Employing predictive techniques to functionally and structurally characterize orphans in *Arabidopsis thaliana*

Priyanka Bhandary*
November 12, 2020

A Ph.D. Research Proposal submitted to the Program of Study Committee (POSC)

Major Professor: Dr. Eve Syrkin Wurtele
Co-major Professor: Dr. Matthew Hufford
Committee: Dr. Karin Dorman, Dr. Basil Nikolau, Dr. Marna Yandeau-Nelson

Bioinformatics and Computational Biology, Iowa State University,
Ames, IA, USA

Abstract. An essential part of understanding a genome is deciphering the function of the genes that it is composed of. Studying gene function can give insight into an organism's developmental and physiological mechanisms. One of the computational tools used to infer the function of an unannotated gene is sequence homology. On the other hand, it has been hypothesized that sets of genes with similar expression patterns across multiple spatial and temporal conditions could potentially be related in function. These group of genes characterized as regulons has been made available in *Arabidopsis thaliana*. Gaining regulon information for unannotated genes can give context to a scientist who can further validate these genes' potential functions. I have developed a machine learning framework to predict the regulon information for functionally unannotated genes using the massive amount of publicly available expression data for *Arabidopsis thaliana.* Another level of understanding the genome is at the protein level, where the protein structure plays a paramount role. There have been efforts to use machine learning methods for predicting protein structure, but most rely on homology-based features. Here, I propose to use an ensemble-based machine learning approach that uses features of the amino acids that the protein is composed of, one of them being the physicochemical properties. Finally, I propose to perform meta-analysis with expression data for unannotated candidate orphan genes using MetaOmGraph. Co-expression and clustering analysis could shed light into the modules that they express in. I propose to use Gene Ontology (GO) enrichment to further augment the functional context of these candidate orphan genes. Additionally, I will apply my optimized machine learning method to predict regulon information for these genes.

Glossary

Machine learning - The subfield of artificial intelligence which focuses on methods to construct computer programs that can learn from patterns in the data with respect to some class and assessed using a performance measure.

Features - Individual, independent variables, the patterns of which a machine learning model can learn from and perform predictions. In the case of Aim 1, the features used in the machine learning framework are the RNA-seq data for *Arabidopsis thaliana* obtained from multiple different conditions.

Classes/labels - Label is the thing that we aim to predict. In supervised learning, the target labels are known for the training dataset but not for the test. In the case of Aim 1, the label used is the regulon information for genes in *Arabidopsis thaliana*.

Classifier - An algorithm that can map input data to a specific label/class.

Regulons - Sets of genes that show similar expression patterns across multiple spatial, temporal, environmental, genetic conditions and likely to be under common transcriptional regulations.

* Preliminary examination for candidacy into the Bioinformatics and Computational Biology (BCB) graduate program at Iowa State University, Genetics Developmental and Cell Biology (GDCB)

Jaccard index - Also known as Jaccard similarity coefficient, is a statistic used in understanding the similarities between sample sets. The measurement emphasizes similarity between finite sample sets, and is formally defined as the size of the intersection divided by the size of the union of the sample sets. In our case, for a particular gene1 and gene 2, Jaccard Index would be the proportion of reads that map to *both* gene 1 and gene 2 which also map to either gene 1 *or* gene 2.

Correlations due to technical factors - Correlations in the RNA-seq data arising from technical biases such as sequencing bias, sampling bias, etc. which can confound the true biological correlation existing within the data. In Aim 1, we propose to address the correlations manifested through multi-mapping done by the aligner.

Functional annotation - Information about a gene's biological identity – it's various aliases, molecular function, biological role(s), subcellular location and protein domains.

Q3 prediction - Residues assigned to one of the three classes of secondary structure of a protein (helix, beta sheet, coil), by percent of total.

Q8 prediction - Residues assigned to one of the eight classes of secondary structure of a protein (3_{10} helix (G), α-helix (H), π-helix (I), β-stand (E), bend(S), bridge (B), turn(T) and others(C)), by percent of total.

Physicochemical properties of amino acids - Physical and chemical properties of naturally occurring amino acids.

Orphan genes - Genes in an organism that bear no similarity with genes or proteins in other evolutionary lineages.

Specific Aims:

This proposal aims to develop predictive methods for the characterization of unannotated genes in *Arabidopsis thaliana* using machine learning methods and data available in the public domain. Characterization at both the gene and protein level as well as at the structural and functional level can give a deeper understanding of an organism.

Aim 1. Construct a machine learning framework that uses RNA-seq expression-based features to classify functionally unannotated protein-coding genes in *Arabidopsis thaliana* into regulons.

Aim 2. Develop an ensembl-based machine learning method to predict secondary structure of orphan proteins using non-homology-based features.

Aim 3. Conduct meta-analysis with expression data using MetaOmGraph and predict regulon information employing the machine learning framework from Aim 1 for characterization of candidate orphan genes identified in *Arabidopsis thaliana.*

Aim 1. Construct a machine learning framework that uses RNA-seq expression-based features to classify functionally unannotated protein-coding genes in *Arabidopsis thaliana* into regulons. Regulons are defined as sets of genes that share similar expression profiles across multiple conditions and are likely to be under common transcriptional regulations. For the characterization of functionally unannotated genes, many different tools are available but there's no universal method developed to decipher regulon information for these genes. I have developed a machine learning method in collaboration with Sagnik Banerjee. This method uses expression data as features and the corresponding available regulon information as labels. Picking up signal from only the regulon-based expression patterns could help classify genes into regulons. Using predictive approaches to characterize gene function can greatly assist researchers to focus their validation on a handful of conditions. The predictive framework can operate with gene count data alone, computed from publicly available RNA-Seq expression data. We also compared this method with other techniques to infer regulon information, such as correlation, Markov Chain clustering (MCL) and modularity maximization.

The outcome will be a machine learning method that uses expression-based features to predict regulon information for functionally unannotated genes.

Aim 2. Develop an ensembl-based machine learning method to predict secondary structure of orphan proteins using non-homology-based features. For this aim, I propose to develop an ensembl-based machine

learning method in collaboration with Sagnik Banerjee which could enhance the prediction of orphan protein secondary structure. There has been extensive research that has been done to predict secondary structure of proteins, with the focus now on using machine learning and deep learning. However, most of these methods use homology-based features which could potentially not work for proteins for which there is no homology information. We propose to use the high-quality, literature-derived physicochemical properties of amino acids as features in an ensembl machine learning method for orphan secondary structure prediction. This could be a high risk, high reward project.
The outcome will be an ensembl-based machine learning tool that could enhance prediction of orphan protein secondary structure.

Aim 3. Conduct meta-analysis with expression data using MetaOmGraph and predict regulon information employing the machine learning framework from Aim 1 for characterization of candidate orphan genes identified in *Arabidopsis thaliana.* Studying orphan genes are a major point of interest to scientists because they could hold a key to understanding how new genes evolve in a species. It also becomes imperative to decipher how these genes could contribute to the adaptation of the organism to the dynamic environment it encounters. Looking at the expression spectrum of these genes could help uncover conditions under which these genes are expressed. For this aim, I will be carrying forward the work done by Jing Li in the lab. She identified unannotated candidate orphan genes in *Arabidopsis thaliana* using her pipeline. The expression patterns of these candidate orphan genes will be studied in MetaOmGraph (MOG). Clustering and Gene Ontology (GO) Enrichment Analysis could potentially lead to a better understanding of the modules that these genes express under. Further, I propose to apply the machine learning method developed from Aim 1 to predict regulon information for these candidate orphan genes. Ultimately, these unannotated candidate orphan genes with transcription evidence will be submitted to The Arabidopsis Information Resource (TAIR). If there is Ribo-seq data available for these genes, I would have evidence of translation.
The outcome will be a comprehensive list of orphan genes and candidate orphan genes in *Arabidopsis thaliana* with evidence of transcription, translation, regulon assignment, and the conditions under which they are expressed.

2. Research Strategy

2.1 Background and strategy

Under Aim 1, I have developed a machine learning classification framework in collaboration with Sagnik Banerjee to functionally characterize unannotated genes in *Arabidopsis thaliana* using publicly available expression data. Unraveling gene function is pivotal to better understanding of an organism. A key challenge to biologists is deciphering functions of genes, especially since the *A. thaliana* genome has been fully sequenced [1]. Understanding gene function could give an insight into the genetic makeup of the organism, and is central to the discernment of its developmental and physiological processes.

Once a genome is sequenced, both structural and functional annotations play an important role in genome annotation. One of the tools that is commonly used to infer function of an unannotated gene is sequence homology. When a genome is newly sequenced, the sequences are searched against databases of other closely related organisms to identify similar sequences, and functions inferred. If two sequences have high homology, then it is assumed that the functions of the two genes could be related. In yeast, 30% of the genes had been identified before and once the genome was sequenced, 30% of the genes were given a putative function based on homology, while the rest of the genes couldn't be characterized. Homology tools could infer molecular function; however, this method fails in the case of genes that bear no homology to other genes. Functional characterization may also be deduced from the expression profiles of genes. Two genes that have the same expression profile under the same conditions could be hypothesized to have the same or related function [2], [3]. This idea forms the basis of the machine learning framework where the relationships between genes is inferred from their expression profiles across multiple conditions.

