Identifying the Structural and Evolutionary Constraints of Post-Translational Modifications

Identifying the Structural and Evolutionary Constraints of Post-Translational Modifications

Sean Landry

A Thesis in the Field of Bioinformatics

for the Degree of Master of Liberal Arts in Extension Studies

Harvard University

November 2019

Abstract

The objective of this work is to investigate structural and evolutionary constraints on post-translational modification (PTM) sites and their flanking regions beyond the phosphoproteome. An analysis of 21 PTM types including phosphorylation, acetylation, ubiquinitation, and methylation was conducted based on 400,000 high quality, *in vivo* PTM sites from PhosphoSitePlus. State of the art computational methods and cloud-based computing aided in the large-scale secondary structure prediction of the entire human proteome and generation of millions of global pairwise alignments between orthologs in 143 eukaryote species ranging from chimpanzee to yeast. We found that the majority of PTMs exhibit a significantly higher level of accessibility and prefer the localization in unstructured regions of a protein. Furthermore, we found that PTMs are more conserved at both the protein and residue level when compared with their unmodified counterparts. The conservation pattern of flanking residues, critical for transferase recognition, agrees with the evolutionary profile observed for PTM sites. Our study provides previously unknown insight into structural and evolutionary constraints of PTMs, laying the foundation for future work such as mutation effect prediction models.

Table of Contents

List of Tables

List of Figures

Chapter I.

Introduction

The human proteome is three times more complex than the expected coding capacity of the genome (Walsh, Garneau-Tsodikova, & Gatto, 2005). Post-translational modifications (PTMs) are major contributors to the complexity of the proteome (Khoury, Baliban, & Floudas, 2011). A PTM is the covalent modification of a protein after translation as the result of an enzymatic process catalyzed by transferase proteins. Phosphorylation, acetylation, ubiquitination, and methylation are the most widely studied PTMs amongst hundreds of PTM classes. PTMs provide an important mechanism for signal transduction and protein-protein interactions, allowing communication within the cell to control fundamental processes such as transcription, metabolism and cell division. Altered PTM-mediated signaling is commonly linked with various diseases including cancer. Mutations of a PTM site (or its flanking region) or alterations of associated regulating transferases can trigger perturbed signaling, leading to uncontrolled cell proliferation and growth (Reimand & Bader, 2013). For example, a single amino acid change from valine to glutamine in the BRAF kinase mimics the normal process of phosphorylation, required to activate the protein (Davies, Bignell, Cox, & Stephens, 2002). As a result, the elevated kinase activity of BRAF affects downstream signaling cascades promoting cellular proliferation.

To better understand the biology of PTMs, previous studies investigated structural and evolutionary constraints of PTM sites, but were limited to the phosphoproteome,

small sample numbers, and low number of examined species for conservation analysis. Using a combination of prediction tools, phosphorylation sites have been shown to exhibit a greater degree of accessibility and preference for highly disordered regions of a protein (Gnad et al., 2007). Evolutionary analyses have demonstrated that phosphoproteins in eukaryotes have a significantly greater proportion of orthologs than non-phosphoproteins (Gnad et al., 2010). Furthermore, at the site level, phosphorylated serines and threonines in human, mouse, and fly are more conserved than non-phosphorylated serines and threonines in other eukaryotes.

The focus of this work was to investigate structural and evolutionary constraints on PTM sites and their flanking regions beyond the phosphoproteome. The large-scale analysis of hundreds of non-phospho PTM sites is now feasible, as recent advances in mass spectrometry-based proteomics led to an overwhelming increase in the number of high quality *in vivo* modification sites (Hornbeck et al., 2019). The most comprehensive resource of PTM sites, PhosphoSitePlus (www.phosphosite.org), contains over 400,000 curated, human *in vivo* modification sites representing 21 different PTMs including phosphorylated serines (S-p), ubiquitinated lysine (K-ub), phosphorylated tyrosine (Y-p), acetylated lysine (K-ac), mono-methylated arginne (R-m1), sumoylated lysine (K-sm), mono-methylated lysine (K-m1), succinylated lysine (K-sc), di-methylated arginine (R-m2), O-GlcNAc-T (threonine), O-GlcNAc-S (serine), O-GalNAc-S (serine), O-GlcNAc-T (threonine), di-methylated lysine (K-m2), caspase cleavage of aspartic acid (D-ca), tri-methylated lysine (K-m3), methylated arginine (R-m), methylated lysine (K-m), neddylated lysine (K-ne), and palmitoylated cysteine (C-pa). In addition to the identification of PTM sites, PhosphoSitePlus provides further information about flanking

regions, evolutionary conservation, upstream and downstream signaling interactions, and impact on biological functions (Hornbeck et al., 2019).

In this study we included all human *in vivo* modification sites from PhosphoSitePlus, and investigated their structural constraints using high accuracy secondary structure and accessibility prediction. Furthermore, we examined their conservation across 143 eukaryotes ranging from chimpanzee to yeast. We found that all PTM sites are enriched on the surface of protein to ensure accessibility to their corresponding transferases. Associated secondary structures, however, vary between PTM types. Most PTM sites including phosphorylation sites are enriched in unstructured loop regions. There are, however, exceptions. For example, succinylation sites are enriched in helices. Our evolutionary analysis made clear that the vast majority of proteins with one or more PTM sites have more orthologs in other species than unmodified proteins. Similarly, at the site level most PTM sites tend to be higher conserved across species when compared to their unmodified counterpart residues. On the flip side, our study makes clear that not all PTM sites are highly conserved, indicating non-functional “background” PTM sites in the proteome. The pattern of conservation extends to the modification site flanking regions with preference for residues important for transferase substrate recognition.

Taken together, our structural and evolutionary findings provide previously unknown insight into the structural and evolutionary constrains imposed on PTMs beyond the phosphoproteome. These constraints are tightly linked with the interpretation of protein functions. In the future, these new insights can be implemented into improved models of mutation effect prediction.

Chapter II.

Materials and Methods

The data sources, computational tools, software development, and analyses are defined in this section. A distributed computing infrastructure and analysis pipeline were developed to overcome the challenges of predicting the secondary structure of more than 20,000 human proteins. In addition, the same infrastructure enabled the completion of over 2 million global pairwise alignments between orthologs. Statistical analyses of the resulting data sets were used to define the structural and evolutionary constraints imposed on PTMs.

Design of the data analysis pipeline

PTM sites were retrieved from PhosphoSitePlus (www.phosphosite.org) (Hornbeck et al., 2019) and merged with ortholog information from Ensembl (www.ensembl.org) (Zerbino et al., 2018), representing 143 eukaryotic species ranging from chimpanzee to yeast. Using the latest innovations in distributed cloud computing (Production-Grade Container Scheduling and Management, 2014/2019), global pairwise alignments were computed for over 2 million orthologous pairs based on the Needleman-Wunsch algorithm (Needleman & Wunsch, 1970), to profile the evolutionary conservation of the modification sites. Furthermore, the secondary structure was

predicted for over 18,000 human proteins using SPIDER3 (Heffernan, Yang, Paliwal, & Zhou, 2017). Downstream analyses were performed using R and Python (Figure 1).

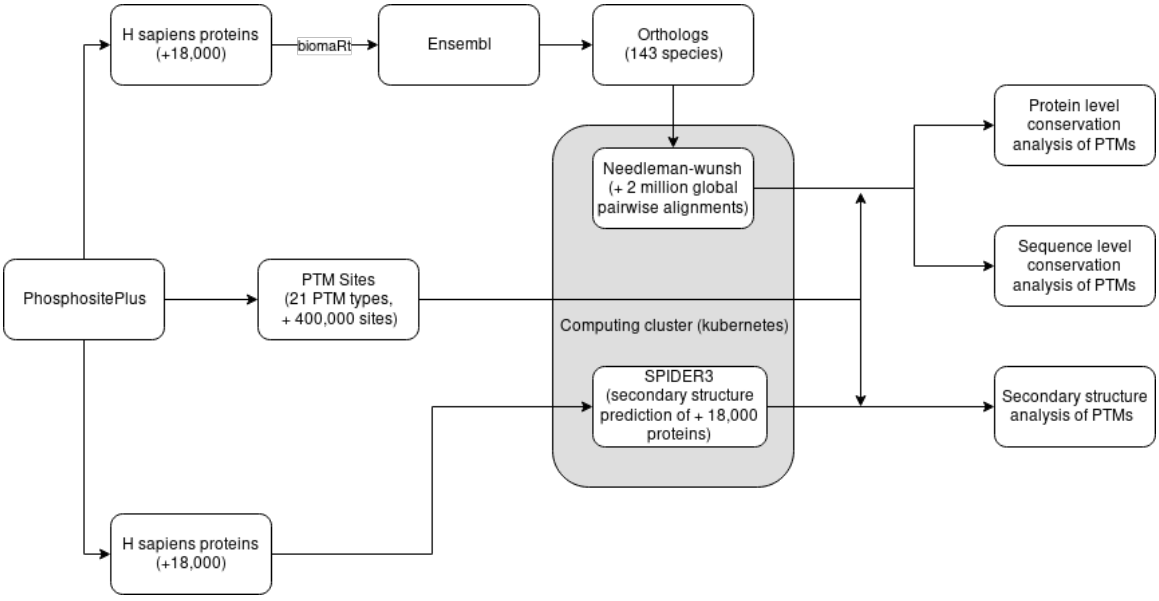


Figure 1. Analysis Workflow

Retrieving the PTM site dataset

All PTM sites included in this study were sourced from PhosphoSitePlus (Hornbeck et al., 2019). The data sets are free to download and available in the form of compressed csv files from the PhosphoSitePlus downloads page at https://www.phosphosite.org/staticDownloads. Table 1 is a summary of the different modification types and their counts included in this study.

Table 1. PTM Types and Frequencies.

Modification Type	Residue	Abbreviation	Count
Phosphorylation	S	S-p	165620
Ubiquitination	K	K-ub	73317
Phosphorylation	T	T-p	69792
Phosphorylation	Y	Y-p	46046
Acetylation	K	K-ac	35247
Mono-Methylation	R	R-m1	10208
Sumoylation	K	K-sm	8443
Mono-Methylation	K	K-m1	4428
Succinylation	K	K-sc	4131
Di-Methylation	R	R-m2	2121
O-GalNAc	T	O-GalNAc-T	1275
O-GlcNAc	S	O-GlcNAc-S	856
O-GalNAc	S	O-GalNAc-S	840
O-GlcNAc	T	O-GlcNAc-T	661
Di-Methylation	K	K-m2	590
Caspase Cleavage	D	D-ca	474
Tri-Methylation	K	K-m3	338
Methylation	R	R-m	118
Methylation	K	K-m	110
Neddylation	K	K-ne	43
Palmitoylation	C	C-pa	25

The post-translation modification types and their frequencies included in this study.

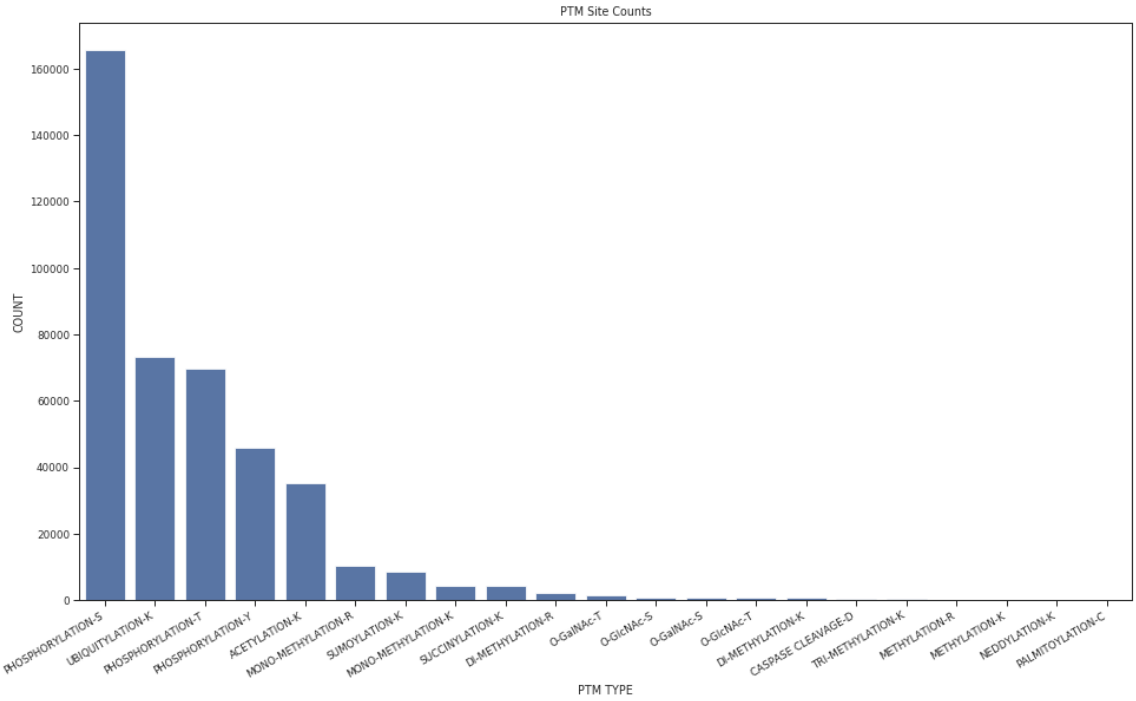


Figure 2. Observed PTM Frequencies

Retrieving Protein Sequences

All human protein sequences included in this study were sourced from PhosphoSitePlus. The sequences can be downloaded in the form of compressed csv files from the PhosphoSitePlus downloads page at https://www.phosphosite.org/staticDownloads (Hornbeck et al., 2019). PhosphoSitePlus includes 20,105 human proteins indexed by Uniprot identifiers (Bateman et al., 2017). Isoforms were omitted for this study.

