Thesis proposal:
Understanding underlying features of N-glycosites and glycan structures in plants through machine learning

Scientific background and significance

N-linked glycosylation is a process that is ubiquitous in eukaryotes and many archaea, and also sometimes occurs in bacteria. It is a subsection of the glycosylation class, which also includes glycosylation forms based on different glycosites (e.g. N-linked and O-linked glycosylation, C-mannosylation). These oligosaccharides perform an array of functions, depending on their structure. These functions generally concern structural and folding integrity for proteins and cell walls/matrices, as well as cell-signaling and the transportation of glycoproteins.

The oligosaccharide, also called a glycan, contains a strongly conserved core structure (at least in all metazoan, plants and yeast) consisting of 9 Mannoses (Man), 2 N-acetylglucosamines (GlcNAc) and 3 Glucoses (Glc), but shows differentiation in the maturation process where a stepwise process of additions and removals of molecules like Fucose, Galactose and Xylose alter its form and function. Glycan precursor synthesis starts as a single mannose linked to the Endoplasmic Reticulum(ER)-bound Dolichol phosphate, cytosol side, which is followed by the step-wise addition of Mannoses and N-acetylglucosamines until Man5GlcNAc2 is formed. From there, a flippase switches the Dolichol-bound oligosaccharide precursor from the cytosolic side to the ER lumen side, where additional molecules are added to form GlcNAc2Man9Glc3. Oligosaccharyltransferase (OST) cuts the precursor from the dolichol phosphate and attaches it to the glycosite on the protein, after which the calnexin/calreticulin cycle checks whether the protein is folded correctly or the protein is targeted for degradation.

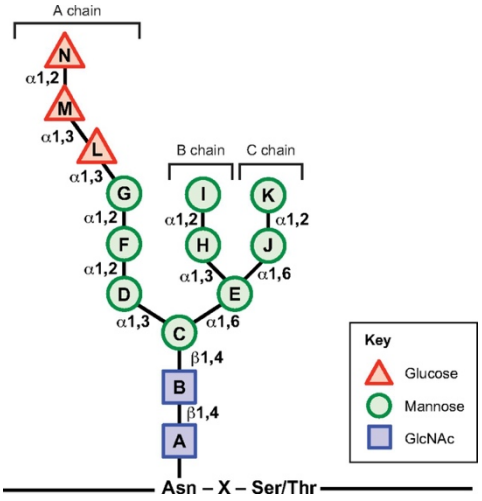

Figure 1: Structure of the core GlcNAc2Man9Glc3 glycan structure and attachment to the Nx[S/T] (Asn-X-Ser/Thr) glycosite. Source: Ferris et al., 2014.

Additional modifications to the glycan might occur; in such case the protein and attached glycan are transferred to the Golgi complex, where a variety of mannosidases and transferases maturate the glycan through addition and removal of residues. The maturated glycans generally fall in one of three categories depending on the additions to the core structure: oligomannose glycan, complex glycan or hybrid glycan. As their names indicate, olgimannose glycans only contain additional mannoses, complex glycans contain a combination of e.g. mannose, fucose and galactose, and hybrid glycans contain a combination of the previous two.

Of the possible N-linked motifs (Nx[S/T/C]), the most abundant one is the Nx[S/T] motif (Essentials of glycobiology, 2009), where Asparagine is followed by anything other than Proline, which in turn is followed by either a Threonine or Serine. This particular sequence of amino acids functions as an attachment site for polysaccharides, including N-linked glycosylation, and is called a glycosite or sequon. Previous research has shown that these sequons form a necessary site for N-glycosylation to happen, but on their own are not adequate enough to accurately predict glycosylation (Gavel *et al.*, 1990). Where Gavel *et al.* did not find any direct features that could predict the presence or absence of N-glycosylation, a more recent study has found that functional

features of the protein could be used to differentiate between proteins that will be glycosylated and those that won't (Li *et al.*, 2014).

With the rise and ever-developing state of computational power, our ability to perform lengthy calculations on massive data matrices grows, which can manifest itself in machine-learning derived models and e.g. allowing scientists to greatly reduce their search space to just a set of highly probably targets. These models can be used to find potential underlying rules that might not be clearly visible in large datasets, and improve our understanding of the process (in this case N-glycosylation in plants). So far, various computational models have been developed for the prediction of glycosites, relying on algorithms like Random Forests, Support Vector Machines and Neural Networks (Table 1). They've made use of a variety of methods for positive and negative collection and labelling, feature generation and feature selection. However, the species scope of these previous models is largely centered around the human proteome, with some having a mixture of species; little is done for plants. Due to the importance of glycoproteins, it will be fruitful to attempt to enhance our understanding of the N-linked glycosylation rules in plants. By learning underlying features and rules of N-glycosylation, these might open up routes for biotechnology to explore.