The NCBI-sequence data archive (SRA) comprises of a burgeoning number of biological samples from numerous experiments conducted under several developmental, genetic and environmental conditions [4]. This repository is ever-growing and is yet to be optimally utilized. These data, when aggregated, could hold the potential to identify genes that maybe co-regulated at the transcriptional level across different conditions [5]–[7]. One needs to utilize the data in such a way so as to tease out information that could be contained within, which could give an insight into how a gene functions. Thus, a promising avenue to infer biological process(es) in which unannotated genes participate is to draw information from co-expressed genes of known function [8]–[11]. This is where regulons come into play. Regulons comprise of genes that exhibit a similar expression profile across a plethora of conditions [3]. A group of genes co-expressed across varied conditions can be termed as a eukaryotic regulon [3], [12]–[14]. Mentzen and Wurtele in their paper published in 2008, used microarray co-expression analysis to assign regulon information for genes in *Arabidopsis thaliana*. In their paper, regulon information was deciphered for genes using correlation information from microarray expression data. Microarray data from Arabidopsis obtained from NASCArrays and PlexDB was used to construct regulons utilizing a Pearson's correlation threshold of 0.7, and co-expression networks were created using MCL clusters. Once genes were assigned to these clusters, functionality was assigned using annotation based on GO enrichment and manual literature survey. Further, the conditions under which genes of a regulon are expressed were also examined for further validation using MetaOmGraph [15]. An important contribution of a regulon-based approach is that to understand a biological process, it is critical to be able to delineate its components, which are the elements involved in the regulation of a biological pathway. Thus, improved prediction of regulon information could facilitate better understanding of that process. Biologists could formulate testable hypotheses pertaining to the functional role of an unannotated gene, if regulon information is provided. Predicting regulon information for a functionally unannotated gene could give context to an experimental biologist in order to narrow down the conditions under which to test its potential function.

Regulon prediction requires a method that can learn from expression patterns of the genes under a plethora of conditions and apply it to functionally unannotated genes. Machine learning (ML) could be a powerful tool for analysis of large datasets, since it can learn from the complex relationships within a dataset and construct hypotheses that can explain these relationships [16].

Machine learning and several other artificial intelligence methods have used gene expression data to classify tumor datasets and cancer prognosis [17]–[20]. There could be a huge amount of associations that can be learnt from this data, which is not easily discernable. A machine learning method that uses only publicly available data hasn't been used to infer regulon information for functionally unannotated genes. The goal here is to harness the power of machine learning to tease out all linear and nonlinear relationships in the expression data which may have been overlooked by previous studies that may have focused on one particular kind of relationship.

In our method, we used a machine learning framework that uses regulon information as ground truth labels and RNA-seq data for *Arabidopsis thaliana* under a range of conditions to predict regulon information for functionally unannotated genes. We compared our method to existing methods that are used for functional annotation. Three methods utilize the three correlation coefficients: Pearson's, Spearman's and Kendall Tau's Correlation. A fourth method uses Markov Chain Clustering (MCL) [22]. The fifth method uses clustering using RenEEL which is based on modularity maximization [23].
This work aims to address the problem of functionally characterizing unannotated genes in *Arabidopsis thaliana* for which conventional methods of homology based functional characterization may not work.

Under Aim 2, I propose to implement an ensemble-based machine learning classification method, in collaboration with Sagnik Banerjee, that could enhance orphan protein secondary structure prediction. Deciphering the structure and conformations that a protein takes up, could accentuate the understanding of its function [24]–[26]. In its most natural state, amino acid residues of a protein can fold into one of three secondary structures, either alpha helices, beta sheets or nonregular coils [28]. The local conformations that an amino acid exhibits is its secondary structure assignment. Hydrophobic forces and side chain interactions, such as hydrogen bonding between amino acids helps bring together the secondary structure elements to form the tertiary structure. The tertiary structure of a protein is described by the coordinates of all the atoms in the protein or by the coordinates of the backbone atoms [26], [28]. The problem of accurate tertiary structure prediction is challenging even with advances in machine learning and deep learning methods [26], [29], [30], so improvements in secondary structure prediction can eventually help with better tertiary structure prediction.

Most initial studies started with solving the Q3 problem. Anfinsen's legacy states that all the information to predict the three-dimensional structure is contained within the primary amino acid sequence [27]. The first-generation methods used stereochemical properties of the amino acids [31], propensities of the amino acids to be in a specific secondary structure [32] and statistical models that use the frequency of the amino acids in secondary structures[33]. An important shift in the secondary structure prediction methods was bought about by the understanding that evolutionary information of an amino acid could increase accuracy [34]. PSI-PRED, one of the predominant secondary structure prediction softwares, utilizes PSI-BLAST to construct Position Specific Scoring Matrices (PSSMs) [35], [36], which represents evolutionary information, extracted from alignments of multiple homologous sequences [34]. With the burgeoning amount of sequence data being available over the years, there has been an increase in methods that use PSSMs as primary features [24], [34], [37]–[41]. Many prediction servers have also been developed over the years which utilize homology based information [41], [42]. Ab initio methods such as threading and Rosetta further use homology comparisons and extend it by using molecular modeling [43].

On the machine learning front, over the years, with algorithmic developments, use of neural networks have been shown to increase performance. Also, machine learning classifiers such as support vector machines (SVM) [24], [44], [45], hidden markov models (HMM) [46], [47], random forests [48]–[50] have been used extensively. Most of the features that have been utilized in these methods have been PSSMs. Some other methods have also been developed that use structural similarity based features in neural networks [51], [52].

We propose to develop an ensemble-based machine learning method that exclusively uses non homology-based features to predict the secondary structure of a protein. This is especially important in cases of orphan proteins that don't bear similarity with any other sequence. In addition to using PSSMs as features, over the years, there have been additional features derived from the amino acids such as physicochemical properties of the residues [24], [37], [38], [53]–[55] and the frequency of amino acids [38], [56]. However, all these are used in conjunction with PSSMs. These methods are not optimized for orphan protein secondary structure prediction. Further, the physicochemical properties used have consistently been the seven representative physicochemical properties as deciphered in [57]. However, since then, there have been many more physicochemical properties discovered in the literature which have not been utilized for secondary structure prediction.

Orphan protein secondary structure prediction will only be successful if there is more emphasis on finding the accurate features for prediction, since homology-based features are not useful. AAindex is a database that houses numerical indices that represent various physicochemical and biochemical properties of amino acids [58], [59]. All the data in this database has been derived from recently published literature We propose to solve the Q3 and consequently, the Q8 prediction problem using an ensemble-based machine learning method that uses physicochemical properties derived from recent published literature.
This work aims to use a machine learning method using non-homology-based features for **prediction of secondary structure of orphan proteins**

Under Aim 3, I aim to conduct an expression analysis study of candidate orphan genes in *Arabidopsis thaliana.* Orphan genes are those genes which code for proteins that bear no homology to proteins in any other species. These genes are important on many levels because it helps understand how an organism adapts to its ever-dynamic environment [60], [61]. In plants, for example, these genes may confer traits that may help it survive in drought stressed or temperature stressed conditions [61], [62]. Furthermore, it could also enhance traits that could help with the increasing food crisis that is faced by the world where there are dwindling resources and increasing populations. For example, the QQS orphan gene in *Arabidopsis thaliana* was found to increase protein content and decrease starch content by regulating carbon and nitrogen partitioning [63], [64]. When this orphan gene was introduced into soybean, rice and maize, it brought about the same increase in protein content and decrease in starch content [63]. This could increase the amount of protein that can be obtained from consuming plant-based foods and contribute widely to sustainability. There have been many studies that have been done to identify orphan genes using RNA-seq data [65], synteny [66], [67] and phylostratigraphy [68]. I propose to use expression data further to be able to functionally characterize the candidate orphan genes identified in *Arabidopsis thaliana*.

Narrowing down the conditions that they are expressed in, can help give context to a biologist who could validate the functions of these candidate orphan genes in the plant. Using the RNA-seq data from myriad conditions, one can investigate the expression patterns of these genes. Suppose one observes that an orphan gene is being upregulated in ten different samples that are sequenced under the same condition, for example cold stress, one can hypothesize that this gene may be involved in responding to this particular condition, specifically cold stress [69]. This also holds true for when investigating a particular tissue or developmental stage. Donoghue et al. in their paper have further shown that orphan genes in *Arabidopsis thaliana* showed upregulation in abiotic stress conditions [70].

