conjoint Triad Descriptor	-	343	-	Shen et al. (2007)
F12	Pseudo-Amino Acid Composition	-	40	-	Chou (2001)
F13	Amphiphilic Pseudo-Amino Acid Composition	-	60	-	Chou (2005), Xiao et al. (2015)
F14	Composition descriptor	-	21	-	Dubchak et al. (1995)
F15	Transition descriptor	-	21	-	Dubchak et al. (1995)
F16	Distribution descriptor	-	105	-	Dubchak et al. (1995)

Table II. Comparison of the performance of several hydroxyproline prediction implementations on the 21-mer and 15-mer test set sequences.

model	test set	Accuracy	Balanced Accuracy	Sensitivity	Specificity	Cohen's kappa	Matthews CC	AUC
RF Hydroxysite		0.850	0.898	0.969	0.828	0.581	0.632	0.955
PredHydroxy		0.908	0.729	0.469	0.989	0.565	0.602	0.827
iHyd PseAAC	21-mer	0.777	0.638	0.438	0.839	0.245	0.249	-
iHyd PseCp		0.859	0.674	0.406	0.943	0.394	0.401	-
ragp 21-mer		0.966	0.954	0.937	0.971	0.875	0.877	0.986
RF Hydroxysite		0.843	0.893	0.962	0.824	0.541	0.598	0.954
PredHydroxy		0.916	0.725	0.462	0.987	0.558	0.591	0.818
iHyd PseAAC	15-mer	0.785	0.649	0.462	0.836	0.246	0.253	-
iHyd PseCp		0.869	0.681	0.423	0.939	0.394	0.397	-
ragp 15-mer		0.953	0.957	0.962	0.952	0.820	0.828	0.986