Predicting Secondary Structure

Predicting the secondary structure for all human proteins was critical to understand the structural constraints and features across orthologs. The tool of choice for predicting the secondary structure was SPIDER3 (Heffernan et al., 2017; Lyons et al., 2014). SPIDER3 is the latest release of a series of tools that specializes in predicting protein secondary structure, backbone angles, solvent accessibility, local and non-local interactions directly from a sequence.

SPIDER3 implements bidirectional recurrent neural networks (Schuster & Paliwal, 1997), or BRNNs, and long short-term memory cells (Hochreiter & Schmidhuber, 1997), or LSTM, to improve the capture rate for non-local interactions when predicting secondary structure. The method consists of a pipeline, in which the PSSM, or position specific substitution matrices, output by PSI-BLAST (Altschul et al., 1997) and the HMM, or hidden markov model, profiles output from HHBlits (Remmert, Biegert, Hauser, & Söding, 2012) are used as input for a series of classifiers constructed using the LSTM-BRNN model (Heffernan et al., 2017). The described approach allows the entire sequence to be used as input.

The PSI-BLAST and HHBlits programs that SPIDER3 rely on require the uniref90 and uniprot20 databases. These databases are publicly available from Uniprot and were downloaded for use locally (Bateman et al., 2017). A few test runs of the SPIDER3 program revealed that it can take a significant amount of time if processing many thousands of sequences. For example, one sequence can take more than 10 minutes when using a machine with 8 cpu and 16gb of memory. The projects architecture was

designed for tackling this type of computational hurdle. The project was deployed on a Kubernetes (Production-Grade Container Scheduling and Management, 2014/2019) cluster hosted in an Amazon Virtual Private Cloud provided by Cell Signaling Technology. The only limits to the number of sequences that can be run in parallel through the SPIDER3 program are the limits imposed to control the number of nodes in the cluster. This approach shortened the time required for SPIDER3 computation from weeks to hours when processing thousands of sequences. SPIDER3 was containerized and deployed on the Kubernetes cluster using the projects framework. Using this distributed computing approach, the secondary structure prediction was conducted for over 20,000 human proteins in about eight days. About 30 SPIDER3 predictions were running in parallel at any given time distributed across the computing cluster.

The data output from SPIDER3 consists of a series of files including the PSI-BLAST summary, the HHBlits summary, the position specific scoring matrix, and a SPIDER3 specific file that includes the protein structure prediction. The SPIDER3 specific file includes structural state for each amino acid position. SPIDER3 uses the label “H” to classify a position’s secondary structure state as a helix shape, which could be a 3_{10}-helix , alpha-helix, or pi-helix. The label “E” is used to classify a position’s state as a beta-bridge or beta-strand. The label “C” is used to classify as position’s state as a high curvature loop, a beta-turn, or a coil (Heffernan et al., 2017). Furthermore, a series of bond angles are included that can be used to describe the proteins backbone structure. Additional features include accessible surface area and half sphere exposure. The half sphere exposure adds a spatial context to the contact number. If a sphere were drawn around an amino acid position, the half sphere exposure captures the number of

interactions at the top of the sphere versus the bottom of the sphere (Heffernan et al., 2017).

Batches of sequences were submitted as jobs to the cluster using the dakub http api via a post to http://example/dakub/api/task/async-apply/spider.spider3.spider3_workflow_v_0_0_1/. The spider3_workflow_v_0_0_1 accepts a list of protein sequences for secondary structure prediction. The workflow submits each sequence as a separate job to SPIDER3 in order to maximize parallel processing. All output files by SPIDER3 were uploaded to Amazon S3 for retrieval. Over 20,000 files output by SPIDER3 were parsed and concatenated into one large pandas data frame containing over 11 million rows (McKinney, 2013). Each row of the data frame represented a single amino acid position within a protein and contains all of its associated secondary structure information. The data frame was saved as a python ".pickle" file for further analysis.

Merging Secondary Structure with Post-Translational Modifications

All PTM sites sourced from PhosphoSitePlus were merged with the secondary structure data frame using a SQL style inner join on Uniprot id and residue position. Quality filtering was performed to omit any PTMs in which the expected modified site residue did not match the sequence residue at the specified position which resulted in an attrition of 1,573 PTMs. An additional column was added to the merged data frame titled "IS_MOD" to denote whether an amino acid is modified or unmodified. The resulting data frame was saved as a python ".pickle" file.

Secondary Structure Analysis

A survey was conducted for each PTM type and secondary structure classification as defined by SPIDER3. For a given PTM type all modified residues were analyzed in conjunction with all unmodified residues to identify significant enrichment. For instance, all phosphorylated serines were compared to unmodified serines. A contingency table was built for each of the three secondary structure classifications, which includes groups H, E, or C. The label "H" is used to classify a position's secondary structure state as a helix shape, which could be a 3_{10}-helix , alpha-helix, or pi-helix. The label "E" is used to classify a position's state as a beta-bridge or beta-strand. The label "C" is used to classify as position's state as a high curvature loop, a beta-turn, or a coil (Heffernan et al., 2017). Table 2 is an example of the resulting contingency table for phosphorylated serines and the structure classification of group C, which is a high curvature loop, beta-turn or a coil.

Table 2. Phosphorylated Serines Secondary Structure Group C Contingency Table

		Modified Residue	
		False	True
Classified as high curvature loop, a beta-turn, or a coil	False	266,099	31,540
	True	624,029	134,080

In comparison, 80.95% of phosphorylated serines and only 70.11% of unmodified serines are classified as either a high curvature loop, beta-turn, or coil. A Pearson Chi-squared test was applied to test the null hypothesis that the enrichment for modified and

unmodified residues is the same within unstructured regions (Pearson, 1900). The same exercise was applied to all of the secondary structure classification groups.

Furthermore, the accessible surface area was analyzed by comparing the reported distributions for modified versus unmodified residues. A two-sided Welch's t-test and a Bonferroni (Dunn, 1959) corrected significance level of $\alpha = 0.05/21 = 0.002$ was performed to test the null hypothesis that modified and unmodified residues have an identical mean accessible surface area (Welch, 1947). The test was performed for all PTM types.

Retrieving orthologs

The biomaRt library available through the Bioconductor package collection was used for retrieving orthology data for each human protein included in this study (Kinsella et al., 2011). Ensembl was queried using a protein's Uniprot id to retrieve orthologs across 143 distinct species. Important information such as orthology type, gene order conservation score, whole genome alignment score, and the protein sequence for each ortholog were retrieved (Zerbino et al., 2018). The quality of orthology is quantified using the gene order conservation score and whole genome alignment score. The gene order conservation score is defined as how well the four nearest neighbors of a gene match. The whole genome alignment score captures the alignment coverage over the orthologs (Zerbino et al., 2018).

The Ensembl Compara project defines orthologs based on protein trees built using a prediction pipeline. The pipeline consists of eight steps including translation to protein sequences, running NBCI Blast+ of one gene against every other gene (Camacho et al.,

2009), clustering of the protein sequences using the hcluster_sg program (H. Li, 2011/2017), performing multiple sequence alignments of the clusters using a combination of M-Coffee3 (Wallace, O'Sullivan, Higgins, & Notredame, 2006) and Mafft4 (Katoh et al., 2002), constructing phylogentic trees using TreeBeST5 (H. Li, 2011/2017), and finally determining orthology types using the constructed trees. An ortholog can be describe as homologous sequences separated by a speciation event. Ensembl further classifies orthologs as either an "ortholog_one2one", "ortholog_one2many", or "ortholog_many2many". A one to one ortholog is defined as a homology event occurring once when comparing a sequence from one species to another species. A one to many ortholog is defined as a homology event occurring more than once when comparing a sequence from one species to another. In other words, a single sequence of one species shows similarity to multiple sequences of another species. Finally, a many to many ortholog is observed when a two or more sequences from one species demonstrate homology to the same sequence or many sequences from another species (Zerbino et al., 2018).

Queries to the Ensembl database are limited to one species at a time. However, a large collection of Uniprot ids can be included with each request. The R script named "retrieve_all_homologs.R" was written to complete the retrieval of all homologs in a parallel fashion. An .rds file was generated for each available species in the Ensembl database, that includes orthologs to all human proteins obtained from PhosphoSitePlus. The .rds files contain Uniprot ids, protein sequences, and all ortholog types with their corresponding protein sequences. Ensembl release 95 was used for this study enabling the retrieval of homologs across 143 distinct species. An attrition of 5% in total number

of proteins was experienced during the collection of orthologs. Of the 20,105 Uniprot identifiers submitted to Ensembl, 1,026 did not return any ortholog match. Another seven proteins were omitted as a result of mapping to a none unique Ensembl peptide identifier.

Generating Global Pairwise Alignments

Global pairwise alignments were generated for each orthologous sequence pair using the Needleman-Wunsch algorithm (Needleman & Wunsch, 1970). Alignments were completed for all orthology types including one to one, one to many, and many to many. The Biopython pairwise2 implementation of the Needleman-Wunsch algorithm was used to complete the global pairwise alignments for over 2 million homologous pairs (Cock et al., 2009). In order to overcome the slow and computational expensive, dynamic programming approach of the Needleman-Wunsch algorithm, a workflow was created and packaged within the project framework to distribute the alignment work load across the computing cluster. Batches of homologous pairs were submitted to the cluster as jobs using the dakub http api via a post to http://example/dakub/api/task/apply/movecore.tasks.batch_needleman_wunsch_v1/. Any homologous pair containing a sequence greater than 30,000 amino acids in length was omitted due to hardware memory constraints. Only Titin was omitted from the data set due to sequence length, which resulted in an attrition of 123 homologous pairs. The batch_needleman_wunsch_v1 workflow accepts a data frame of homologous pairs as input with algorithm parameters specified in additional columns. The alignments were completed using a gap open penalty of 10 and a gap extension penalty of 0.5. Upon completion of all alignments within a batch, the results are uploaded to Amazon S3

storage for retrieval. After 2,429,949 global pairwise alignments were complete, the results for all batches were downloaded and reassembled into one data frame. The species files generated from ortholog retrieval were traversed and the alignments were mapped to their corresponding homologous pair.

Analysis of PTM Conservation

For the orthology analysis only one to one orthologs were considered. The species files containing orthologs and global pairwise alignments were traversed and one to one orthologs were extracted. The resulting information was concatenated into one data frame. All PTMs were merged with the one to one orthology data frame using a SQL style inner join on Uniprot identifiers. Using this approach all PTMs for each protein were analyzed for conservation in each of its orthologs across all 143 species at the protein and sequence level. In addition to modification site, seven flanking amino acids were included to provided full PTM site profiles.

Evolutionary Analysis at the protein level

The number of modified and unmodified human proteins with and without a one to one ortholog were counted for each of the 21 different PTM types across all 143 species. A contingency table was built for each PTM type within each species. A Pearson Chi-squared test and a Bonferroni (Dunn, 1959) corrected significance level of $\alpha = 0.05/21 = 0.002$ was applied to each contingency table to test the null hypothesis that modified proteins and unmodified proteins have the same number of observed one to one orthologs (Pearson, 1900). For example, Table 3 illustrates the contingency table generated for the conservation phosphorylated and non-phosphorylated serines in gorilla.

For example, at an $\alpha = 0.002$ the resulting significance ($p = 3.917506e^{-30}$) indicates strong evidence to reject the null hypothesis. The observed higher proportion of modified proteins with a one to one ortholog at 85.22% versus 75.27% for unmodified proteins is not due to chance.

Table 3. *G. gorilla* Contingency Table

		HOMOLOG	
		False	True
IS_MOD	False	487	1,482
	True	2,470	14,237

The number of proteins with one or more phosphorylated serine versus a protein without a phosphorylated serine.

Evolutionary Analysis at the residue level

For the analysis of flanking residues, modification sites within seven amino acid positions of the start or end of a sequence were omitted. As a result, an attrition of 13,554 PTMs was observed, which represents about 3% of all PTMs included in this study. Conservation matrices were generated for each PTM observed within an orthologous pair using their global pairwise alignments. A conserved residue was assigned a 1 and a non-conserved residue was assigned a 0, resulting in a 15 kmer of 1s and 0s for each PTM. Furthermore, the same exercise was completed for all corresponding unmodified residues. Figure 3 is an illustration of a subset of rows in the resulting conservation matrix built for each PTM type.

Figure 3. Theoretical Conservation Matrix

7 aa upstream							PTM	7 aa downstream						
1	1	1	1	0	1	0	1	0	1	1	1	1	1	1
0	1	0	0	1	0	1	1	1	0	0	1	1	1	1
1	1	1	1	0	0	1	1	0	1	1	1	0	1	0

Example conservation matrix, 1 = conserved, 0 = not conserved.

The matrices were split into modified and unmodified residue data sets. These data sets were then grouped by species and the number of conserved residues for each position in the 15 kmer was obtained using a straight forward summation of each individual column. The proportion of conserved residues was calculated using the column summation divided by the total number of rows.

A contingency table was built for each kmer position for each PTM type within each species. A Pearson Chi-squared test was applied to test the null hypothesis that there is no difference in level of conservation for modified residues and unmodified residues (Pearson, 1900). The same test was applied to all positions within seven amino acids upstream and downstream of a modification site.

Software Architecture & Engineering

The project was built using the python web framework known as Django (*Django*, 2018). Django is fast, scales well, and provides an infrastructure built on modular, portable applications. Every Django project consists of a collection of applications. A Django application can be taken from one project and plugged into

another with minimal effort. This project followed the Django design principal by organizing related components into independent and reusable applications.