Table 1: Concise overview of several computational prediction models for glycosites, sorted by date. Sources: 1) Gupta et al. 2004, 2) Caragea et al. 2007, 3) Hamby et al. 2008, 4) Chuang et al. 2012, 5) Li et al. 2015, 6) Li et al. 2019.

Name	Date	Glycosite type	Species	Model
NetNGlyc[1]	2002	N-linked	Human	Neural Network
EnsembleGly[2]	2007	C, N, O-linked	Multiple species	Ensemble SVM
GPP[3]	2008	O-linked	Mammals	Random Forest
NGlycPred[4]	2012	N-linked	Eukaryotes	Random Forest
GlycoMine[5]	2014	C, N, O-linked	Human	Random Forest
GlycoMine_PU[6]	2019	C, N, O-linked	Human	PU-learning

Objectives

The thesis aims to answer a set of three, closely related questions regarding the specificity and underlying rules of glycosites and glycan modification in plants. Empirical and experimentally verified datasets, generated through mass-spectrometry, will form the backbone for machine learning applications to discover and understand the underlying rules or features that can answer the following three objectives:

i) Find features that allow for prediction of whether a protein is N-glycosylated or not;
ii) Find features that allow for prediction of whether an individual glycosite is N-glycosylated or not;
iii) Find features that allow for the prediction of mature glycan modification.

By learning these features and making models that adhere to these, we seek to make accurate predictions about whether proteins will be N-glycosylated or not, whether potential glycosites will be N-glycosylated or not and what glycan modifications will occur, based on sequence characteristics. So far, various papers employed statistical and machine learning oriented approaches for glycosite prediction, often using the human proteome or a mixture of organisms.

By using an aggregate dataset originating from *Arabidopsis thaliana*, more specialized models can be generated that will aid in plant- or *A. thaliana*-specific applications that involve N-glycosylation.

Approach

Note: A schematic overview can be seen in Figure 3.

Data acquisition and curation

For this research, a group of five experimentally verified datasets will be used. These datasets originate from five studies utilizing mass-spectrometry analysis of *A. thaliana* samples (Zielinska *et al.* 2012, Song *et al.* 2013, Xu *et al.* 2016, Ma *et al.* 2016, Zeng *et al* 2018). The five sets can be aggregated on the aspects of protein accession number and sequon. From these accession numbers the protein sequence can be retrieved for protein and glycosite predictions. The datasets also have their individual features, like mass-spectrometry related measurements or glycan structure. These features could potentially be used on smaller-scale models for e.g. glycan prediction, but the precise methodology will be addressed further down the thesis.

Careful curation of the data will be applied to keep data integrity. As some of the studies from which the datasets originate are several years old, finding outdated accession numbers is unavoidable. The older identifiers will be updated to today's standards where applicable (Wein *et al.*, 2012), and discarded if no longer supported. Likewise, overlapping entries will be discarded to avoid artificial overrepresentation of entries, as the studies reference each other in their datasets. Here the combination of identifier and sequon will be considered as a unique entry, but the complete protein sequence per identifier will be scanned for potential multiple occurrences of the sequon, although the chance is extremely low. Additionally, data validation will be applied, where scripts independent of previous data curation will be used to ascertain the aggregated data is accurate in relation to the original datasets.

Table 2 displays the number of sequons per paper, which serve as a source for the dataset. At the time of writing, the dataset contains 3410 curated sequons that link to a current TAIR identifier. Roughly 1085 were lost after the redundancy filter and 213 could not be used because either the linked identifier did not exist anymore, or the old identifier linked towards multiple possible newer identifiers. The presented numbers are likely not final as we finish up the curation of the data, but should give a reasonable indication of our data scope.

Table 2: Overview of the five Arabidopsis thaliana datasets, with the amount of glycosites(sequons) per data set. Raw indicates the summation of all the datasets. Curated total indicates the total when accounting for the loss of outdated and/or ambiguous entries. The non-redundant number represents the curated total minus the duplicate entries.