For this aim, I will be carrying forward the work conducted by Jing Li in the lab. She has utilized her pipeline to identify unannotated candidate orphan genes in *Arabidopsis thaliana* using RNA-seq evidence and ab initio methods. I propose to investigate the expression of these candidate orphan genes and study their associations with other genes under several different conditions. Exploration of this data will be carried out in MetaOmGraph [15], also known as MOG. MOG is a java software that can be used to interactively explore and visualize large datasets. It incorporates a lot of features, such as correlation analysis, differential expression analysis, mutual information generation and distance computation. I propose to further use clustering and Gene Ontology (GO) enrichment which could enrich our understanding of the conditions that these candidate orphan genes express in. Additionally, I propose to apply the optimized machine learning method from Aim 1 to predict regulon information for these genes.
This work aims to examine the expression patterns of candidate orphan genes in *Arabidopsis thaliana* across different conditions. Along with regulon information, gene ontology information and clustering, it could provide a rich context to a biologist to develop testable hypotheses for potential functions of these genes.

Innovation
The proposed project aims to infer regulon information for functionally unannotated genes in *Arabidopsis thaliana* using a machine learning method utilizing expression data as features and regulons as labels. Obtaining regulon information for functionally unannotated genes in *Arabidopsis thaliana* could narrow down the conditions under which to test the function of the gene. Using expression data in a machine learning approach can tease out all relationships in the data which other methods may miss out on. The proposed novel method relies on publicly available expression data and regulon information of genes available in the public domain.

The proposed project will provide an ensemble-based machine learning method that can be used for protein secondary structure prediction that uses non-homology-based features which can be eventually validated on orphan proteins. Although protein secondary structure prediction has been carried out over the years, evolutionary information has been one of the predominant features in most machine learning methods used. Secondary structure prediction of orphan proteins would be difficult with these features. A novel ensemble-based machine learning method that uses non-homology-based features would be developed to enhance prediction of orphan protein secondary structure.

The proposed project will enrich understanding of the conditions in which orphan genes in Arabidopsis thaliana express and provide a contextual setting to study their potential function. After identification of orphan genes, in order to functionally annotate them, it's important to study what conditions these genes express under. For a biologist who can further validate them and their function, it's imperative to narrow down the conditions under which to test them. Prediction of regulon information using the machine learning method from Aim 1and Gene Ontology enrichment studies could provide more insight into their potential function.

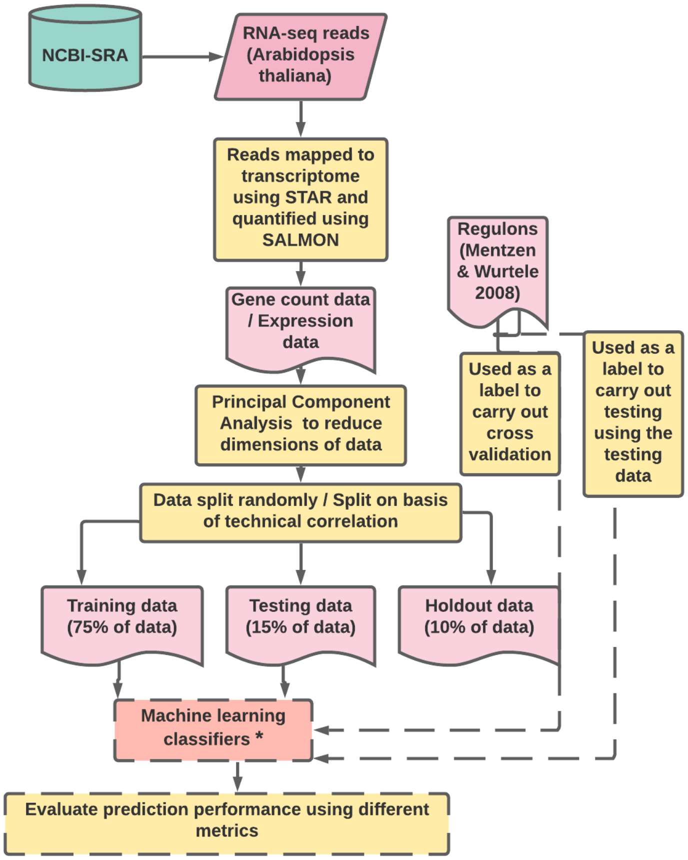
NCBI-SRA
RNA-seq reads (Arabidopsis thaliana)
Reads mapped to transcriptome using STAR and quantified using SALMON
Gene count data / Expression data
Regulons (Mentzen & Wurtele 2008)
Principal Component Analysis to reduce dimensions of data
Used as a label to carry out cross validation
Used as a label to carry out testing using the testing data
Data split randomly / Split on basis of technical correlation
Training data (75% of data)
Testing data (15% of data)
Holdout data (10% of data)
Machine learning classifiers *
Evaluate prediction performance using different metrics

Fig 1: An overview of the approach for using machine learning classifiers to predict regulon information for unannotated genes using expression data as features and regulons as labels

Approach

Aim 1: Construct a machine learning framework that uses RNA-seq expression-based features to classify functionally unannotated protein-coding genes in *Arabidopsis thaliana* into regulons.

To achieve this, I downloaded all publicly available paired-end transcriptomic RNA-seq reads for *Arabidopsis thaliana* Col-0 from NCBI-SRA. I aligned the 5210 samples obtained to the transcriptome using the ultra-fast aligner, STAR [71]. I quantified these alignments using the SALMON software [72] and the Transcripts Per Million (TPM) file was used for further analysis.

To check the feasibility of using expression data as features in the machine learning method, normalized microarray data compiled by Wieslawa Mentzen [3] were used as features. Regulons discovered in the same paper were utilized as features. For the initial classification on the microarray data, we used SVM classifier and photosynthesis regulon was used as labels. Seeing the promising results, I collaborated with Sagnik Banerjee for the rest of this project. We developed a machine learning method with the RNA-seq samples as features. While the microarray data has 22,746 genes, the RNA-seq data has 32,833 genes. The labels used for the final machine learning method were the regulons inferred from the 2008 paper [3]. We define a regulon as large (>400 genes), medium (100-400 genes) or small (<100 genes). For the machine learning method, we used five machine learning classifiers for classification: Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), Random Forest (RF), Multi-layer Perceptron (MLP), Gradient Boosting Trees (GBT). Fig. 1 is an overview of the approach used to build the machine learning framework.

We used Principal Component Analysis (PCA) to reduce the dimensionality of the data, such that the machine learning method can be trained and tested using lesser number of components and still retain the most variability in the data. It was found that ~250 samples explained 95% of the variability in the data. For the analysis, we used 20-100, 150, 200, 250 and 500 components.

An important consideration when developing a machine learning method is the approach used to split data into training, testing and holdout datasets. We split the data randomly into these datasets but figured that there could be potential redundancies between the datasets. Here, redundancy can be explained by considering two genes that are functionally unrelated but are technically correlated due to multi-mapping by the aligner. Several genes may operate in unison to carry out specific biological functions and could exhibit high correlation of gene counts. This could assist in the recognition of genes involved in the same pathway. However, functionally unrelated but positionally overlapping genes that share sequence similarity could be a precursor to incorrect gene quantification. Reads mapping to common regions of two genes may contribute to the counts of both genes even when only one of them expresses. This is due to multi-mapping and could spike up the read counts of highly similar genes. The count data obtained after quantification could potentially represent the reads mapping due to sequence similarity also and not due to expression of the gene alone. This leads to features that have both technical and biological correlation among them. If one of these genes belongs to a regulon and is in the training set, a highly similar gene which is in the testing set could also be assigned to the same regulon just because of technical correlation. This could potentially affect the performance of the machine learner. To identify such cases, we propose using a combination of sequence similarity search (BLAST) and probing of the number of reads mapping to both genes that share nucleotide similarity. To identify gene pairs that have a high sequence similarity, we propose using an all-vs-all BLAST search with transcript sequences of *Arabidopsis thaliana*. Next, we would compute the number of reads that map to both the gene pairs and divide it by the number of reads that map to either of the genes. This ratio, called Jaccard index, would be calculated for each of the samples for each of these highly similar genes. However, each sample has its own technical effects. For a particular gene pair, the Jaccard index obtained from one sample can't be compared to the Jaccard index from other samples. We could compare several Jaccard index thresholds to select the gene pairs that will be considered for making the final split. If we use a lower threshold, we could eradicate more of the technical correlations. Eventually, to construct non redundant datasets such that technical correlation is low, the training, testing and holdout datasets will be constructed by stratifying using the pairs of genes that have Jaccard indices above a certain threshold. Stratification will be done so as to keep these technically correlated genes together, in either training, testing or holdout such that performance of the machine learner is not affected. This is an empirical approach that could potentially remove technical biases manifested through multi-mapping. We will conduct an initial analysis with 25 RNA-seq samples with highest expression. This would eventually be carried out using all 5210 samples that were used to construct the features

For comparing the machine learning method to already used methods, we used direct correlation methods and clustering methods such as Markov Chain Clustering (MCL) [22] and RenEEL [23].These are methods which have been used to functionally characterize unannotated genes. For the direct correlation method, we computed three correlation coefficients (Pearson's, Spearman's and Kendall-Tau’s). Since the sampling distribution of these correlations are not normally distributed, Fisher's transformation was used to convert the skewed distributions of the sample correlation r into a distribution that is approximately normal. Fisher's z transformation for correlation coefficient r_{ij} between gene i and j is

$$z_{ij} = arctanh(r_{ij})$$

For iid (independent and identically distributed data) bivariate normally distributed data, z_{ij} is approximately normal. This observation can then be used to formulate a standardized measure of association between a gene $G_i$and a regulon R.