Engineering a pipeline from the collection of tools was accomplished using the python library Celery. Celery is straightforward to integrate into Django projects. It provides a method of asynchronous task and job queue using message passing (*Celery*, 2009/2018). Furthermore, Celery includes a useful feature called canvas that provides a way to define work flows that chain tasks together in a serial and/or parallel fashion. This project included work flows defined using Celery canvas that are composed of tasks defined in one or more of its applications.

Celery does not enforce any restrictions on what content is included in messages that are passed between work flow tasks. In addition, Celery does not provide a way to easily track work flows defined using the canvas feature. The results of individual tasks are captured in a traditional database or cache back-end. However, tracking the progress of a canvas work flow and working with a collection of related tasks efficiently is difficult with currently available tools. A custom Django application known as Dakub (DatA worklfows with Celery, Django, Docker, and KUBernetes) was implemented to tackle these issues.

Dakub is a Django application that can be used in any Django project that incorporates Celery. It provides python classes such as the DakubMessageV1 class to help standardize the messaging protocol between work flow tasks. A task that implements the DakubMessageV1 protocol should be able to decipher the message sent by another task that implements DakubMessageV1 protocol. Dakub also includes custom python decorators (Smith, Jewett, Montanaro, & Baxter, 2003) that can be used to

decorate any task to instantly add functionality that will track its result as part of a work flow. Dakub includes a RESTful API for the execution of any task or work flow registered with a Celery application. A basic web user interface is implemented for browsing available work flows and results. An important feature of Dakub is a built-in scheduler for interaction with the Kubernetes API when an application is deployed on a Kubernetes cluster.

Kubernetes is a system for the orchestration, deployment, scaling, and management of containerized applications (Production-Grade Container Scheduling and Management, 2014/2019). The dakub scheduler takes advantage of the existing Kubernetes job object to deploy Celery workers. The scheduler monitors the Celery task queues and schedules Kubernetes jobs (Celery workers) accordingly. Each queue can have a unique job configuration, or the same job configuration can be shared among multiple queues. One can define CPU and memory claims in a job configuration to utilize available computing resources efficiently. For example, a task for predicting protein secondary structure may be more memory intensive compared to a task that dumps data to a file on disk. Furthermore, the dakub scheduler will shut down the Celery workers when there are no more messages to be consumed in its corresponding queue. Kubernetes can then reclaim the computing resources for other functions or scale the number of nodes in the cluster accordingly.

The project was fully containerized using Docker. Docker is an open source project that includes multiple tools for the delivery and creation of containerized applications (Docker CE., 2017/2019). Pursuing the containerized approach allows for deployment locally on a single machine and/or on a Kubernetes cluster hosted on a cloud

provider such as Amazon Web Services (AWS) or the Google Cloud Platform (GCP). Containerization has the added bonus of making the application easier to use and avoid any complicated installations. The project is flexible, modular, and very easy to scale.

Chapter III.

Results

Most but not all PTM classes are enriched in unstructured regions

To elucidate the structural constraints of PTMs, we compared modified and unmodified residues within each PTM type across all three secondary structure classification groups based on secondary structure and accessibility predictions (Materials). Figure 4 profiles the proportions of all PTM types and the corresponding unmodified residues located in a non-regular secondary structure including a high curvature loop, a beta-turn, or a coil as classified by SPIDER3. A Pearson Chi-squared test and a Bonferroni (Dunn, 1959) corrected significance level of $\alpha = 0.05/21 = 0.002$ was used to test the null hypothesis that modified and unmodified residues are equally enriched in unstructured regions. O-GlcNAc-S, O-GalNAc-T, O-GlcNAc-S, D-ca, O-GlcNAc-T, S-p, R-m2, R-me, T-p, R-m1, K-sm, K-me, and K-m1 demonstrated significant enrichment in unregular structures. In contrast, in the case of K-ac, Y-p, K-ub, K-sc, the corresponding unmodified residues demonstrated significant enrichment. Furthermore, C-pa, K-m3, K-m2, and K-ne and their corresponding unmodified residues did not show any significant enrichment in loops or turns. Overall, for the majority of PTMs type sets, more than half of their modification sites are located in unstructured regions including O-GlcNAc-S, O-GalNAc-T, O-GlcNAc-S, D-ca, O-GlcNAc-T, S-p, R-m2, R-me, T-p, R-m1, K-sm, K-me, K-m1, C-pa, K-m3, and K-m2. In comparison,

only half of the K-ac, Y-p, K-ub, K-sc, and K-ne sites are located within an unstructured region.

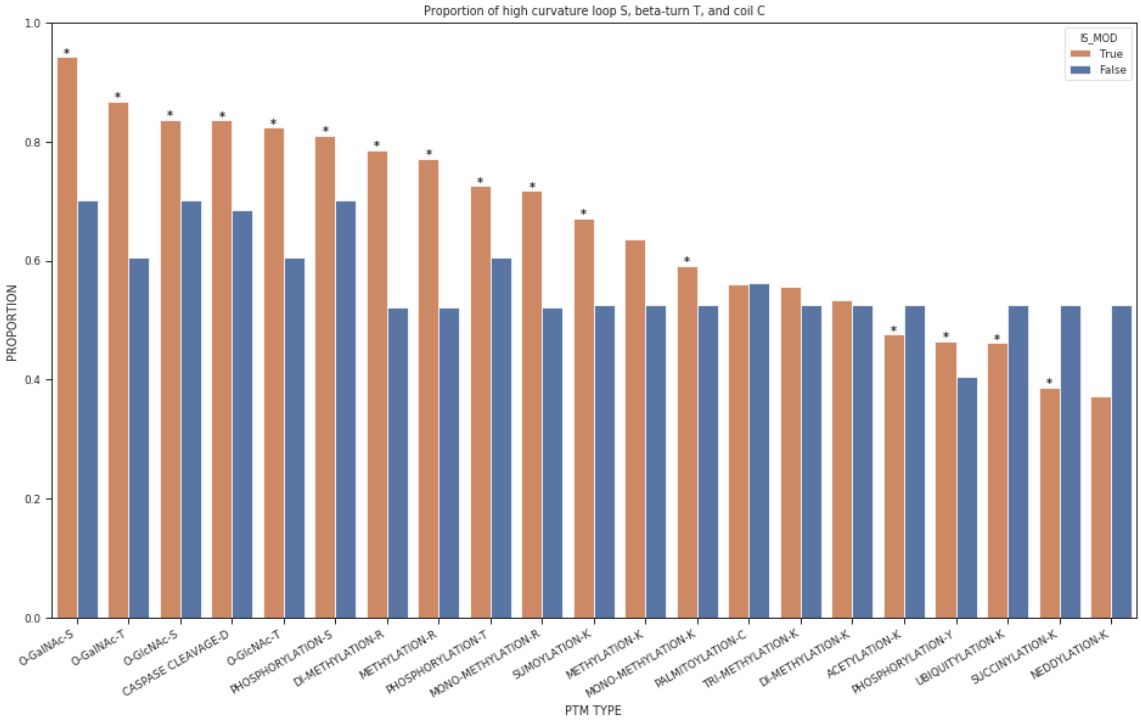


Figure 4. Proportion of high curvature loop S, beta-turn T, and coil C.

Proportion of modified residues, shown in orange, and unmodified residues, shown in blue, enriched in a high curvature loop, a beta-turn, or a coil as classified by SPIDER3

In addition to unstructured regions, we investigated localization patterns in helical regions. Figure 5 illustrates the proportions of all PTM types and the corresponding unmodified residues located in a 3_{10}-helix , alpha-helix, or pi-helix as classified by SPIDER3. We found that K-sc, K-ub, and K-ac are significantly enriched within helical

regions. Consistent with their enrichment in unstructured regions, T-p, K-m1, K-me, K-sm, R-m1, R-me, Y-p, R-m2, D-ca, S-p, O-GlcNAc-S, O-GlcNAc-T, O-GalNAc-T, and O-GalNAc S, were underrepresented in helices compared to their unmodified counterpart residues. C-pa and K-ne and their corresponding unmodified C and K residues showed no significant enrichment within helical regions. In comparison with unstructured regions, the proportion of modification sites located within a helical region falls below 40% for the majority of PTM types.

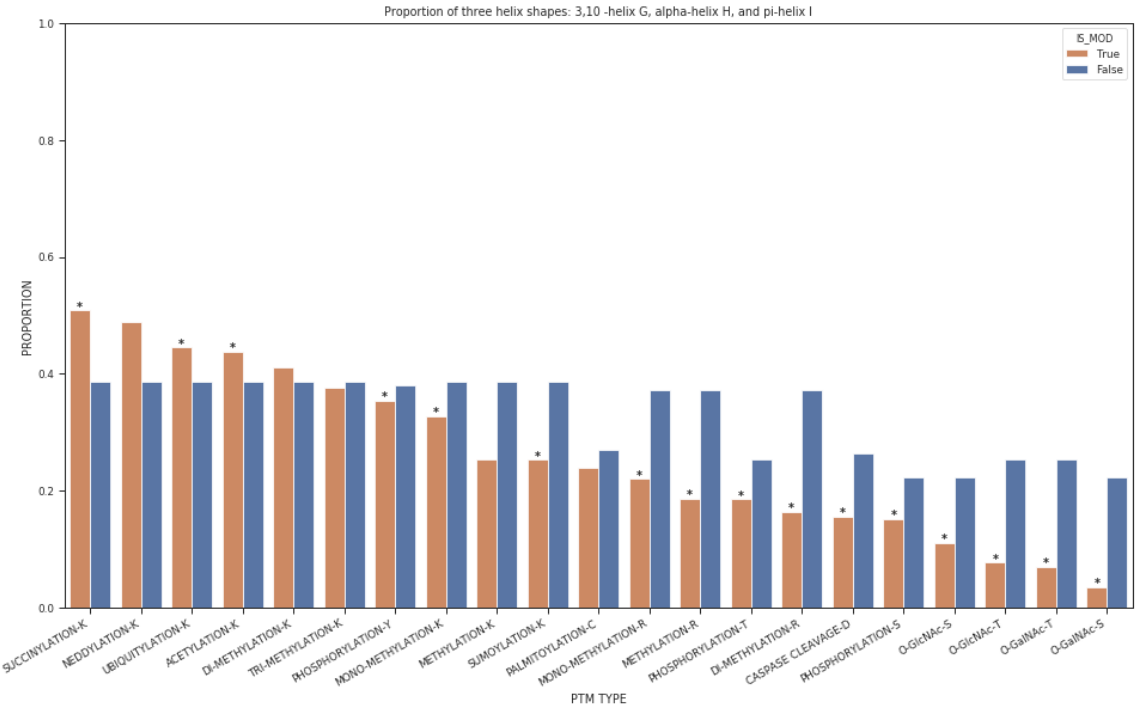


Figure 5. Proportion of three helix shapes: 3,10-helix G, alpha-helix H, and pi-helix I

Proportion of modified residues, shown in orange, and unmodified residues, shown in blue, enriched in a 3_{10}-helix , alpha-helix, or pi-helix as classified by SPIDER3.

Finally, beta-bridges and strands were investigated. Figure 6 shows the proportions of all PTM types and the corresponding unmodified residues located in a beta-bridge or beta-strand as classified by SPIDER3. Of the 21 different PTM types included in this study only K-sc demonstrated significant enrichment within a beta-bridge or beta-strand. Modification sites located within beta-bridges or beta-strands are the least frequent representing only 10% of most PTM types.

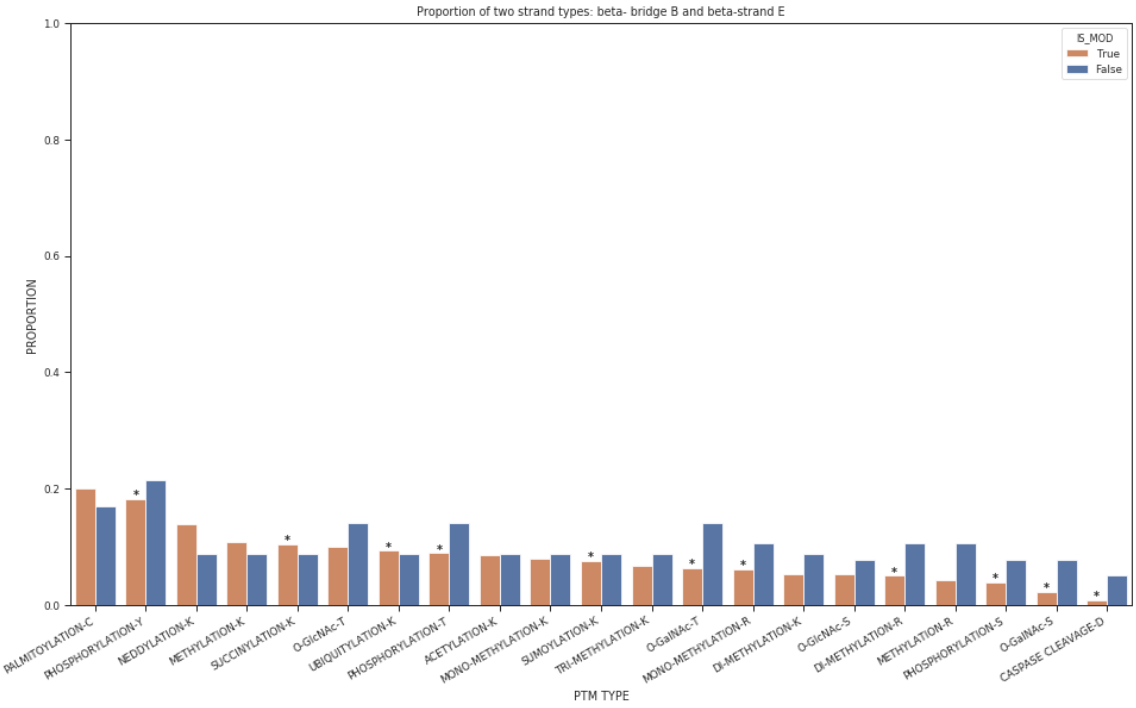


Figure 6. Proportion of two strand types beta-bridge B and beta-strand E

Proportion of modified residues, shown in orange, and unmodified residues, shown in blue, enriched a beta-bridge or beta-strand as classified by SPIDER3.