Source	**Sequons**
Zielinska 2012	2185
Song 2013	390
Xu 2016	862
Ma 2016	161
Zeng 2018	1110
Raw total	4708
Curated total	4495
Of which non-redundant	**3410**

A point of contention could be the identification and inclusion of a negative glycosylation set. One method could focus on glycosite level, where the Nx[S/T] motif is recognized but, to our knowledge, does not result in glycosylation. One could speculate that a simple motif finder would suffice to find all potential sites and intersect this collection to our positives (the experimentally verified glycosites). The remainder would then be the negative set. However, it will prove difficult to ascertain that these are in fact negative, as the assignment of 'negative' relies on the presumption that our positives provide good coverage. Secondly, when looking at protein level (is the protein glycosylated or not?), one could retrieve the entire proteome of *A. thaliana* and deem the proteins that do not occur in the experimentally verified set as negatives. Another potential route is the usage of Positive-labelled learning (PU-learning), where the negative set is inferred from the positive and unlabeled dataset through various statistical methods. A recent study (Li *et al.* 2019) made successful use of this method and could maybe also prove fruitful for us. The precise handling of negatives will have to be carefully determined after a more thorough literature study on the characteristics of each method.

Feature selection and classifier models

After data preprocessing, the focus will lie on preliminary analysis of the feature list. Given the fact that there will be a wide array of potentially influencing features or characteristics and only several thousand sequons, it is not unreasonable to assume problems from 'large p small n' dimensionality might occur. Additionally, it is not a given that the five aggregated datasets are truly comparable. Therefore, we seek to perform various unsupervised methods like K-means clustering, Hierarchical clustering and Principal Component Analysis (PCA). K-means and Hierarchical clustering could show whether the datasets are generally overlapping or are largely distinct from one-another, while a PCA could show where the majority of the variation in the data comes from (e.g. inter- or intragroup). The application of these unsupervised methods requires features. Some potential features could be sequence alignments (either sequon or complete protein), or sequence k-mer mapping.

There are various machine learning algorithms that could provide a model for glycosylation and glycan prediction. Previously, methods like Support Vector Machines (SVM), PU-learning and Neural Networks (NN's) have been used (Table 1). Regularization will be applied through algorithms like LASSO/Ridge regression/Elastic Net and generalization will be estimated with cross-validation. While currently undecided yet about which algorithm to utilize, it will have to cope well with a large array of features and have a runtime that is reasonable for n-fold cross-validation (n to be determined). The performance of the generated models will be supplied through the archetypical confusion matrices/receiver-operator characteristics.

Workplan

A schematic overview of the steps described in the Approach section can be seen in Figure 3. Outside the activities listed in the schematic overview, a midterm progress presentation, a final presentation and a thesis defense will be held. The exact dates are to be determined.

Time schedule	
ECTS	36
Hours	1008
Weeks	25.2
Starting date	04-03-2019
Ending date	05-09-2019

As roughly the first 4 weeks were used for the writing of this proposal, roughly 21 weeks remain to distribute. Below are the main steps from the schematic overview (Figure 3) and their crudely estimated duration (overview in Figure 2):

1. Data curation & validation 2 weeks
2. Negative set integration 2 weeks
3. Feature generation 4 weeks
4. Unsupervised learning 2 weeks
5. Creating ML models 7 weeks
6. Feature selection 1 week
7. Writing report 3 weeks

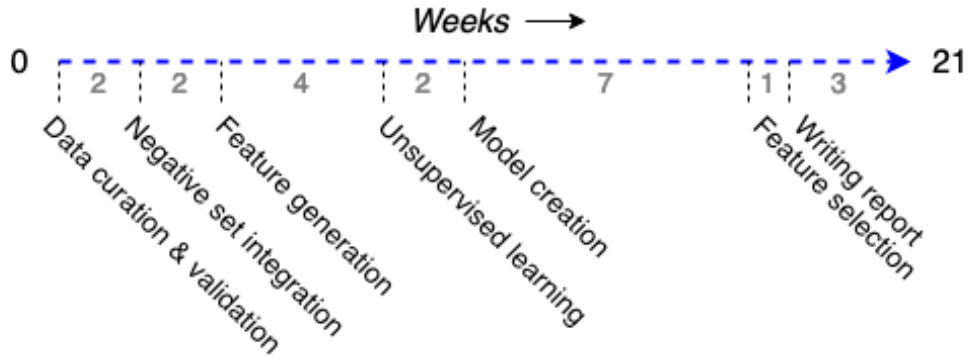

Figure 2: Schematic overview of the planning. Displayed are the main steps from the workflow, with their duration indicated in grey numbers. The timeline starts from the finishing of the proposal. Image generated with draw.io.