$$d_{corr}(G_i) = \frac{\overline{z}_{i+} - \overline{z}_{i-}}{s_p\sqrt{\frac{1}{|T_R^+|} + \frac{1}{|T_R^-|}}}$$

Equation 1

$$\overline{z}_{i+} = \frac{1}{|T_R^+|}\sum_{j\in T_R^+} z_{ij} \quad and \quad \overline{z}_{i-} = \frac{1}{|T_R^-|}\sum_{j\in T_R^-} z_{ij}$$

and

$$s_p = \sqrt{\frac{(|T_R^+| - 1)s_+^2 + (|T_R^-| - 1)s_-^2}{|T_R^+| + |T_R^-| - 2}}$$

for sample variances

$$s_+^2 = \frac{1}{|T_R^+| - 1} \sum_{j \in T_R^+} (z_{ij} - \bar{z}_{i+})^2 \quad and \quad s_-^2 = \frac{1}{|T_R^-| - 1} \sum_{j \in T_R^-} (z_{ij} - \bar{z}_{i-})^2$$

$|T_R^+|$ is the training set of genes in the regulon,
$|T_R^-|$ is the training set of genes not in the regulon
The degree of association of each gene in the testing dataset was compared to the ground truth in order to estimate the different performance metrics.

The three correlations (computed in the previous step) and thresholds from 0.1 to 0.9 with increments of 0.1 was used for MCL. We varied one of the parameters from 0.1 to 0.5 with increments of 0.1 to produce various MCL clusters. We analyzed each set of clusters to assess the predictive ability. Once the clusters were reported by MCL, we probed each cluster to obtain the proportion of genes in that cluster that belonged to the positive training set. This proportion was used as a degree of association for genes belonging to the same cluster but residing in the testing dataset. The proportion was calculated as

$$d_{mcl}(G_i) = \frac{|\{g \in C(G_i) : g \in T_R^+\}|}{|\{g \in C(G_i) : g \in T_R^+\}| + |\{g \in C(G_i) : g \in T_R^-\}|}$$

Equation 2

$C(G_i)$ = cluster that contains gene G_i
$|T_R^+|$ is the training set of genes in a regulon
$|T_R^-|$ is the training set of genes not in regulon
The degree of association of each gene in the testing dataset was compared to the ground truth in order to estimate the different performance metrics.
Different components of the expression data were supplied to these two methods in order to accurately compare them to the machine learning method.

Results for Aim 1:
For the machine learning method, we propose to remove the technical correlations in the data for testing on the final holdout set. Here, I present the results of the machine learning method where the datasets are split randomly and the test dataset was used for testing. The training, testing and holdout datasets were split randomly to make up 75%, 15% and 10% of the data. We used stratified 5-fold cross validation to optimize the various parameters of the machine learning classifiers. The five metrics which we used to assess performance were the precision, recall, Matthews correlation coefficient (MCC) and AUCPR score (see Appendix). For each regulon, different PCA components were trained and tested using the different classifiers to assess prediction. Here, I present the results for the largest regulon (photosynthesis) and a medium regulon (embryo maturation fruit and seed preferential).

Components	Classifiers	Precision		Recall		F1-score		MCC		AUC-PR	
		P	EM	P	EM	P	EM	P	EM	P	EM
40	GBT	**0.86**	0.83	**0.67**	0.38	**0.75**	0.52	**0.75**	0.55	**0.86**	0.51
	MLP	**0.87**	0.67	**0.73**	0.44	**0.79**	0.53	**0.79**	0.54	**0.85**	0.54
70	GBT	0.85	0.77	0.69	0.38	0.76	0.51	0.78	0.53	0.81	0.48
	MLP	**0.86**	0.59	**0.71**	0.47	**0.78**	0.52	**0.79**	0.52	**0.85**	0.45
150	GBT	0.77	0.71	0.69	0.38	0.73	0.49	0.72	0.51	0.70	0.46
	MLP	0.8	0.55	0.76	0.44	0.78	0.49	0.77	0.48	0.82	0.48
500	GBT	0.82	0.76	0.67	0.34	0.74	0.47	0.73	0.51	0.80	0.46
	MLP	0.84	0.60	0.70	0.47	0.77	0.53	0.76	0.52	0.75	0.49

Table 1: Performance metrics for photosynthesis regulon (P) and embryo maturation fruit and seed preferential regulon (EM) using the machine learning method. Comparison is shown among the range of components used and the two best classifiers that showed the best performance across components and regulons. The metrics for best performing components has been marked in red.

Machine learning results: For the photosynthesis regulon which is the biggest regulon with 1162 genes, table 1 shows the results with testing carried out with the test dataset. For better presentation of the results, four components were chosen over the range of components used. As is seen in the table, increasing components doesn't necessarily improve the performance. The metrics used to assess performance show consistent performance with 40 components as well as with 500 components. Among all five classifiers used, GBT and MLP performed well for all components, and hence I have only shown the metrics for these two classifiers. Fig. 2 shows the area under the curve for precision and recall using different thresholds for photosynthesis. The best performance was achieved with 40 components using the GBT and MLP classifier which is shown in table 1 as well as in figure 2b.

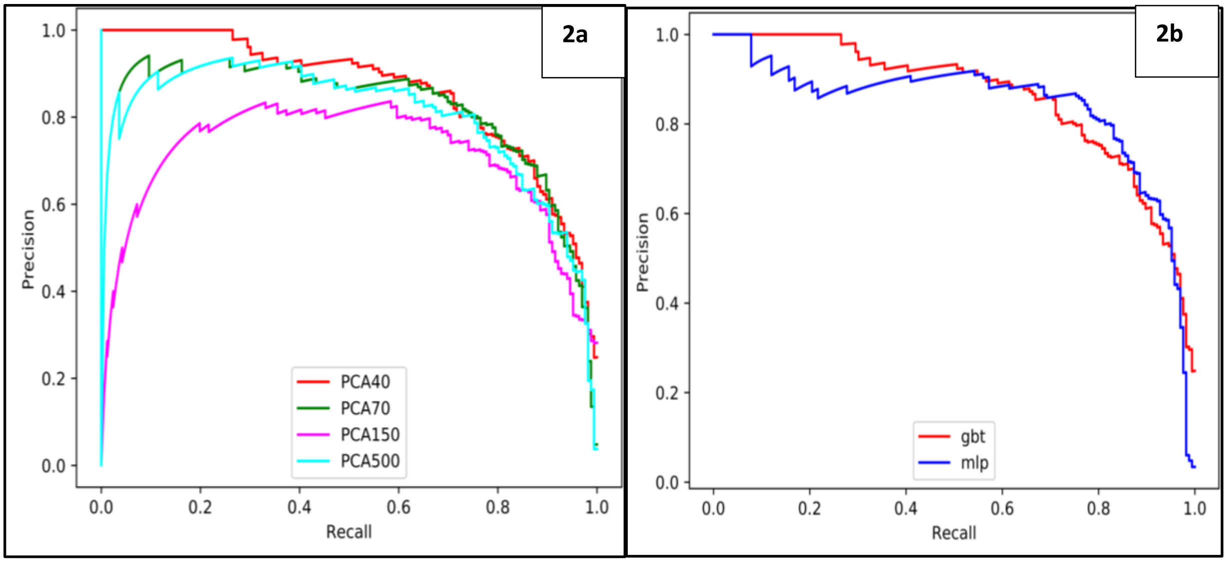

Fig. 2: The AUC-PR curves for the photosynthesis regulon using the machine learning method. 2a shows the AUC-PR curves for the four components in table 1. 2b shows the AUC-PR curves for the best performing classifiers with best performing component.

For the medium regulon, embryo maturation fruit and seed preferential with 430 genes, table 1 shows the performance metrics with the same components used for testing with the photosynthesis regulon. The metrics for none of the components for this regulon were encouraging. This may be because the genes in the medium and small regulons have fewer genes and the machine learner may not be able to distinctly classify genes into regulons from learning from such a small positive set and a huge negative set. We propose to use a multi-class classification approach for training and testing using these medium and small regulons. With a multi class approach, one could provide genes from large and medium regulons as labels, the information of which could help better classify the genes belonging to the smaller regulons.

Markov Chain Clustering result: Markov Chain Clustering (MCL) is currently running for the different components. However, we had count data for RNA-seq samples downloaded in November 2019. Initially, 2944 samples were aligned to the transcriptome using STAR and quantified using Salmon. MCL clustering was carried with all these samples and equation 2 (from Approach above) was used to assess its performance in predicting genes belonging to photosynthesis regulon. The same amount of data was provided to the machine learning method. Fig. 3 shows the ROC curves comparing the performance of the machine learning method with the MCL method. The graph also displays the area under the curve ROC score for both in order to compare the performance quantitatively. One can see the stark difference in prediction performance between the machine learning method and MCL.