PTM sites are localized on the protein surface

In addition to secondary structure, we examined the accessible surface area predicted by SPIDER3. Figure 7 illustrates the accessible surface area distributions for each of the 21 PTM types. A two-sided Welch's t-test and a Bonferroni corrected significance level of $\alpha = 0.05/21 = 0.002$ was used to test the null hypothesis that modified and unmodified residues have the same mean accessible surface area. Except for K-ne, K-sc, and C-pa, all PTM types demonstrated a significantly higher surface accessibility when compared with their corresponding unmodified residues. K-me exhibited the largest mean accessible surface area of 119.92, and C-pa showed the lowest observed mean accessible surface area of 30.73.

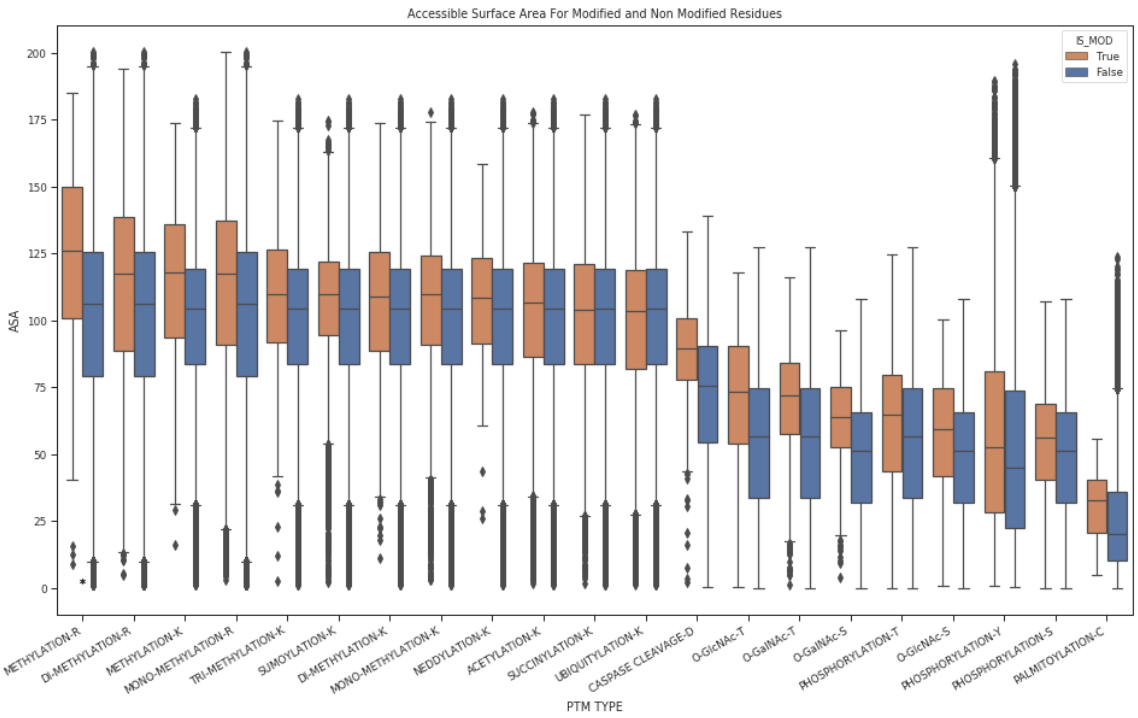


Figure 7. Accessible surface area for modified and unmodified residues

The accessible surface area distributions for modified residues, shown in orange, and unmodified residues, shown in blue.

Evolutionary Constraints of Post-Translationally Modified Proteins

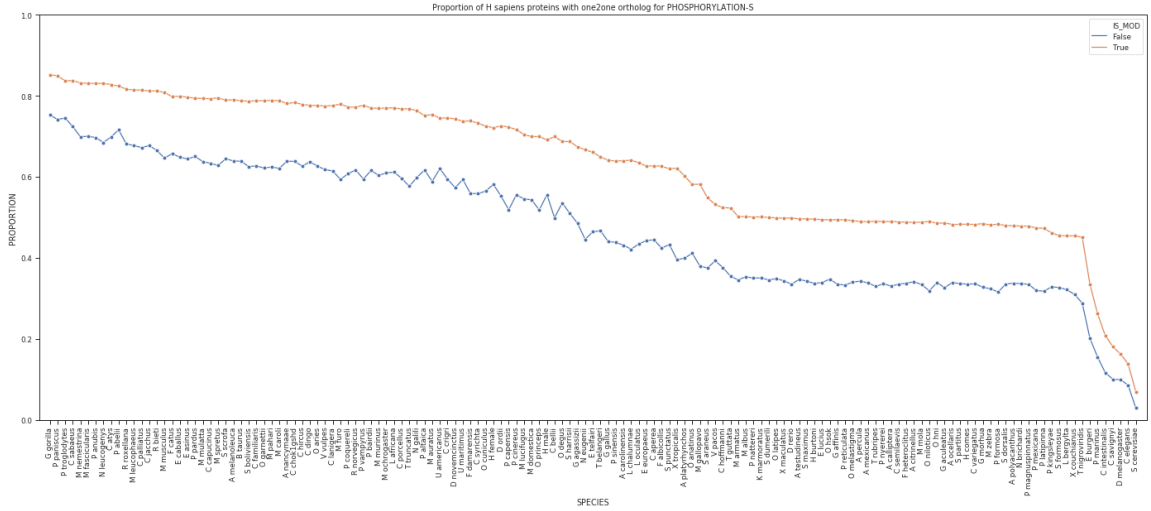
Using orthology data from Ensembl we tested the null hypothesis that modified and unmodified proteins have the same proportion of one to one orthologs in all 143 species included in this study. Significance was determined using a Pearson Chi-squared test and Bonferroni corrected significance level of $\alpha = 0.05/21 = 0.002$. Figure 8 profiles the proportions for the five most frequently observed post-translation modifications within the data set including S-p, K-ub, T-p, Y-p, and K-ac. Figures for the additional 16 PTM types can be found in Appendix 1.

Phosphorylation on serines is the most frequently observed modification within the data set. Proteins that have one or more S-p sites, are more likely to have an ortholog in another species than proteins that without any record of S-p sites (Figure 8). This pattern was observed across all 143 species. For instance, within gorilla (*G. gorilla*), 85.22% of human proteins with at least one or more S-p site have an one to one ortholog compared to 75.26% for human proteins without an S-p site. The species *S cerevisiae*, commonly known as brewer's yeast, includes the fewest one to one orthologs for human with proportions of 6.91% for S-p proteins and 2.89% for proteins without an S-p site. Human proteins with one or more K-ub, T-p, Y-p, and/or a K-ac site demonstrate a similar trend to S-p proteins with a proportion significantly greater than their corresponding unmodified groups across all 143 species (Figure 8).

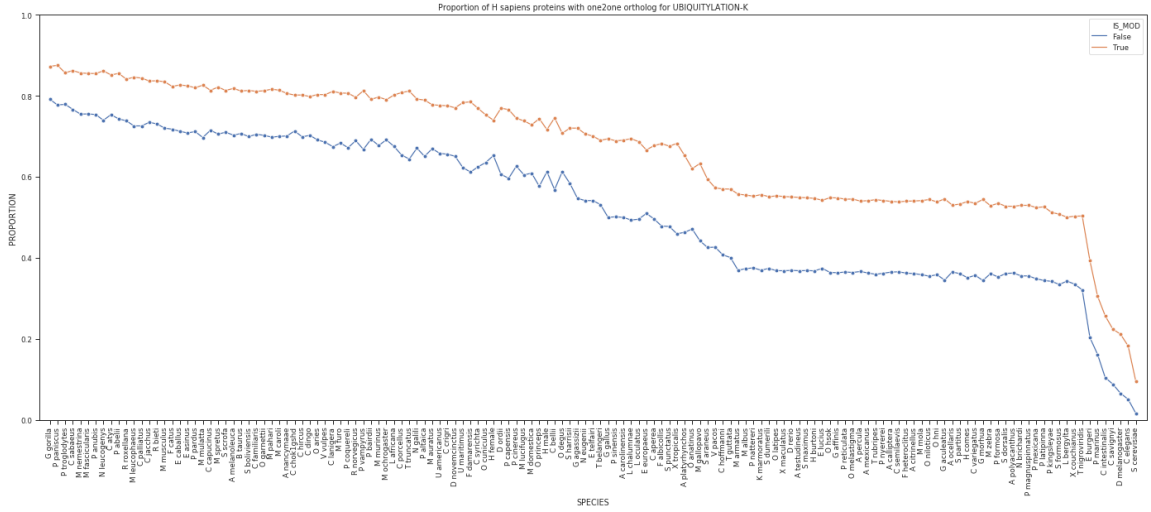
R-m1 sites are the 6th most frequent within the dataset falling in behind K-ac sites with 35,247 and 10,208 sites respectively. Similar to K-ub, T-p, Y-p, and K-ac modified proteins, R-m1 modified proteins show a significantly greater proportion of one to one orthologs in all species surveyed. K-m1, D-ca, and R-m2 exhibited significantly greater proportions of one to one orthologs within 94%, 85% and 80% of the species surveyed respectively. They show a greater proportion across most species with only sporadic dips of insignificance. O-GalNAc-T, O-GalNAc-S, O-GlcNAc-S, and O-GlcNAc-T modified proteins have a significantly greater proportion of one to one orthologs in roughly 60% of the species surveyed. The group consists almost entirely of species more closely related to humans. For K-sc significance is not observed for the top 50% of the most closely related species but for those with fewer one to one orthologs such as *X. tropicalis.* K-sm demonstrates significance in 11% of species surveyed from both close and distantly

related members. K-me, K-m3 and K-m2 exhibit significance in 3%, 3%, and 1% of species included with no pattern of relationship. Finally, R-me, K-ne, and C-pa exhibit no pattern of significance in any of the 143 species surveyed (Appendix 1 – Supplementary Figures).

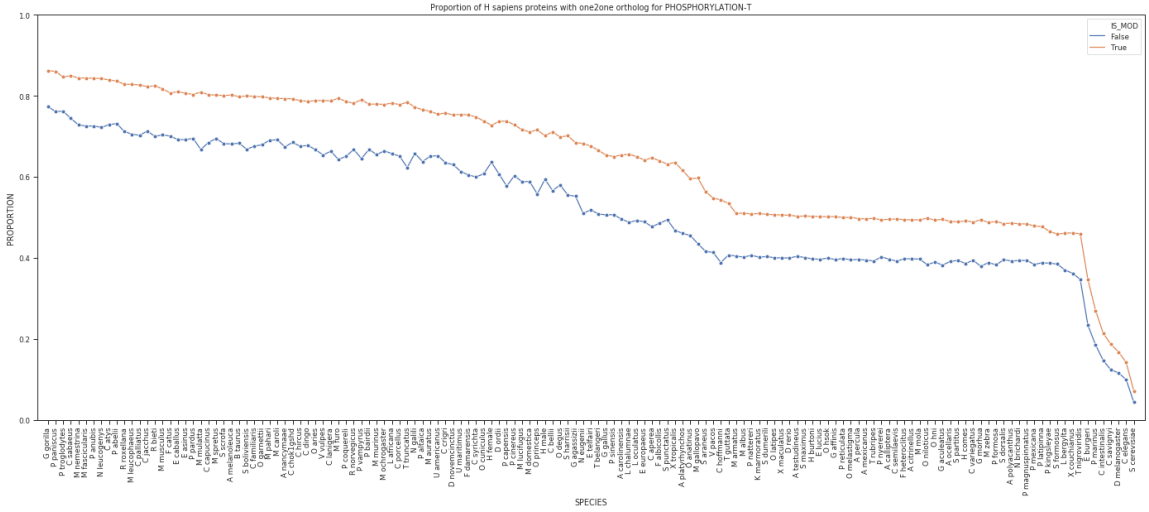
a)


Proportion of H sapiens proteins with one2one ortholog for PHOSPHORYLATION-S
PROPORTION
SPECIES
IS_MOD
False
True

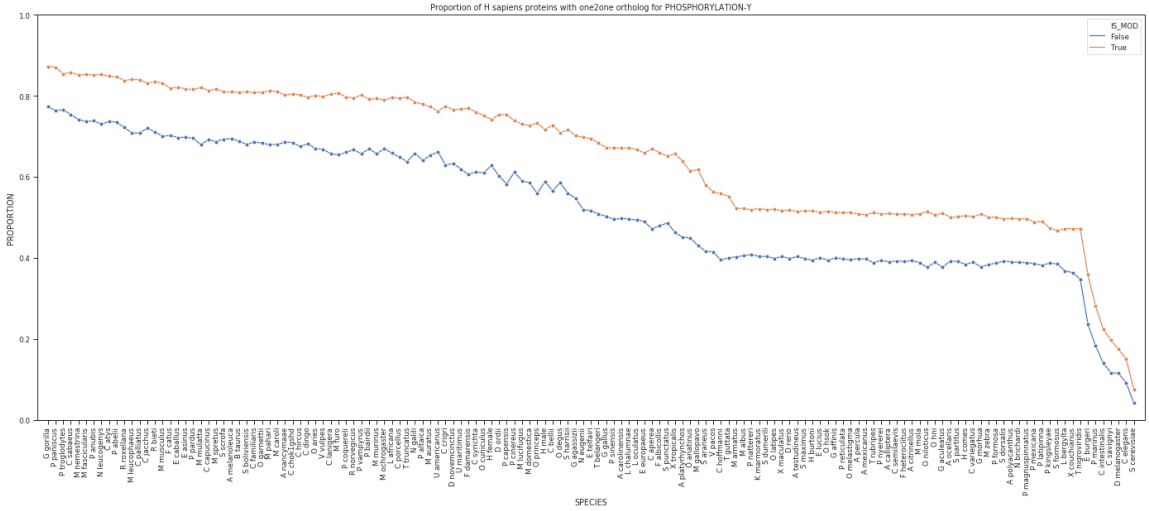
b)