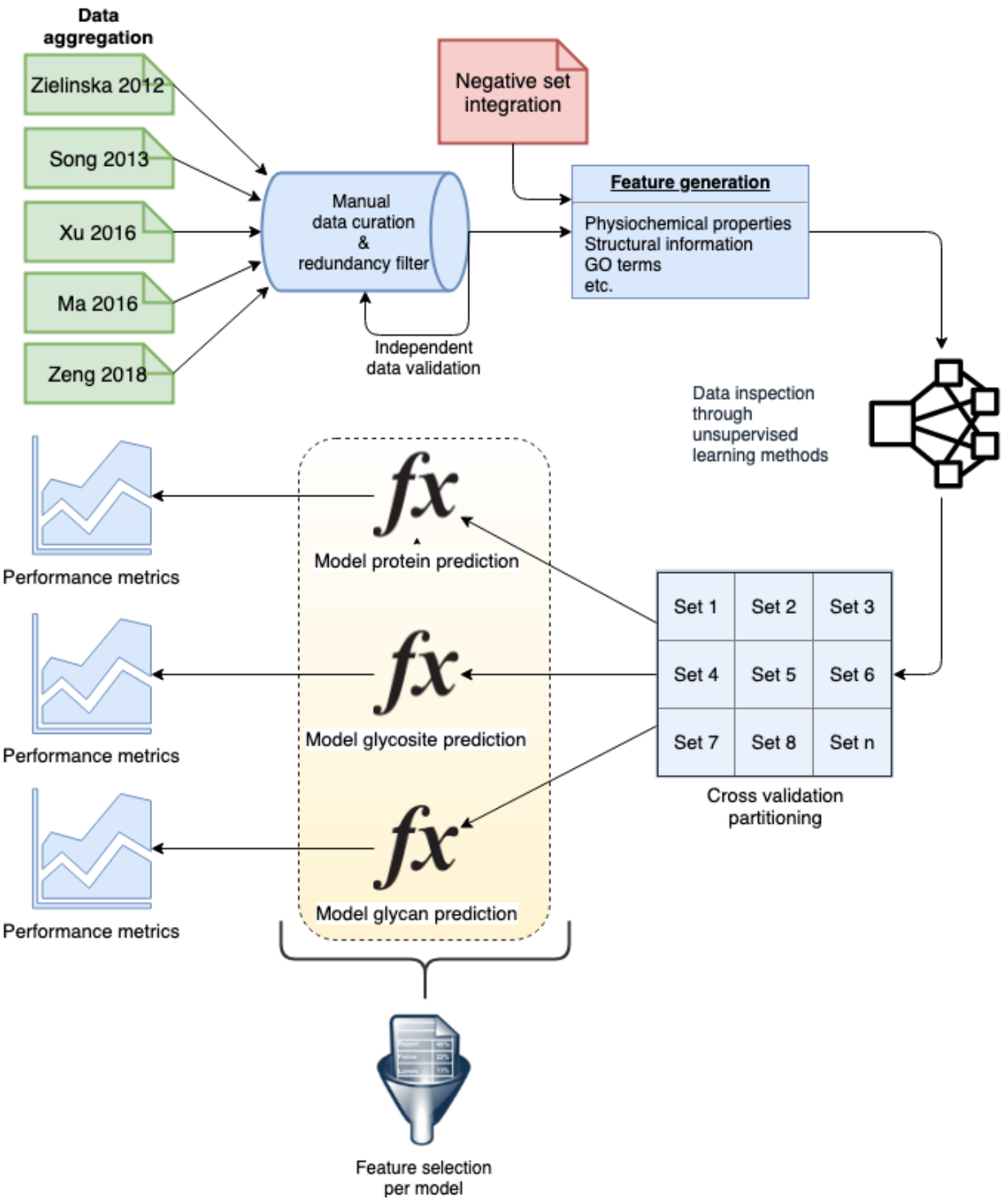

Figure 3: Schematic overview of the thesis as described in the approach. Five Arabidopsis datasets will be aggregated and curated/validated. A negative set will be added depending on negative-approach. Various features will be added depending on the sequence window size. Preliminary investigation will be done with unsupervised methods and feature selection will be applied to find potential causes of differences. The dataset will be partitioned for cross-validation and three individual models will be generated. Image generated with draw.io.

References

Essentials of glycobiology:
https://www.ncbi.nlm.nih.gov/books/NBK1908/
Varki, A., Cummings, R. D., Esko, J. D., Freeze, H. H., Stanley, P., Bertozzi, C. R., ... & Etzler, M. E. (2009). Nematoda--Essentials of Glycobiology. Cold Spring Harbor Laboratory Press.

Difference in motifs:
https://academic.oup.com/peds/article-abstract/3/5/433/1453569
Gavel, Y., & Heijne, G. V. (1990). Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering. Protein Engineering, Design and Selection, 3(5), 433-442.