Outcomes: The outcome would be a machine learning framework that would enrich regulon information for functionally unannotated genes. Classification of genes into regulons could offer insight to develop hypotheses for a potential context to validate their function. Since this method uses expression data and no homology-based features, this method can be applied to gain information on genes whose function is difficult to infer through sequence similarity. Further, we propose to incorporate a method in the model that could remove technical correlation in the expression data manifested through multi-mapping due to the aligner used.

Challenges: An important challenge is the technical correlations in the expression data. These can be caused by multi-mapping (which we propose to address), and other technical per sample-biases such as GC-bias, 3' end mapping bias, amplification bias, random hexamer priming bias, sequencing bias. These latter biases may be difficult to identify and remove. This could confound the biological correlation in the data that could impact the performance of the machine learner.

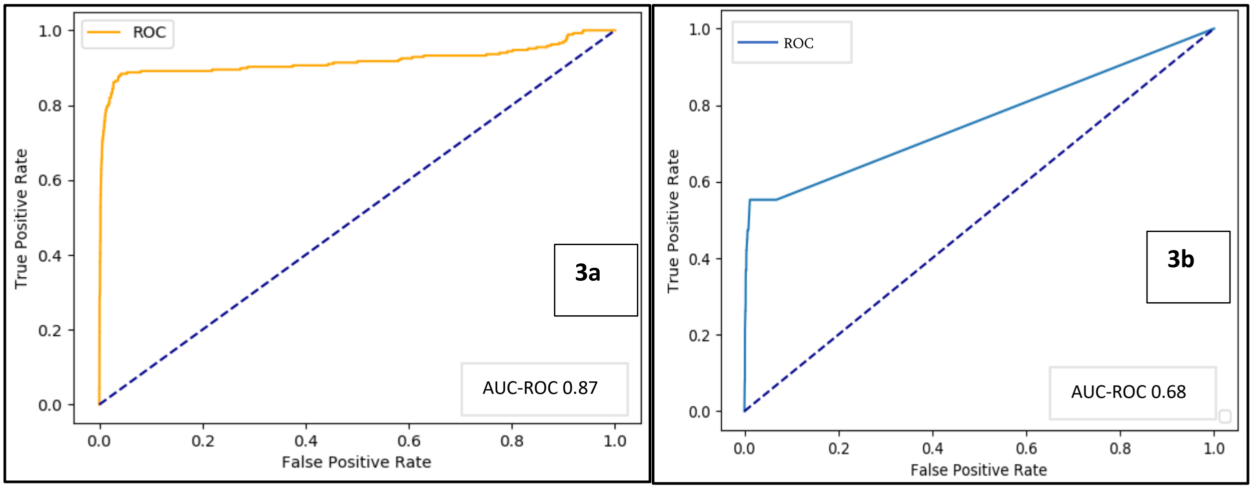

Fig. 3: The AUC-ROC curves for the photosynthesis regulon. 3a shows the AUC-ROC curve using the random forest classifier and 3b shows the AUC-ROC curve using the markov chain clustering method.

Aim 2: **Develop an ensembl-based machine learning method to predict secondary structure of orphan proteins using non-homology-based features**

To solve the Q3 prediction problem, I downloaded the sequences of all available structures for *Arabidopsis thaliana* from Protein DataBank (PDB). I obtained the corresponding fasta sequences and the secondary structure assignments for these sequences from DSSP (Dictionary of Secondary Structure of Proteins). For the machine learning model, the assumption here is that the secondary structure of an amino acid residue depends on the residues in its proximity. For this aim, I developed the features in collaboration with Sagnik Banerjee. For the machine learning model, the features we used are the physiochemical properties of the residues lying in a constant scope, which here are represented in the form of a sliding window. To select the most important features from the physicochemical properties, we propose to use datasets that has been built using fuzzy clustering of the physicochemical and biochemical properties of amino acids [73] from the AAIndex database [58], [59]. The aforementioned paper uses consensus fuzzy clustering (CFC) to find 8 fuzzy clusters which were used to construct 3 high quality datasets (HQI8, HQI24 and HQI40). For HQI8, the dataset contains the medoids of the 8 clusters that were constructed. For HQI24, two indices farthest from the cluster medoid was selected for each cluster such that diversity of an amino acid is captured. For HQI40, the two indices closest to the medoid was selected for each cluster.

For a preliminary feasibility study, we used window sizes corresponding to odd numbers ranging from 9-23. We considered the label as the three-class classification (Helix (H) / Beta sheet (E) / Coiled Coil (C)) of the peptide residues in the middle of the window. The corresponding features used were the eight physicochemical properties under the HQI8 indices for all the residues spanning the window. For example, for a window of size 9 is used, the label would be the secondary structure assignment of the fifth peptide residue while the features would be the physicochemical properties corresponding to the 8 indices for all the residues (8 X 9 = 72 features) in the window. Fig. 4 illustrates the construction of these features.

For the Q3 prediction problem, we carried out a k fold cross validation where the number of cross validations was 5. Stratified cross validation was carried out to divide the dataset in turn into training and testing datasets while keeping the amount of data in each fold consistent. The classifiers used here were Stochastic Gradient Descent (SGD), Random Forest (RF) and Gradient Boosting Trees (GBT). The metrics which we used to assess performance were the precision, recall and the F1-score.

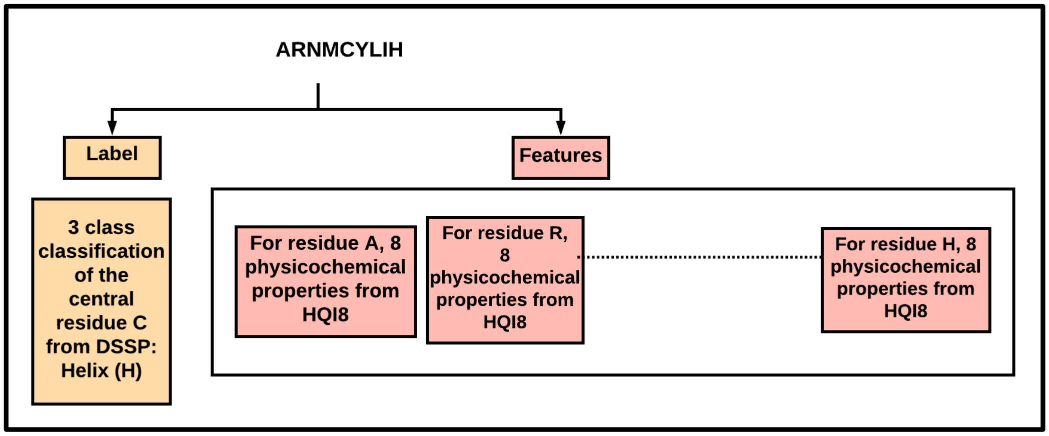
ARNMCYLIH
Label
Features
3 class classification of the central residue C from DSSP: Helix (H)
For residue A, 8 physicochemical properties from HQI8
For residue R, 8 physicochemical properties from HQI8
For residue H, 8 physicochemical properties from HQI8

Fig. 4: The label and features for a peptide sequence with window size of nine

Preliminary results for Aim 2:

For the preliminary results, I present in table 2, the performance metrics of the machine learning method with the varying odd window sizes from 9-23. With varying window sizes, it was seen that there wasn't a marginal difference between very close window sizes. For the alpha helix classification, if one compares window size 9 with 23, there is an improvement in the F1 score. One could conclude that window size does affect classification. If the classification of alpha helix is compared to that of beta sheet, there is marked difference in performance. Alpha helices depend on interactions between local residues while beta sheets are formed by non-local interactions which explains why it's easier to predict alpha helices. 50% of residues in a peptide sequence form the alpha helices and beta sheets while the rest form coiled coils [74]. The overall performance can be improved by predicting beta sheets better, which would automatically help predict coiled coils better.

Window size	Class predicted	Recall	Precision	F1-score
9	Helix	0.77`	0.60	0.67
	Beta-sheet	0.35	0.57	0.42
	Coiled coil	0.59	0.59	0.59
11	Helix	0.78	0.61	0.68
	Beta-sheet	0.37	0.60	0.43
	Coiled coil	0.60	0.60	0.60
13	Helix	0.79	0.62	0.70
	Beta-sheet	0.38	0.62	0.45
	Coiled coil	0.60	0.60	0.60
15	Helix	0.80	0.61	0.70
	Beta-sheet	0.38	0.63	0.46
	Coiled coil	0.61	0.61	0.61
17	Helix	0.80	0.62	0.71
	Beta-sheet	0.39	0.64	0.47
	Coiled coil	0.61	0.62	0.61

19	Helix	0.81	0.62	0.71
	Beta-sheet	0.40	0.65	0.48
	Coiled coil	0.61	0.62	0.62
21	Helix	0.81	0.62	0.71
	Beta-sheet	0.40	0.66	0.50
	Coiled coil	0.62	0.62	0.62
23	Helix	**0.82**	**0.63**	**0.71**
	Beta-sheet	0.41	0.67	0.49
	Coiled coil	0.62	0.63	0.63

Table 2: Cross validation results using window sizes 9-23 for secondary structure classification of central residues of the windows of peptide sequences.