Proportion of H sapiens proteins with one2one ortholog for UBIQUITYLATION-K
PROPORTION
SPECIES
IS_MOD
False
True

c)


Proportion of H sapiens proteins with one2one ortholog for PHOSPHORYLATION-T
IS_MOD
False
True
PROPORTION
SPECIES

d)


Proportion of H sapiens proteins with one2one ortholog for PHOSPHORYLATION-Y
IS_MOD
False
True
PROPORTION
SPECIES

e)

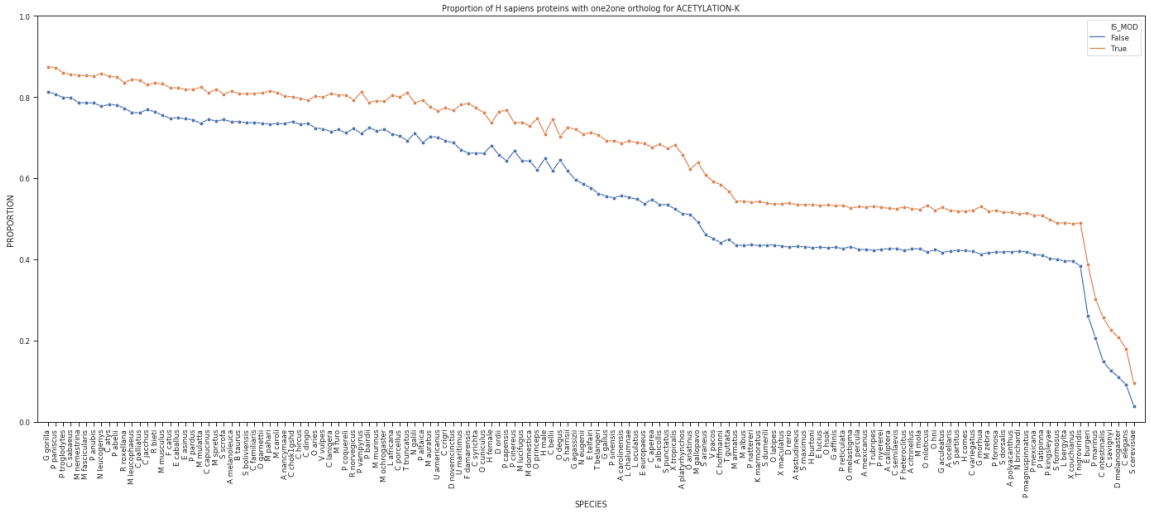


Figure 8. Conservation of post-translationally modified proteins

The proportion of human proteins with a one to one ortholog observed in a given species. The species are sorted in descending order along the x-axis by proportion of proteins, both modified an unmodified, that have a one to one ortholog within the given species. Proteins with at least one or more S-p (a), K-ub (b), T-p (c), Y-p (d), and K-ac (e) site are shown in orange. The unmodified proteins for each group are shown in blue.

Evolutionary Constraints of Post-Translational Modification Sites

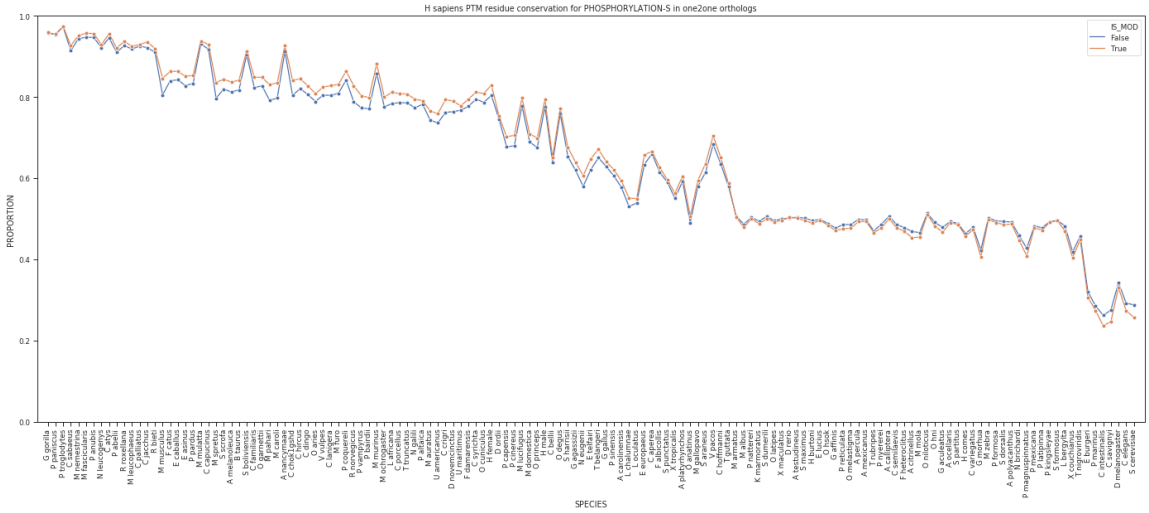
Conservation of post-transnational modification sites was examined using global pairwise alignments of all human proteins and their corresponding one to one orthologs. The null hypothesis is tested that modified and unmodified residues are equally conserved within one to one orthologs. Figure 9 includes the proportion of conserved residues for the five most frequently observed post-translational modifications within the data set including S-p, K-ub, T-p, Y-p, and K-ac. Significance is reported using a Pearson Chi-squared test and Bonferroni corrected significance level of $\alpha = 0.05/21 = 0.002$.

For proteins that include one or more S-p site, a significantly greater proportion of modified residues are conserved within roughly 62% of the species (89/143) surveyed. The 62% include 89 of the most closely related species to human. Notably, proteins with one or more K-ub site demonstrated a significantly greater proportion of modified residues conserved within all species surveyed. Proteins with one or more T-p site exhibit a significant proportion of modified residues conserved within ~49% of the species and those with one or more Y-p site within ~46% of species. Similar to S-p, the T-p and Y-p proteins demonstrate significance in species the most closely related to humans. Proteins with one or more K-ac sites demonstrate strong conservation with a significantly greater proportion of modified residues conserved in ~99% of species surveyed.

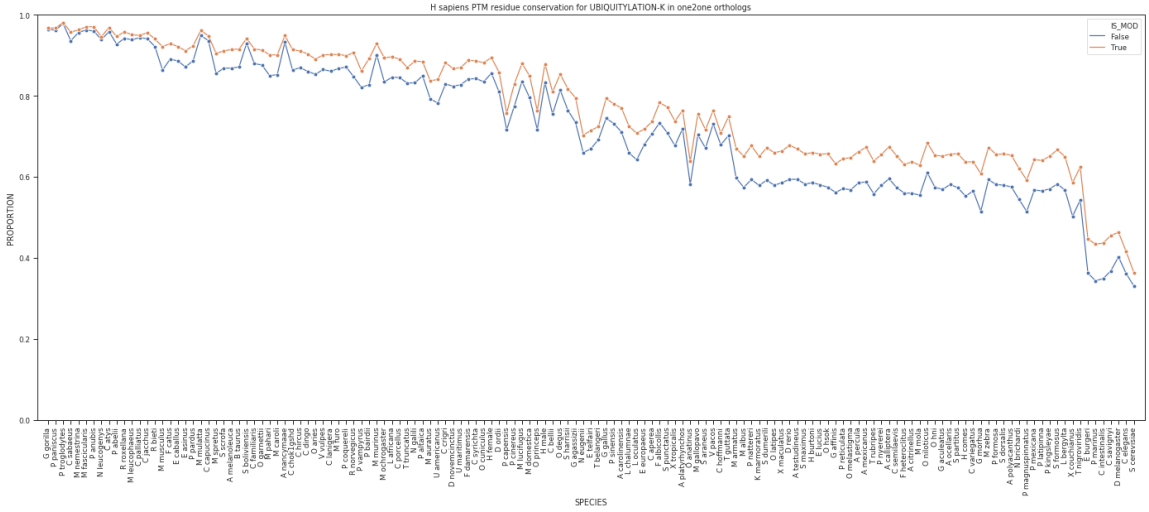
Proteins with one or more K-sc sites (Appendix 1 – Supplementary Figures) have significantly higher proportions of conserved modified sites in ~90% of the species

surveyed. K-sc proteins demonstrate no distinguishable difference within the most closely related species such as *G. gorilla* and *P. paniscus*. However, the difference in conserved modified and unmodified residues increases as more distant species are investigated. K-sm proteins (Appendix 1 – Supplementary Figures) exhibit a similar trend to K-sc with no significance within the most closely related species. However, K-sm proteins have a significant proportion of modified residues conserved within ~72% of species, which includes species up to *C. hoffmanni* before fading. R-m1 and R-m2 (Appendix 1 – Supplementary Figures) have trends that resemble K-sm proteins with a significant proportion of modified residues conserved in ~70% and ~49% of species excluding the most closely related and the most distant. K-m1 (Appendix 1 – Supplementary Figures) have a significant proportion of modified residues conserved in ~45% of species spread throughout the data set, consisting of closely related and the most distant of relatives. O-GlcNAc-T and O-GlcNAc-S proteins (INSERT APPENDIX HERE) exhibit a significant proportion of conserved modified residues in ~42% and 27% of species respectively. In both cases the closest species to human are omitted and significance is observed before beginning to dissipate around *D. novemcinctus*. R-me, K-ne, K-m2, K-m3, and K-me proteins (INSERT APPENDIX HERE) represent a group for which the fewest species demonstrate a significant conservation of modified residues with only ~13%, ~5%, ~2%, <1%, and <1% respectively. Finally, D-ca, O-GalNAc-S, O-GalNAc-T, and C-pa do not exhibit any significant proportion of conserved, modified residues in any of the species surveyed.

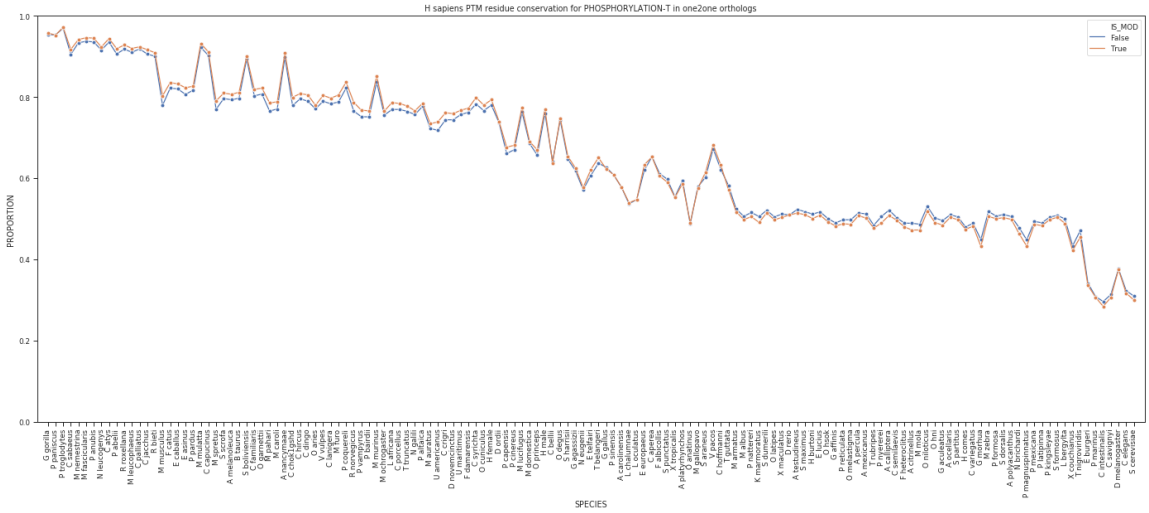
a)


H sapiens PTM residue conservation for PHOSPHORYLATION-S in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

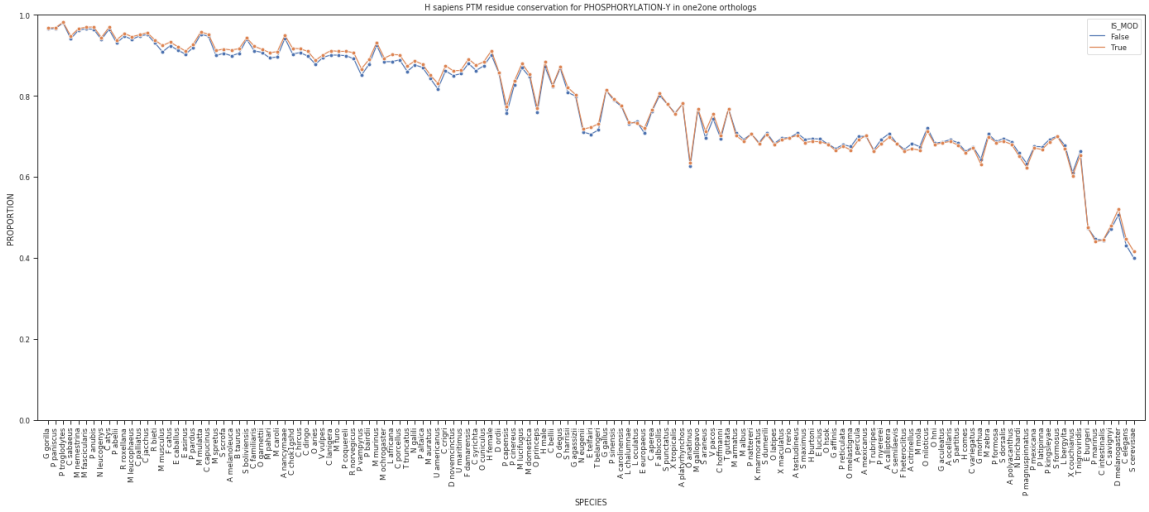
b)