Accession nr conversion:
https://academic.oup.com/nar/article/40/W1/W276/1070987
Wein, S. P., Coˆte, R. G., Dumousseau, M., Reisinger, F., Hermjakob, H., & Vizcaíno, J. A. (2012). Improvements in the protein identifier cross-reference service. Nucleic acids research, 40(W1), W276-W280.

Ferris 2014, glycan structure:
http://dmm.biologists.org/content/7/3/331
Ferris, S. P., Kodali, V. K., & Kaufman, R. J. (2014). Glycoprotein folding and quality-control mechanisms in protein-folding diseases. Disease models & mechanisms, 7(3), 331-341.

Glycosylation in plants:
https://academic.oup.com/glycob/article/26/9/926/2388896
Strasser, R. (2016). Plant protein glycosylation. Glycobiology, 26(9), 926-939.

Models:
Glycomine:
https://academic.oup.com/bioinformatics/article/31/9/1411/200596
Li, F., Li, C., Wang, M., Webb, G. I., Zhang, Y., Whisstock, J. C., & Song, J. (2015). GlycoMine: a machine learning-based approach for predicting N-, C-and O-linked glycosylation in the human proteome. Bioinformatics, 31(9), 1411-1419.

EnsembleGly:
https://www.ncbi.nlm.nih.gov/pubmed/17996106
Caragea, C., Sinapov, J., Silvescu, A., Dobbs, D., & Honavar, V. (2007). Glycosylation site prediction using ensembles of Support Vector Machine classifiers. BMC bioinformatics, 8(1), 438.

GPP:
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-500
Hamby, S. E., & Hirst, J. D. (2008). Prediction of glycosylation sites using random forests. BMC bioinformatics, 9(1), 500.

NetNGlyc (no 2004 publication):
https://www.researchgate.net/publication/290163207_Prediction_of_N-glycosylation_sites_in_human_proteins
2002 source:
https://psb.stanford.edu/psb-online/proceedings/psb02/gupta.pdf

Gupta, R & Jung, E & Brunak, Søren. (2004). Prediction of N-glycosylation sites in human proteins. 46. 203-206.

NGlycPred:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3426846/
Chuang, G. Y., Boyington, J. C., Joyce, M. G., Zhu, J., Nabel, G. J., Kwong, P. D., & Georgiev, I. (2012). Computational prediction of N-linked glycosylation incorporating structural properties and patterns. Bioinformatics, 28(17), 2249-2255.

Positive-unlabelled learning:
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-019-2700-1
Li, F., Zhang, Y., Purcell, A. W., Webb, G. I., Chou, K. C., Lithgow, T., ... & Song, J. (2019). Positive-unlabelled learning of glycosylation sites in the human proteome. BMC bioinformatics, 20(1), 112.

Data sets:

Zielinska 2012:
https://www.ncbi.nlm.nih.gov/pubmed/22633491
Zielinska, D. F., Gnad, F., Schropp, K., Wiśniewski, J. R., & Mann, M. (2012). Mapping N-glycosylation sites across seven evolutionarily distant species reveals a divergent substrate proteome despite a common core machinery. Molecular cell, 46(4), 542-548.

Song 2013:
https://www.sciencedirect.com/science/article/pii/S1874391913004375
Song, W., Mentink, R. A., Henquet, M. G., Cordewener, J. H., van Dijk, A. D., Bosch, D., ... & van der Krol, A. R. (2013). N-glycan occupancy of Arabidopsis N-glycoproteins. Journal of proteomics, 93, 343-355.

Xu 2016:
https://www.mcponline.org/content/15/6/2048.short
Xu, S. L., Medzihradszky, K. F., Wang, Z. Y., Burlingame, A. L., & Chalkley, R. J. (2016). N-glycopeptide profiling in arabidopsis inflorescence. Molecular & Cellular Proteomics, 15(6), 2048-2054.

Ma 2016:
https://nph.onlinelibrary.wiley.com/doi/full/10.1111/nph.14014
Ma, J., Wang, D., She, J., Li, J., Zhu, J. K., & She, Y. M. (2016). Endoplasmic reticulum-associated N-glycan degradation of cold-upregulated glycoproteins in response to chilling stress in Arabidopsis. *New Phytologist*, *212*(1), 282-296.

Zeng 2018:
https://www.mcponline.org/content/17/3/413.short
Zeng, W., Ford, K. L., Bacic, A., & Heazlewood, J. L. (2018). N-linked Glycan Micro-heterogeneity in Glycoproteins of Arabidopsis. Molecular & Cellular Proteomics, 17(3), 413-421.