For future work, we propose to incorporate more non homology-based features such as the physicochemical properties from the other high-quality indices (HQI24 and HQI40) [58][59], [73], the unigram, bigram and trigram frequencies of amino acids and the positions of the corresponding amino acid in the peptide sequence. We also propose to further develop the method with these features to predict the eight classes of secondary structure. From earlier studies, it has been shown that using neural networks for training performed better than classifiers such as Support Vector Machine (SVM) and trees-based classifiers such as Random Forests (RF) and Gradient Boosting Trees (GBT). We propose to use Multilayer perceptrons, and train them using the non-homology-based features mentioned above. Further, I propose to use ensemble methods that work by incorporating the results from many classifiers in such a way to produce one optimal predictive model. Each classifier has its own drawbacks and ensemble methods work by combining the classifiers in such a way to reduce the errors of individual classifiers and combining the best of the classifiers. Several ensemble methods such as Boosting / AdaBoost, Bagging, Mixture of Experts could be used to improve prediction. Further, considering how the window approach performs with using more physicochemical properties, taking a consensus of the windows rather than using them individually, could also enhance performance. Also, for improving beta-sheet classification, I could also use longer window sizes which could take into consideration the non-local interactions.

Outcomes: An ensemble-based machine learning method using non-homology-based features that could enhance the prediction of orphan protein secondary structure would be the outcome. This method will be initially tested on sequences with known structures. It will then be further tested on orphan protein sequences in order to predict their secondary structure.

Challenges: There are no 3D structures available in the public domain for orphan proteins in *Arabidopsis thaliana*. Until such structures are not available, there will be no ground truth available and validation of this method for orphan secondary structures will be difficult. This method will be tested on orphan protein sequences and predictions will be made, but we won't have any available structures to corroborate our predictions. Once few structures are available, we would have some ground truth for validation.

Aim 3: **Conduct meta-analysis with expression data using MetaOmGraph and predict regulon information employing the machine learning framework from Aim 1 for characterization of candidate orphan genes identified in *Arabidopsis thaliana*.**

For this aim, I propose to utilize the publicly available RNA-seq data for *Arabidopsis thaliana* to investigate the expression levels of candidate orphan genes. The RNA-seq data in NCBI-SRA have been collected from a range of spatial, temporal, developmental and stress conditions and could help unravel functions of candidate orphan genes that express only in some conditions. Observing co-expression patterns of the candidate orphan genes could shed light into the functions or regulatory pathways that they may be involved in.

To achieve this goal, I downloaded 8765 paired-end, illumina transcriptomic RNA-seq reads for *Arabidopsis thaliana* from NCBI-SRA and aligned them to the genome using STAR [71], [75]. STAR is a relatively fast aligner which performs well. The reads will be then quantified using SALMON [72]. The metadata for the

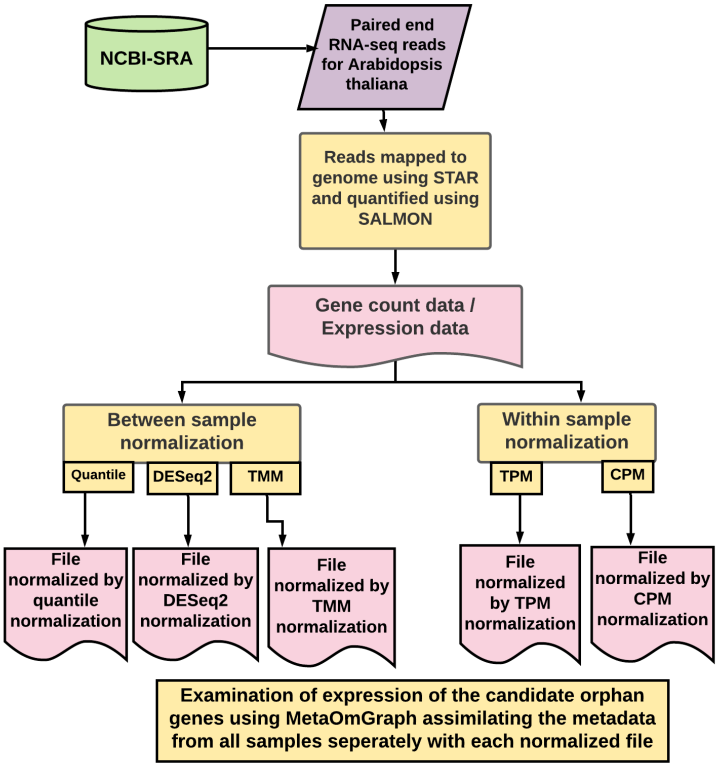
NCBI-SRA
Paired end RNA-seq reads for Arabidopsis thaliana
Reads mapped to genome using STAR and quantified using SALMON
Gene count data / Expression data
Between sample normalization
Within sample normalization
Quantile
DESeq2
TMM
TPM
CPM
File normalized by quantile normalization
File normalized by DESeq2 normalization
File normalized by TMM normalization
File normalized by TPM normalization
File normalized by CPM normalization
Examination of expression of the candidate orphan genes using MetaOmGraph assimilating the metadata from all samples seperately with each normalized file

Fig 5 : An overview of the approach for carrying out expression analysis on candidate orphan genes using MetaOmGraph

corresponding studies will be downloaded using epost under Entrez Direct in NCBI. The expression data will be normalized using both between-sample and within-sample normalization. For within sample normalization, I will be using transcripts per million (TPM) normalization and counts per million (CPM) normalization (see Appendix). Between sample normalization would be conducted using quantile normalization, DESeq2 and trimmed mean of M-values (TMM) (see Appendix). Gene metadata downloaded from Ensembl Biomart will be used. Each of these normalized files, along with gene metadata and metadata for each of the samples will be utilized in MOG to explore and analyze the expression patterns of the candidate orphan genes with evidence of transcription provided by Jing's pipeline. Further, the machine learning approach from Aim 1 will be applied here to predict regulon information for these candidate orphan genes using the expression data from 8765 samples. Co-expression analysis could help narrow down the potential function of these transcribing orphan genes. Examining the expression of an orphan gene under various experiments of the similar condition can give an insight into its function which can be further validated by the clusters that it belongs to. This can help in deciphering the module that it may be functional in. I will also carry out a GO enrichment analysis to further narrow down the contexts under which these candidate orphan genes express. Fig. 5 shows an overview for the approach used to generate a dataset for conducting meta-analysis using MetaOmGraph (MOG).

Outcomes: For the list of unannotated and annotated candidate orphan genes, a set of conditions under which they would tend to be expressed would be obtained. Also, regulon information for these genes would be predicted using the machine learning method from Aim 1. Further, if there is available Ribo-seq data, it could provide translational evidence for these genes. This would provide context to a biologist to develop testable hypotheses for the potential functions of these candidate orphan genes.

Challenges: Using MOG, there are several between-sample and within-sample normalization techniques being applied. However, there could be other batch effects and per sample biases that could confound studying the expression of these candidate orphans across multiple different samples. Also, for applying the machine learning method, the same challenges mentioned in Aim 1 would apply.

References:

ADDIN Mendeley Bibliography CSL_BIBLIOGRAPHY [1] P. Poczai, I. Cernák, I. Varga, and J. Hyvönen, "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.," *Genet. Resour. Crop Evol.*, vol. 61, no. 1, pp. 796–815, 2014.

[2] S. Uygun, C. Peng, M. D. Lehti-Shiu, R. L. Last, and S. H. Shiu, "Utility and Limitations of Using Gene Expression Data to Identify Functional Associations," *PLoS Comput. Biol.*, vol. 12, no. 12, pp. 1–27, 2016.

[3] W. I. Mentzen and E. S. Wurtele, "Regulon organization of arabidopsis," *BMC Plant Biol.*, vol. 8, pp. 1–22, 2008.

[4] Y. Kodama, M. Shumway, and R. Leinonen, "The sequence read archive: Explosive growth of sequencing data," *Nucleic Acids Res.*, vol. 40, no. D1, pp. 2011–2013, 2012.

[5] S. van Dam, U. Võsa, A. van der Graaf, L. Franke, and J. P. de Magalhães, "Gene co-exprecoexpressions for functional classification and gene-disease predictions," *Brief. Bioinform.*, vol. 19, no. 4, pp. 575–592, 2018.

[6] A. Bhar, M. Haubrock, A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and E. Wingender, "CoexpresCoexpressionegulation analysis of time-series gene expression data in estrogen-induced breast cancer cell," *Algorithms Mol. Biol.*, vol. 8, no. 1, pp. 1–11, 2013.

[7] Z. Zhu *et al.*, "Co-expression Network Analysis Identifies Four Hub Genes Associated With Prognosis in Soft Tissue Sarcoma," *Front. Genet.*, vol. 10, no. February, pp. 1–10, 2019.