H sapiens PTM residue conservation for UBIQUITYLATION-K in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

c)


H sapiens PTM residue conservation for PHOSPHORYLATION-T in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

d)


H sapiens PTM residue conservation for PHOSPHORYLATION-Y in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

e)

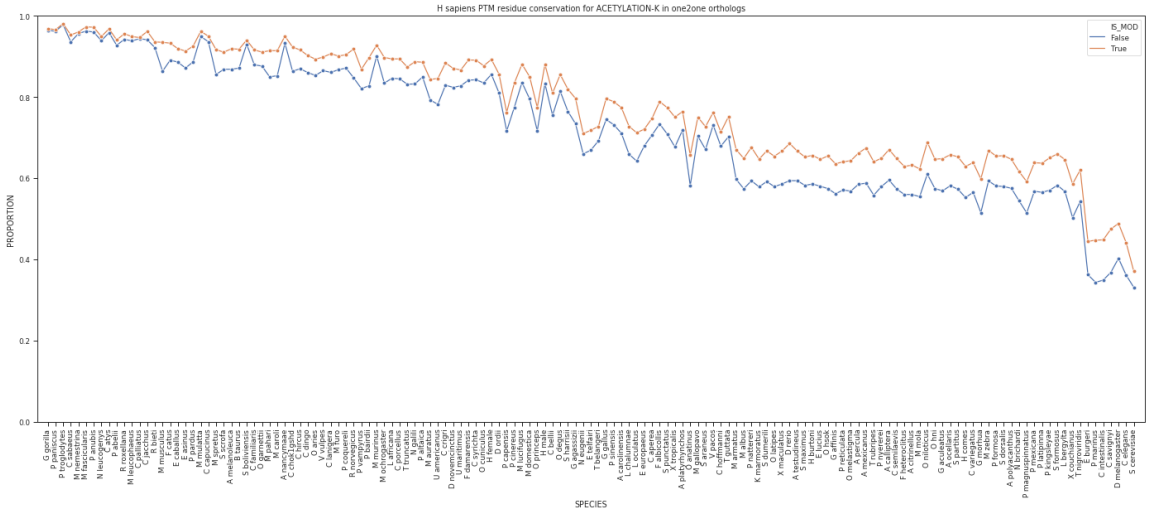


Figure 9. Conservation of PTM sites

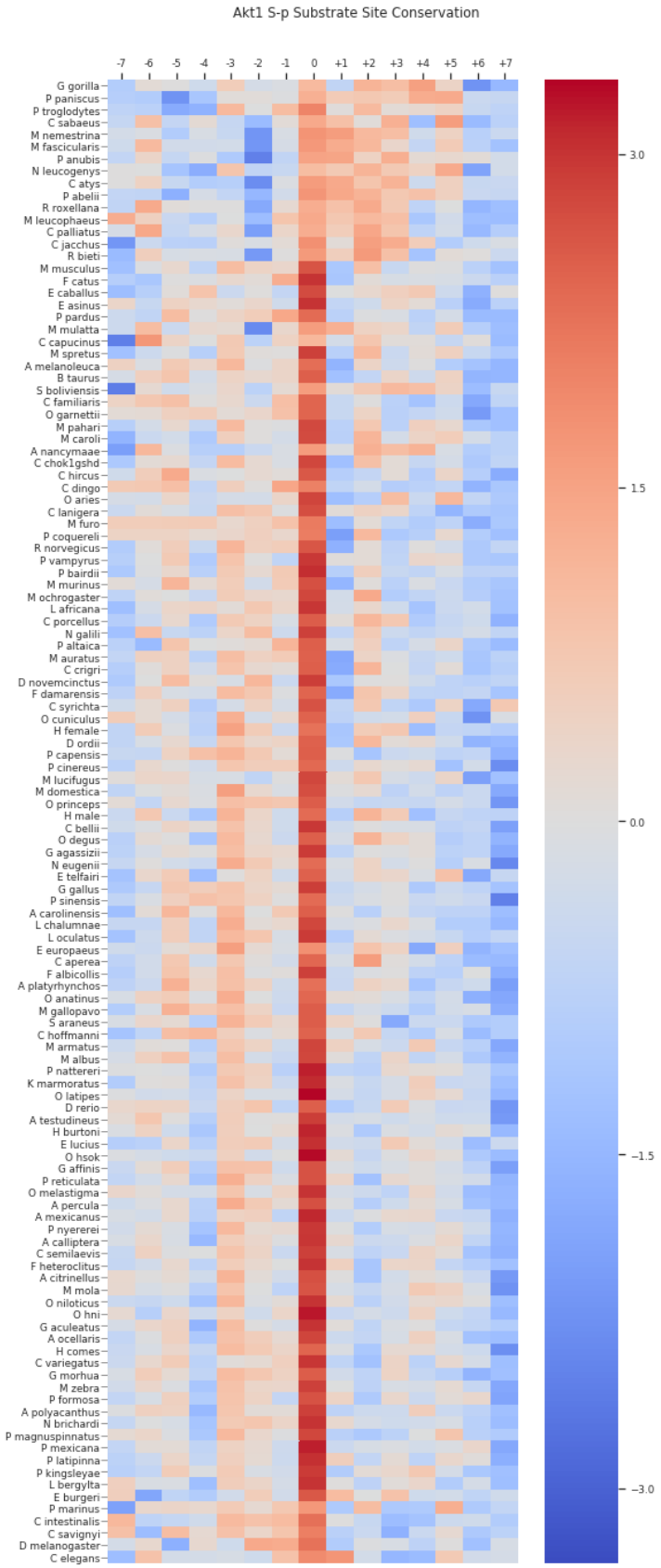
The proportion of modified (orange) and unmodified (blue) residues conserved for proteins with one or more S-p (a), K-ub (b), T-p (c), Y-p (d), and K-ac (e) site.

Sequence motifs are highly conserved

In addition to the conservation of PTM sites, we explored the conservation of their flanking regions including seven residues upstream and downstream of all modification sites surveyed. The AGC kinase Akt1 is used as an example to demonstrate the conservation of the associated recognition motif within its substrates. The Akt1 substrate data set was derived from PhosphoSitePlus. The set consists of 180 S-p and 65 T-p sites covering 183 distinct substrates. Akt1 recognizes sequence motifs consisting of arginine in the -3 and -5 positions within the flanking regions (Figure 12). Resulting heat maps generated (Figure 11) illustrate the conservation of flanking regions only for conserved modified residues. One to one orthologs are observed for Akt1 substrates in 122 out of the 143 species. The proportion of conserved residues at each position is normalized by subtracting the proportion of non-conserved residues. The z-scores are calculated for the resulting differences at each position within a species. The modified residue is considered position 0. In concurrence with S-p and T-p sites observed across the entire proteome, Akt1 substrates exhibit strong conservation for modified residues.

In the case of S-p substrates (Figure 11), the preference in conservation for the -3 and -5 positions is difficult to resolve for the most closely related species to human such as *P. paniscus, C. sabaeus, and M. fascicularis*. However, a greater proportion of conserved residues for the -3 and -5 positions is observed in more distantly related species such as *R. norvegicus, A. carolinesis, and G. aculeatus.* The evolutionary profile is similar for T-p substrates. The conservation of the -3 and -5 motif can be more easily resolved even in species most closely related to human. Overall,these results make clear that not only the PTM sites, but also their recognition motifs are highly conserved.

a)


Akt1 S-p Substrate Site Conservation
-7 -6 -5 -4 -3 -2 -1 0 +1 +2 +3 +4 +5 +6 +7
G gorilla
P paniscus
P troglodytes
C sabaeus
M nemestrina
M fascicularis
P anubis
N leucogenys
C atys
P abelii
R roxellana
M leucophaeus
C palliatus
C jacchus
R bieti
M musculus
F catus
E caballus
E asinus
P pardus
M mulatta
C capucinus
M spretus
A melanoleuca
B taurus
S boliviensis
C familiaris
O garnettii
M pahari
M caroli
A nancymaae
C chok1gshd
C hircus
C dingo
O aries
C lanigera
M furo
P coquereli
R norvegicus
P vampyrus
P bairdii
M murinus
M ochrogaster
L africana
C porcellus
N galili
P altaica
M auratus
C crigri
D novemcinctus
F damarensis
C syrichta
O cuniculus
H female
D ordii
P capensis
P cinereus
M lucifugus
M domestica
O princeps
H male
C bellii
O degus
G agassizii
N eugenii
E telfairi
G gallus
P sinensis
A carolinensis
L chalumnae
L oculatus
E europaeus
C aperea
F albicollis
A platyrhynchos
O anatinus
M gallopavo
S araneus
C hoffmanni
M armatus
M albus
P nattereri
K marmoratus
O latipes
D rerio
A testudineus
H burtoni
E lucius
O hsok
G affinis
P reticulata
O melastigma
A percula
A mexicanus
P nyererei
A calliptera
C semilaevis
F heteroclitus
A citrinellus
M mola
O niloticus
O hni
G aculeatus
A ocellaris
H comes
C variegatus
G morhua
M zebra
P formosa
A polyacanthus
N brichardi
P magnuspinnatus
P mexicana
P latipinna
P kingsleyae
L bergylta
E burgeri
P marinus
C intestinalis
C savignyi
D melanogaster
C elegans
3.0
1.5
0.0
−1.5
−3.0

b)

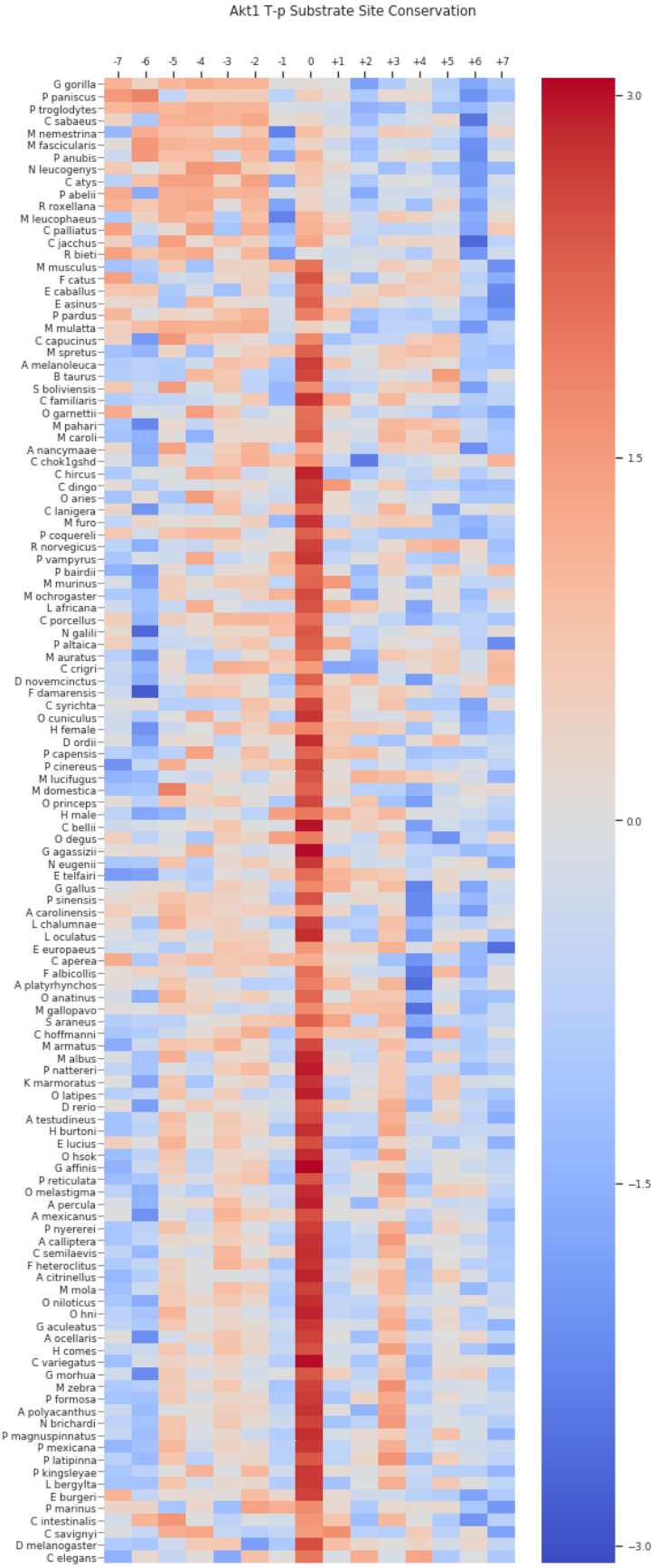

Akt1 T-p Substrate Site Conservation
-7 -6 -5 -4 -3 -2 -1 0 +1 +2 +3 +4 +5 +6 +7
G gorilla
P paniscus
P troglodytes
C sabaeus
M nemestrina
M fascicularis
P anubis
N leucogenys
C atys
P abelii
R roxellana
M leucophaeus
C palliatus
C jacchus
R bieti
M musculus
F catus
E caballus
E asinus
P pardus
M mulatta
C capucinus
M spretus
A melanoleuca
B taurus
S boliviensis
C familiaris
O garnettii
M pahari
M caroli
A nancymaae
C chok1gshd
C hircus
C dingo
O aries
C lanigera
M furo
P coquereli
R norvegicus
P vampyrus
P bairdii
M murinus
M ochrogaster
L africana
C porcellus
N galili
P altaica
M auratus
C crigri
D novemcinctus
F damarensis
C syrichta
O cuniculus
H female
D ordii
P capensis
P cinereus
M lucifugus
M domestica
O princeps
H male
C bellii
O degus
G agassizii
N eugenii
E telfairi
G gallus
P sinensis
A carolinensis
L chalumnae
L oculatus
E europaeus
C aperea
F albicollis
A platyrhynchos
O anatinus
M gallopavo
S araneus
C hoffmanni
M armatus
M albus
P nattereri
K marmoratus
O latipes
D rerio
A testudineus
H burtoni
E lucius
O hsok
G affinis
P reticulata
O melastigma
A percula
A mexicanus
P nyererei
A calliptera
C semilaevis
F heteroclitus
A citrinellus
M mola
O niloticus
O hni
G aculeatus
A ocellaris
H comes
C variegatus
G morhua
M zebra
P formosa
A polyacanthus
N brichardi
P magnuspinnatus
P mexicana
P latipinna
P kingsleyae
L bergylta
E burgeri
P marinus
C intestinalis
C savignyi
D melanogaster
C elegans
3.0
1.5
0.0
−1.5
−3.0

Figure 11. Akt1 substrate site conservation

Flanking regions of conserved Akt1 S-p (a) and T-p (b) substrate modification sites. Z-scores are calculated for the difference in proportions of conserved and non-conserved residues. A higher score indicates a greater proportion of conserved residues. For S-p substrates (a) of 180 sites are included covering 146 proteins. For T-p substrates (b) 65 sites are included covering 60 proteins

Figure 12. Akt1 sequence logo

Substrate sequence logo for Akt1 generated using 748 input sequences (Hornbeck et al., 2019).

Chapter IV.

Discussion and Future Directions

Previous studies have shown the importance of post-translational modifications, in particular phosphorylation, in protein-protein interactions and signaling. The goal of our study was to elucidate structural and evolutionary patterns in PTM sites, relying on data sets from PhosphoSitePlus. Many of the modification sites included in the data set were first observed with mass spectrometry (MS). Consideration must be given to the bias MS can introduce due to the enormous variation of protein abundance within a sample. Often the bias of abundance is combated with techniques such as immuno-depletion to remove the high abundant proteins prior to analysis (Bylund & Henriksson, 2015). Nevertheless, in the context of this study it must be considered. The potential exists that a larger proportion of modification sites belong to high abundance proteins.

Since the study relies on millions of predictions and calculations, high accuracy large-scale computing in the cloud was required. The findings presented in this work would not have come to fruition without significant effort in developing a framework for scaling the secondary structure predictions and global pairwise alignments to a distributed computing environment. The powers of existing open source technologies such as Docker, Django, and Kubernetes are harnessed into a unique framework that is capable of managing a diverse set of computational workloads. The approach yielded

significant time savings shortening the estimated time of computation for both the secondary structure predictions and alignments from months to days.

The first objective of this study was to understand structural constraints imposed on human PTMs. We found that most PTM types are enriched in unstructured regions. K-ac, Y-p, K-ub, K-sc, and K-ne modifications are the only PTM types that have less than 50% of their sites located within unstructured regions.

Concordant with the enrichment in loops and turns, our analysis revealed significantly higher mean accessible surface areas for all PTM types, excluding K-ne, K-sc, and C-pa, when compared to their corresponding unmodified residues. These findings make clear that substrate sites have to be accessible to their upstream regulators. For downstream effects their localization on the surface is also required for protein-protein interactions and modified domain functions. Unexpectedly, some PTM types showed enrichment in regular structures including helices and beta strands, which needs further investigation.

The second objective of this study was to understand evolutionary constraints imposed on human PTMs. Taken together we found a strong conservation of PTMs – both at the protein and site level. Human proteins that include frequently observed modification types such as phosphorylation, ubiquitination, and acetylation have a significantly greater proportion of one to one orthologs than their corresponding unmodified groups in all species included in this study. Overall, the majority of PTM types exhibit a significant proportion of one to one orthologs in more closely related species. The PTMs that lack a significant orthology trend are K-me, K-m3, K-m2, R-me, K-ne, and C-pa.

Notably, our site level analysis identified a high degree of conservation for ubiquitination, acetylation, and succinylation in more than 90% of the species. Similarly, sumoylation, mono-methylation, and phosphorylation exhibited site conservation in the majority of the species (>50%). It is difficult to conclude any significant conservation within the most closely related species due to the high level of similarity between aligned sequences. For example, equal amounts of conservation are observed for Y-p sites and their corresponding unmodified Y residues within the three most closely related species to human including *G. gorilla*, *P. paniscus*, and *P. troglodytes*. While the high sequence similarity of closely related species made it difficult to observe differences at the residue level, the alignment quality of distantly related species made it difficult to accurately align PTM sites. Notably, PTM sites have a preference for unstructured loop regions of a protein, and the structural regions of higher disorder further contributes to the difficulties of measuring conservation in distantly related species.

These conservation constraints extend to the flanking regions of PTMs. For example, the serine/threonine kinase Akt1 interacts with 249 unique sites across 185 distinct substrates within the human proteome. It plays a critical role in cell survival and metabolism regulation. In addition, Akt1 is a substrate with 60 documented modification sites of its own (Hornbeck et al., 2019). A pattern of conservation within the flanking regions of Akt1 substrates is observed that mimics the expected sequence logo. In addition to the modification site, the residues in the -3 and -5 positions are more conserved than any other position. The structural and evolutionary constraints imposed on the upstream regulators and downstream substrates of a kinase such as Akt1

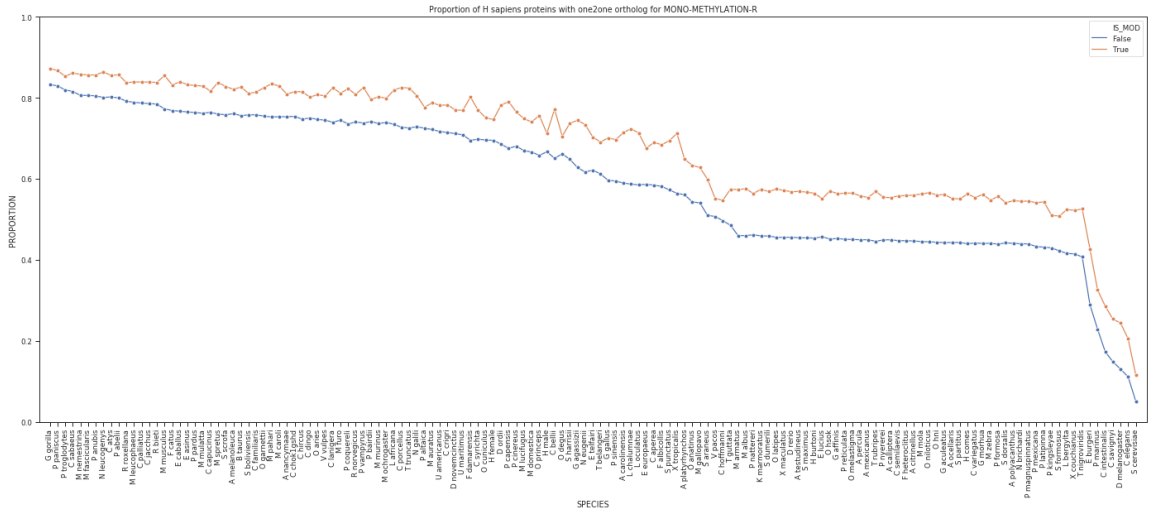
contributes significantly to the understanding of the pathways involved. Many of which are linked to variety of cancers such as breast, prostate, and liver.

In summary, the identification of the structural and evolutionary constraints imposed on PTMs did not only reveal previously unknown patterns, but also presents a unique opportunity for enhanced mutation effect prediction. Many tools utilize a variety of approaches to score the impact of a mutation on protein function such as SIFT (P. C. Ng, 2003), Mutation Assessor (Reva, Antipin, & Sander, 2011), Polyphen-2 (Adzhubei et al., 2010), Condel (González-Pérez & López-Bigas, 2011), and CHASM (Carter et al., 2009). Despite using similar techniques such as evolution, sequence homology, and protein structure, it has been demonstrated that they lack consensus when applied to an independent data set (Martelotto et al., 2014). We propose that including features derived from the evolutionary and structural constraints of PTMs in conjunction with improved secondary structure predictions and restricting sequence alignments to global pairwise alignments will improve the accuracy of a mutation effect predictor.

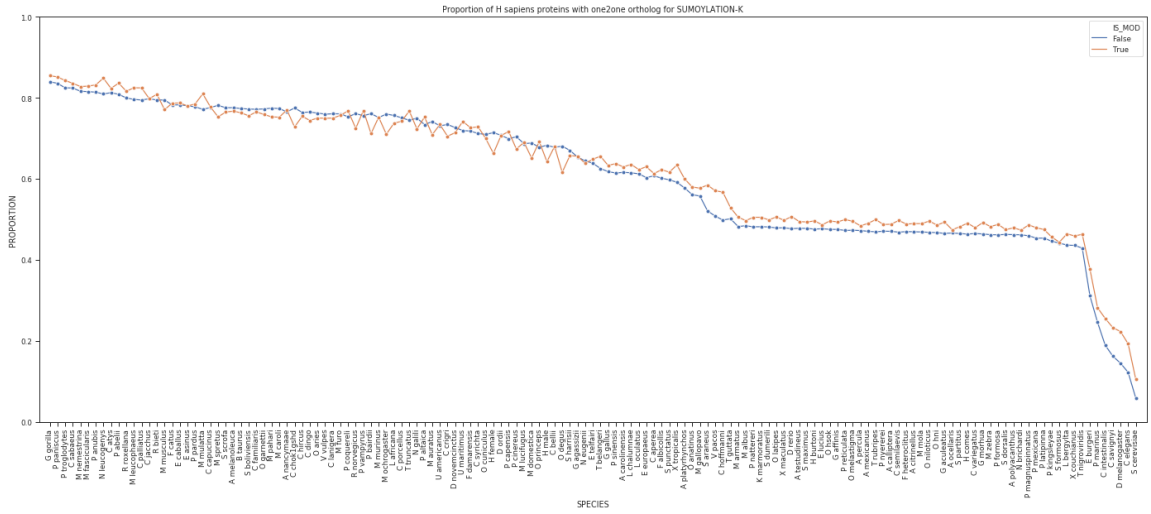
Appendix 1.

Supplementary Figures

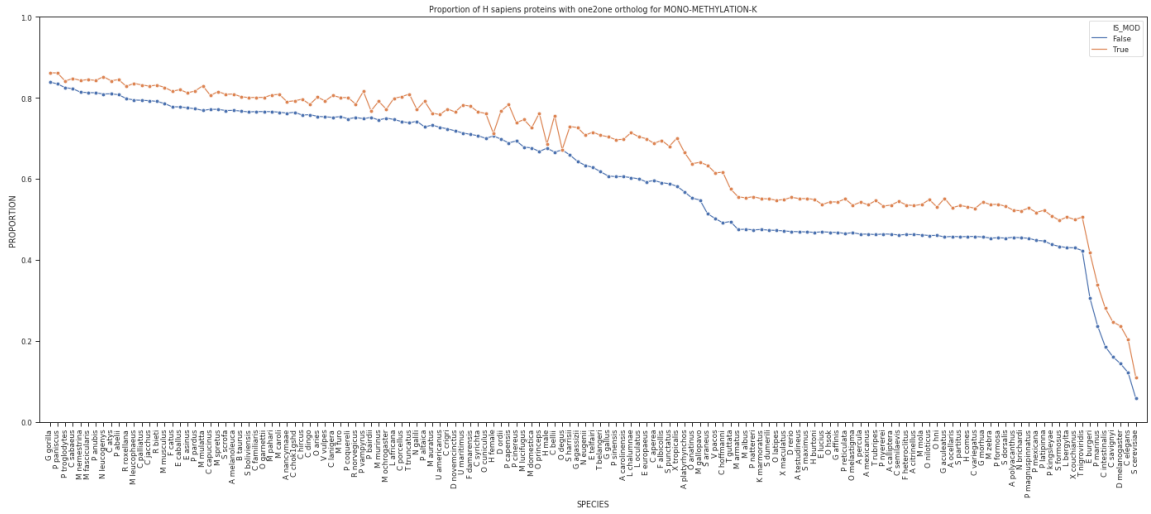
a)