[8] Z. Gerring, E. Gamazon, and E. Derks, "A Gene Co-expression Network-based Analysis of Multiple Brain Tissues Reveals Novel Genes and Molecular Pathways Underlying Major Depression," *bioRxiv*, p. 591693, 2019.

[9] S. M. Salleh, G. Mazzoni, P. Løvendahl, and H. N. Kadarmideen, "Gene co-exprecoexpressions from RNA sequencing of dairy cattle identifies genes and pathways affecting feed efficiency," *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–15, 2018.

[10] X. Ma *et al.*, "Co-expression Gene Network Analysis and Functional Module Identification in Bamboo Growth and Development," *Front. Genet.*, vol. 9, no. November, pp. 1–15, 2018.

[11] H. Dai, J. Zhou, and B. Zhu, "Gene co-exprecoexpression analysis identifies the hub genes associated with immune functions for nocturnal hemodialysis in patients with end-stage renal disease," *Med. (United States)*, vol. 97, no. 37, pp. 1–8, 2018.

[12] K. R. Jaglo-Ottosen, S. J. Gilmour, D. G. Zarka, O. Schabenberger, and M. F. Thomashow, "Arabidopsis CBF1 overexpression induces COR genes and enhances freezing tolerance," *Science (80-.).*, vol. 280, no. 5360, pp. 104–106, 1998.

[13] J. D. Keene and P. J. Lager, "Post-transcriptional operons and regulons co-ordinating gene expression," *Chromosom. Res.*, vol. 13, no. 3, pp. 327–337, 2005.

[14] A. Joshi, Y. Van de Peer, and T. Michoel, "Structural and functional organization of RNA regulons in the post-transcriptional regulatory network of yeast," *Nucleic Acids Res.*, vol. 39, no. 21, pp. 9108–9117, Aug. 2011.

[15] U. Singh, M. Hur, K. Dorman, and E. S. Wurtele, "MetaOmGraph: A workbench for interactive exploratory data analysis of large expression datasets," *Nucleic Acids Res.*, vol. 48, no. 4, p. E23, 2020.

[16] A. C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification.," *Appl. Bioinformatics*, vol. 2, no. 3 Suppl, pp. 1–10, 2003.

[17] D. G. P. van IJzendoorn, K. Szuhai, I. H. Briaire-De Bruijn, M. Kostine, M. L. Kuijjer, and J. V. M. G. Bovée, "Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas," *PLoS Comput. Biol.*, vol. 15, no. 2, pp. 1–19, 2019.

[18] Y. Yuan and Z. Bar-Joseph, "Deep learning for inferring gene relationships from single-cell expression data," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 116, no. 52, pp. 27151–27158, 2019.

[19] M. Pirooznia, J. Y. Yang, M. Q. Qu, and Y. Deng, "A comparative study of different machine learning methods on microarray gene expression data," *BMC Genomics*, vol. 9, no. SUPPL. 1, pp. 1–13, 2008.

[20] Y. Kong and T. Yu, "A Deep Neural Network Model using Random Forest to Extract Feature Representation for Gene Expression Data Classification," *Sci. Rep.*, vol. 8, no. 1, pp. 1–9, 2018.

[21] W. I. Mentzen, J. Peng, N. Ransom, B. J. Nikolau, and E. S. Wurtele, "Articulation of three core metabolic processes in Arabidopsis: Fatty acid biosynthesis, leucine catabolism and starch metabolism," *BMC Plant Biol.*, vol. 8, pp. 1–15, 2008.

[22] S. van Dongen, "Graph stimulation by flow clustering," *Graph Stimul. by flow Clust.*, vol. PhD thesis, p. University of Utrecht, 2000.

[23] J. Guo, P. Singh, and K. E. Bassler, "Reduced network extremal ensemble learning (RenEEL) scheme for community detection in complex networks," *Sci. Rep.*, vol. 9, no. 1, pp. 1–11, 2019.

[24] P. Chatterjee, S. Basu, M. Kundu, M. Nasipuri, and D. Plewczynski, "PSP-MCSVM: Brainstorming consensus prediction of protein secondary structures using two-stage multiclass support vector machines," *J. Mol. Model.*, vol. 17, no. 9, pp. 2191–2201, 2011.

[25] B. Zhang, J. Li, and Q. Lü, "Prediction of 8-state protein secondary structures by a novel deep learning architecture," *BMC Bioinformatics*, vol. 19, no. 1, pp. 1–13, 2018.

[26] C. M. Freeman, A. M. Gorman, and S. M. Levine, "Structure prediction," *Acta Crystallogr. Sect. A Found. Crystallogr.*, vol. 52, no. a1, pp. C91–C91, 1996.

[27] C. B. Anfinsen, "Principles that Govern the Folding of Protein Chains," *Science (80-.).*, vol. 181, no. 4096, pp. 223 LP – 230, Jul. 1973.

[28] J. D. Watson, Baker, Bell, Gann, Levine, and Losick, *Molecular Biology*. 2014.

[29] Y. Yang *et al.*, "Sixty-five years of the long march in protein secondary structure prediction: The final stretch?," *Brief. Bioinform.*, vol. 19, no. 3, pp. 482–494, 2018.

[30] M. Torrisi, G. Pollastri, and Q. Le, "Deep learning methods in protein structure prediction," *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 1301–1310, 2020.

[31] V. I. Lim, "Algorithms for Prediction of a-Helical and /I-Structural Regions in Globular Proteins," 1974.

[32] P. Y. Chou and G. D. Fasman, "Prediction of Protein Conformation," *Biochemistry*, vol. 13, no. 2, pp. 222–245, 1974.

[33] J. Garnier, J. F. Gibrat, and B. Robson, "[32] GOR method for predicting protein secondary structure from amino acid sequence," *Methods Enzymol.*, vol. 266, no. 1995, pp. 540–553, 1996.

[34] D. T. Jones, "<Jones 1999 - PSIpred.pdf>," pp. 195–202, 1999.

[35] L. J. McGuffin, K. Bryson, and D. T. Jones, "The PSIPRED protein structure prediction server," *Bioinformatics*, vol. 16, no. 4, pp. 404–405, 2000.

[36] D. W. A. Buchan and D. T. Jones, "The PSIPRED Protein Analysis Workbench: 20 years on," *Nucleic Acids Res.*, vol. 47, no. W1, pp. W402–W407, 2019.

[37] B. Zhang, J. Li, and Q. Lü, "Prediction of 8-state protein secondary structures by a novel deep learning architecture," *BMC Bioinformatics*, vol. 19, no. 1, p. 293, 2018.

[38] R. Heffernan, K. Paliwal, J. Lyons, J. Singh, Y. Yang, and Y. Zhou, "Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning," *J. Comput. Chem.*, vol. 39, no. 26, pp. 2210–2216, 2018.

[39] Y. Ma, Y. Liu, and J. Cheng, "Protein Secondary Structure Prediction Based on Data Partition and Semi-Random Subspace Method OPEN."

[40] Y. Loewenstein *et al.*, "Protein function annotation by homology-based inference.," *Genome Biol.*, vol. 10, no. 2, p. 207, 2009.

[41] A. Drozdetskiy, C. Cole, J. Procter, and G. J. Barton, "JPred4: A protein secondary structure prediction server," *Nucleic Acids Res.*, vol. 43, no. W1, pp. W389–W394, 2015.

[42] G. Yachdav *et al.*, "PredictProtein - An open resource for online prediction of protein structural and functional features," *Nucleic Acids Res.*, vol. 42, no. W1, pp. 337–343, 2014.

[43] S. Wu and Y. Zhang, "Protein structure prediction," *Bioinforma. Tools Appl.*, vol. 383, no. 2003, pp. 225–242, 2007.

[44] S. Hua and Z. Sun, "A novel method of protein secondary structure prediction with high segment overlap measure: Support vector machine approach," *J. Mol. Biol.*, vol. 308, no. 2, pp. 397–407, 2001.

[45] J. Guo, H. Chen, Z. Sun, and Y. Lin, "A Novel Method for Protein Secondary Structure Prediction Using Dual-Layer SVM and Profiles," *Proteins Struct. Funct. Genet.*, vol. 54, no. 4, pp. 738–743, 2004.

[46] K. Asai, S. Hayamizu, and K. Handa, "Prediction of protein secondary structure by the hidden Markov model," *Bioinformatics*, vol. 9, no. 2, pp. 141–146, Apr. 1993.

[47] Z. Aydin, Y. Altunbasak, and M. Borodovsky, "Protein secondary structure prediction for a single-sequence using hidden semi-Markov models," *BMC Bioinformatics*, vol. 7, no. 1, p. 178, 2006.

[48] C. Kathuria, D. Mehrotra, and N. K. Misra, "Predicting the protein structure using random forest approach," *Procedia Comput. Sci.*, vol. 132, no. Iccids, pp. 1654–1662, 2018.

[49] A. Dehzangi, S. Phon-Amnuaisuk, and O. Dehzangi, "Using random forest for protein fold prediction problem: An empirical study," *J. Inf. Sci. Eng.*, vol. 26, no. 6, pp. 1941–1956, 2010.

[50] X. Hu, H. Long, C. Ding, S. Gao, and R. Hou, "Using random forest algorithm to predict super-secondary structure in proteins," *J. Supercomput.*, vol. 76, no. 5, pp. 3199–3210, 2020.

[51] C. N. Magnan and P. Baldi, "SSpro/ACCpro 5: Almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity," *Bioinformatics*, vol. 30, no. 18, pp. 2592–2597, 2014.

[52] Z. Aydin, N. Azginoglu, H. I. Bilgin, M. Celik, and A. Valencia, "Developing structural profile matrices for protein secondary structure and solvent accessibility prediction," *Bioinformatics*, vol. 35, no. 20, pp. 4004–4010, 2019.

[53] E. Asgari, N. Poerner, A. McHardy, and M. Mofrad, "DeepPrime2Sec: Deep Learning for Protein Secondary Structure Prediction from the Primary Sequences," pp. 1–8, 2019.

[54] E. Saghapour and M. Sehhati, "FULL LENGTH Iranian Physicochemical Position-Dependent Properties in the Protein Secondary Structures," *Biomed. J.*, vol. 23, no. 4, pp. 253–261, 2019.

[55] G. Pok, C. H. Jin, and K. H. Ryu, "Correlation of Amino Acid Physicochemical Properties with Protein Secondary Structure Conformation," in *2008 International Conference on BioMedical Engineering and Informatics*, 2008, vol. 1, pp. 117–121.

[56] J. M. Otaki, M. Tsutsumi, T. Gotoh, and H. Yamamoto, "Secondary structure characterization based on amino acid composition and availability in proteins," *J. Chem. Inf. Model.*, vol. 50, no. 4, pp. 690–700, 2010.

[57] J. Meiler, M. Müller, A. Zeidler, and F. Schmäschke, "Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks," *J. Mol. Model.*, vol. 7, no. 9, pp. 360–369, 2001.

[58] S. Kawashima and M. Kanehisa, "AAindex: Amino acid index database," *Nucleic Acids Res.*, vol. 28, no. 1, p. 374, 2000.

[59] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa, "AAindex: Amino acid index database, progress report 2008," *Nucleic Acids Res.*, vol. 36, no. SUPPL. 1, pp. 202–205, 2008.

[60] A. R. Carvunis *et al.*, "Proto-genes and de novo gene birth," *Nature*, vol. 487, no. 7407, pp. 370–374, 2012.

[61] Z. W. Arendsee, L. Li, and E. S. Wurtele, "Coming of age: orphan genes in plants," *Trends Plant Sci.*, vol. 19, no. 11, pp. 698–708, Nov. 2014.

[62] C. Yao, H. Yan, X. Zhang, and R. Wang, "A database for orphan genes in poaceae," *Exp. Ther. Med.*, vol. 14, no. 4, pp. 2917–2924, 2017.

[63] L. Li *et al.*, "QQS orphan gene regulates carbon and nitrogen partitioning across species via NF-YC interactions," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 112, no. 47, pp. 14734–14739, 2015.

[64] L. Li and E. S. Wurtele, "The QQS orphan gene of Arabidopsis modulates carbon and nitrogen allocation in soybean," *Plant Biotechnol. J.*, vol. 13, no. 2, pp. 177–187, 2015.

[65] J. Li, Z. Arendsee, U. Singh, and E. Syrkin, "Recycling RNA-Seq Data to Identify Candidate Orphan Genes for Experimental Analysis," pp. 1–7, 2019.

[66] Z. Arendsee, J. Li, U. Singh, P. Bhandary, A. Seetharam, and E. S. Wurtele, "Fagin: Synteny-based phylostratigraphy and finer classification of young genes," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–14, 2019.

[67] Z. Arendsee, A. Wilkey, U. Singh, J. Li, M. Hur, and E. S. Wurtele, "Synder : Inferring Genomic Orthologs From Synteny Maps," *bioRxiv*, p. 554501, 2019.

[68] Z. Arendsee, J. Li, U. Singh, A. Seetharam, K. Dorman, and E. S. Wurtele, "Phylostratr : a Framework for Phylostratigraphy ," *Bioinformatics*, no. March, pp. 1–11, 2019.

[69] P. Bhandary, A. S. Seetharam, Z. W. Arendsee, M. Hur, and E. S. Wurtele, "Raising orphans from a metadata morass: A researcher's guide to re-use of public' omics data," *Plant Sci.*, vol. 267, pp. 32–47, Feb. 2018.

[70] M. T. Donoghue, C. Keshavaiah, S. H. Swamidatta, and C. Spillane, "Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana," *BMC Evol. Biol.*, vol. 11, no. 1, pp. 1–23, 2011.

[71] A. Dobin *et al.*, "Mapping RNA-seq with STAR," *Curr Protoc Bioinforma.*, vol. 51, no. 4, pp. 586–597, 2016.

[72] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, and C. Kingsford, "Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference," *Nat. Methods*, vol. 14, no. 4, p. 417, 2017.

[73] I. Saha, U. Maulik, S. Bandyopadhyay, and D. Plewczynski, "Fuzzy clustering of physicochemical and biochemical properties of amino Acids," *Amino Acids*, vol. 43, no. 2, pp. 583–594, 2012.

[74] M. Singh, "Predicting Protein Secondary and Supersecondary Structure," pp. 29-1-29–29, 2005.

[75] A. Dobin *et al.*, "STAR: Ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 201ses Dobin *et al.*, "STAR: Ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, 201

Appendix:

RNA-Seq Normalization methods

Normalization of RNA-seq data is done to remove differences between samples that arise due to library size or depth of sequencing, the platform used or other technical effects as well as within sample effects related to the gene length or the GC content. There are many methods available that normalize RNA-seq data.

Suppose there are M RNA-seq experiments, where M represents a single RNA-seq sample, which produces sequencing data for a particular RNA-seq sample. From these RNA-seq studies, we have M fastq files which have the short reads produced by the RNA-seq samples within the studies. These short reads can be mapped to the transcriptome of interest using an aligner software such as Salmon. After mapping the short reads from M fastq files to the transcriptome:

Suppose R_{ij} is the number of reads that mapped to the ith transcript from jth samples. Let $Xj = \Sigma_i R_{ij}$ be the total number of reads that mapped to the transcriptomes from jth sample. This is also called the library size of the jth sample.

TPM, CPM, FPKM are some methods used for normalizing for depth and length within a sample.

TPM (transcripts per million) divides the raw counts by transcript length and the library size and and scales this ratio by one million.

$$\mathrm{TPM} = 10^6 \frac{R_{ij}}{L_i} (\frac{1}{\sum_k \frac{R_k j}{L_k}})$$

CPM (counts per million) is where the scaling factor is the number of fragments sequenced times one million.

$$\mathrm{CPM} = 10^6 \frac{R_{ij}}{X_j}$$

FPKM (fragments per kilobase million) is the expected number of fragments per kilobase of transcript per million reads.

$$\mathrm{FPKM} = 10^9 \frac{R_{ij}}{X_j L_i}$$

TMM, DESeq2 and Quantile normalization are some between sample normalization methods:

TMM (trimmed mean of m-values) calculated scaling factors for each sample taking one samples as a reference sample. This scaling factor is computed as the weighted mean of log ratios for each sample with respect to the reference sample, after removing genes with largest log ratios and the most expressed genes

DESeq2 calculates DESeq2 scaling factor for a given sample as the median of the ratio, for each genes, of its read count over the geometric mean across all samples

Quantile normalization makes the distribution of normalized data to be the same for all samples by replacing each quantile with the mean or median of that quantile across all samples.

Machine learning metrics:

Certain metrics are used to assess performance of a machine learning model. These helps assess whether the predictions are correct and how much of the right classes are being correctly predicted by the machine learning model.

Precision: Precision is the proportion of relevant instances among all predicted instances. It is calculated by taking the true positives (relevant and predicted) divided by the sum of true positives and false positives.

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Recall: Recall is the proportion of correct instances which have been predicted. It is calculated by taking the true positives divided by the sum of true positives and false negatives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

F1-score: It is the harmonic mean of the recall and precision. It takes into account both false positives and false negatives and is a weighted average of recall and precision.

$$\text{F1-score} = \frac{2(\text{Recall*Precision})}{\text{Recall} + \text{Precision}}$$

Mathews Correlation Coefficient: The MCC is in essence a correlation coefficient between the observed and predicted binary classifications; it returns a value between −1 and +1. A coefficient of +1 represents a perfect prediction, 0 no better than random prediction and −1 indicates total disagreement between prediction and observation.

$$\text{MCC} = \frac{\text{TP*TN-FP*FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

AUC-PR: Average of precision scores calculated for each recall threshold. The precision recall area under curve (PR AUC) is just the area under the PR curve. The higher it is, the better the model is.

ROC: AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represents degree or measure of separability. It tells how much model is capable of distinguishing between classes.