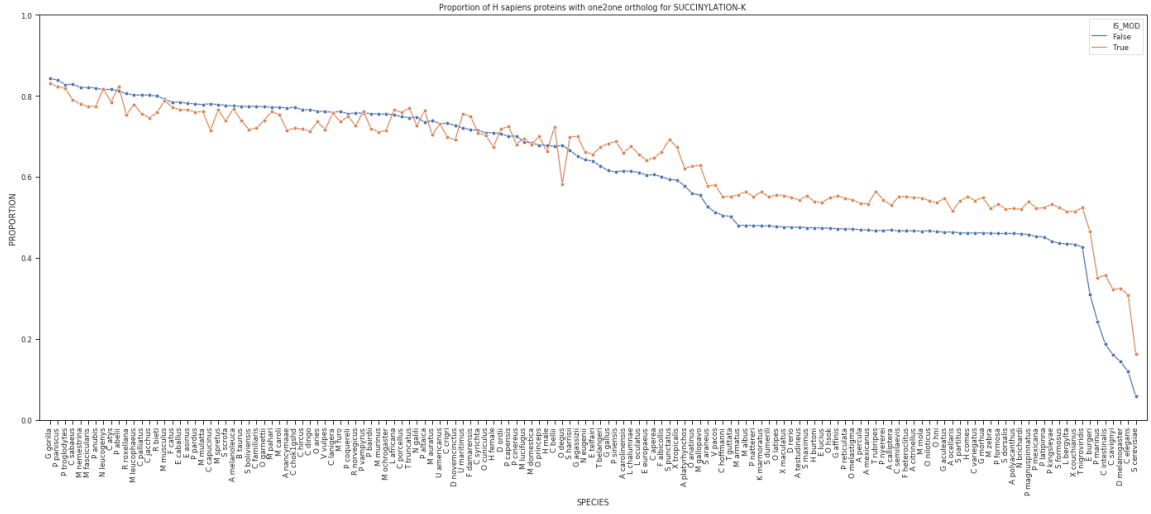
b)



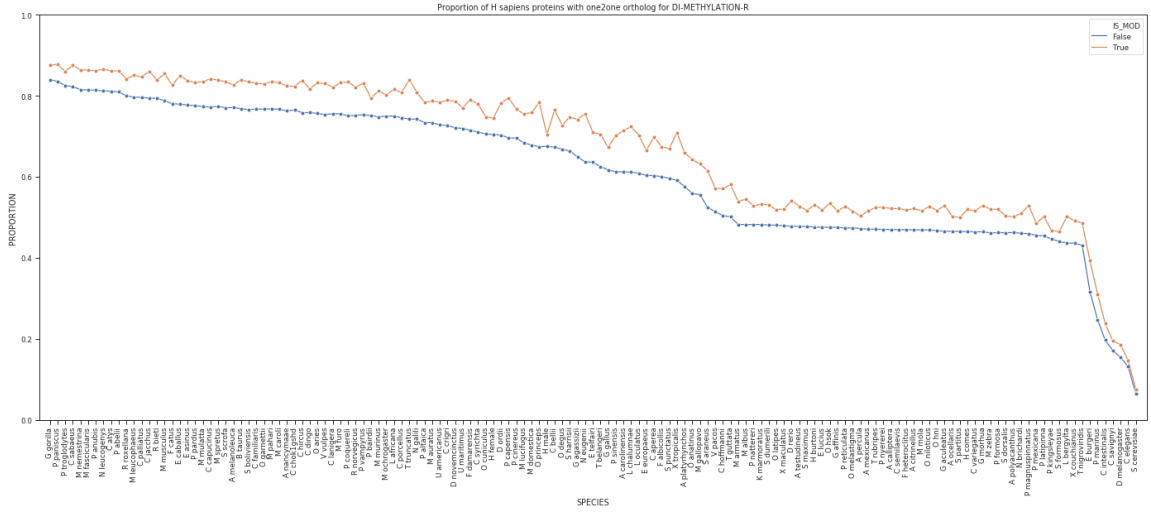
c)


Proportion of H sapiens proteins with one2one ortholog for MONO-METHYLATION-K
IS_MOD
False
True
PROPORTION
SPECIES

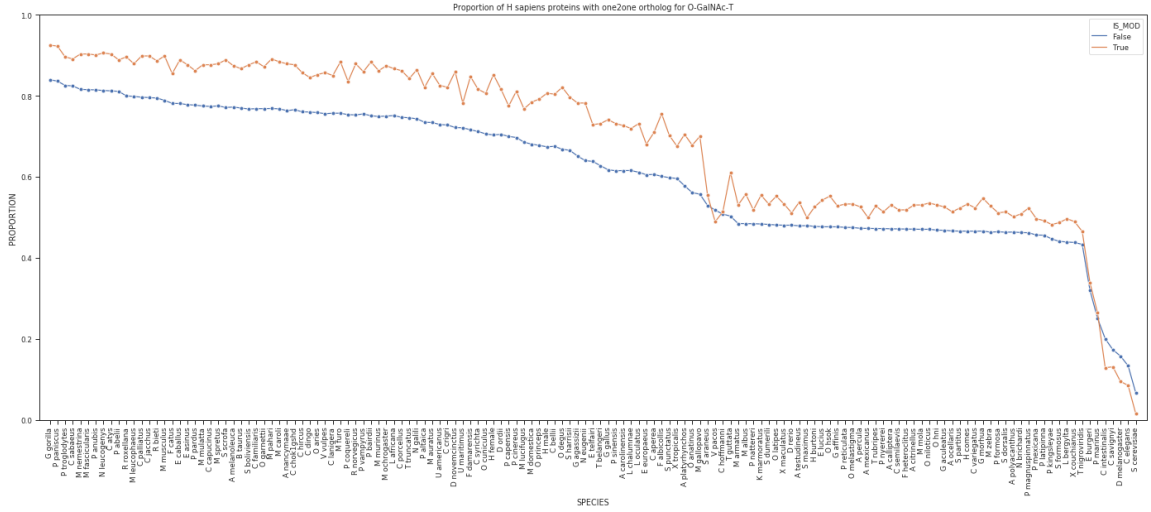
d)


Proportion of H sapiens proteins with one2one ortholog for SUCCINYLATION-K
IS_MOD
False
True
PROPORTION
SPECIES

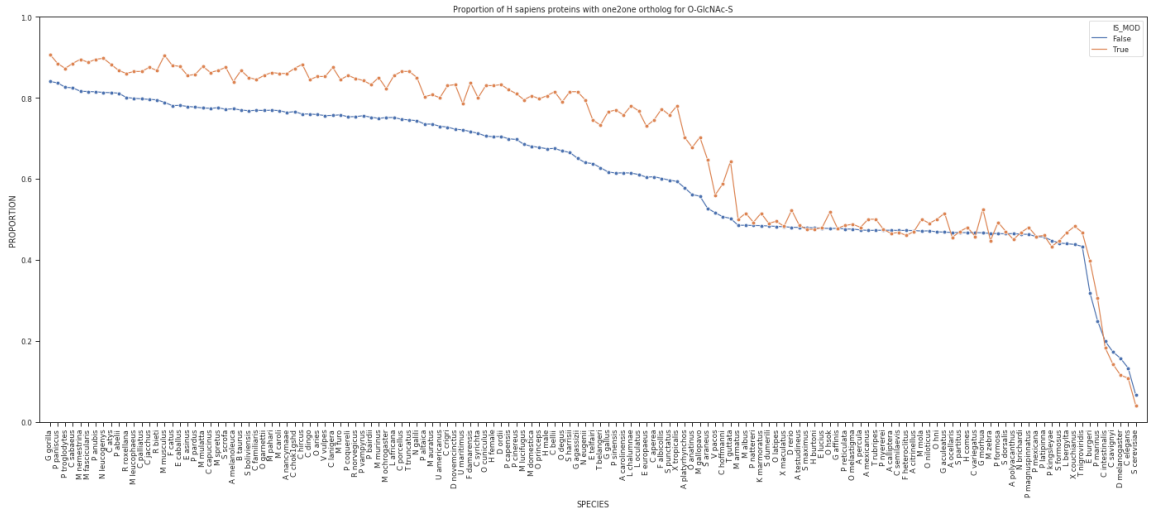
e)


Proportion of H sapiens proteins with one2one ortholog for DI-METHYLATION-R
IS_MOD
False
True
PROPORTION
SPECIES

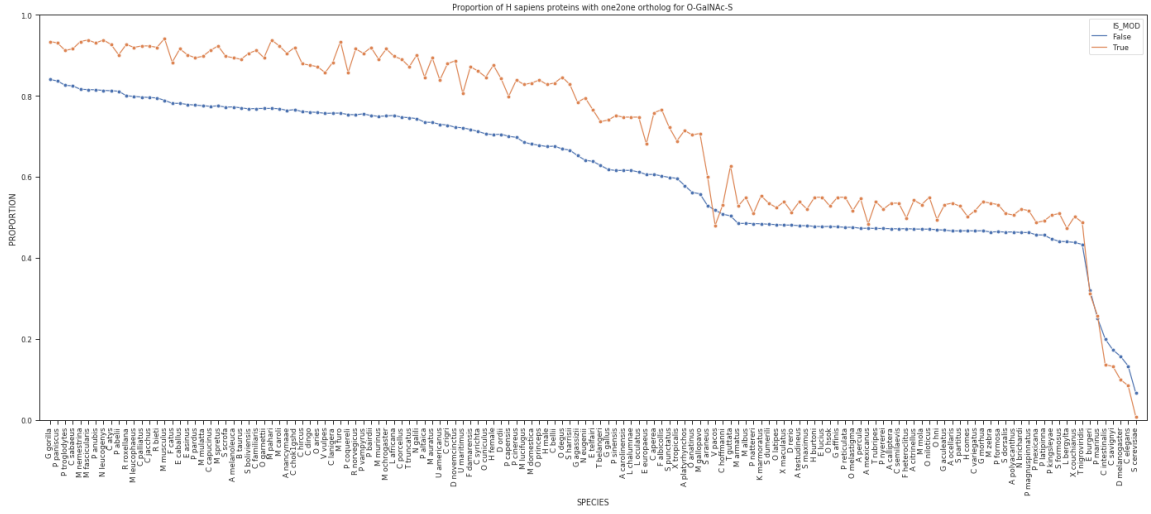
f)


Proportion of H sapiens proteins with one2one ortholog for O-GalNAc-T
IS_MOD
False
True
PROPORTION
SPECIES

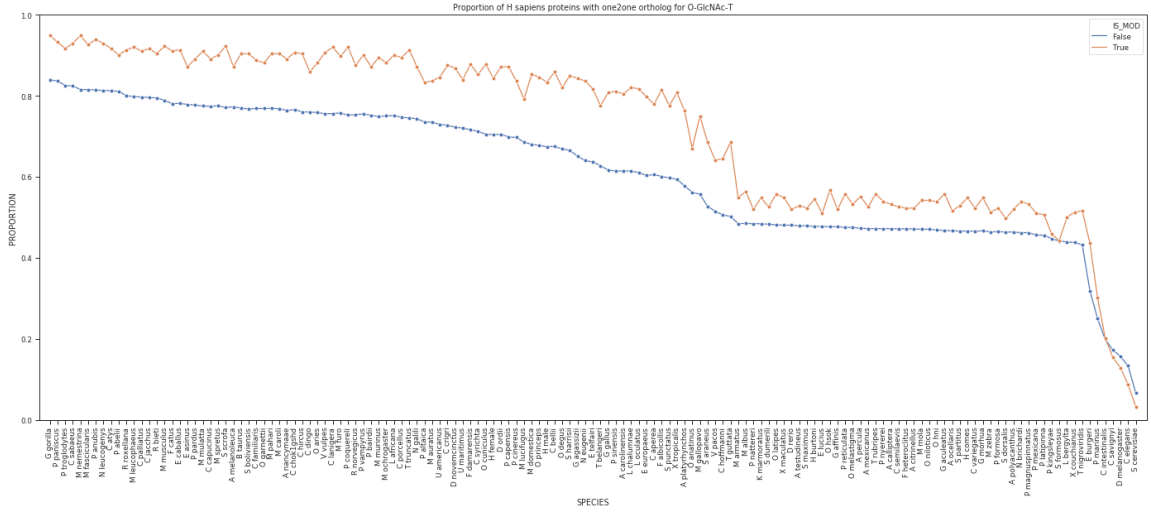
g)


Proportion of H sapiens proteins with one2one ortholog for O-GlcNAc-S
IS_MOD
False
True
PROPORTION
SPECIES

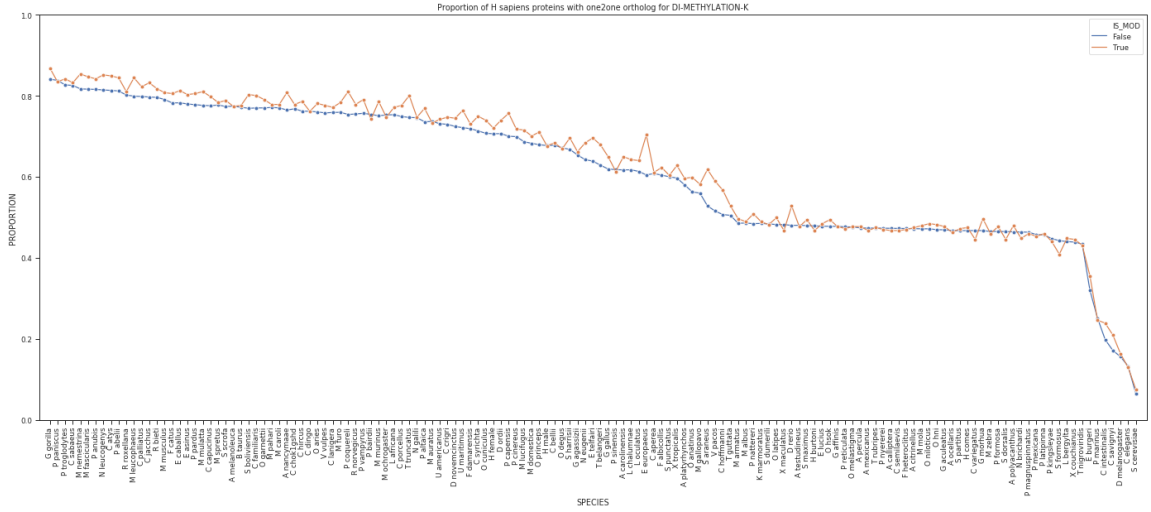
h)


Proportion of H sapiens proteins with one2one ortholog for O-GalNAc-S
IS_MOD
False
True
PROPORTION
SPECIES

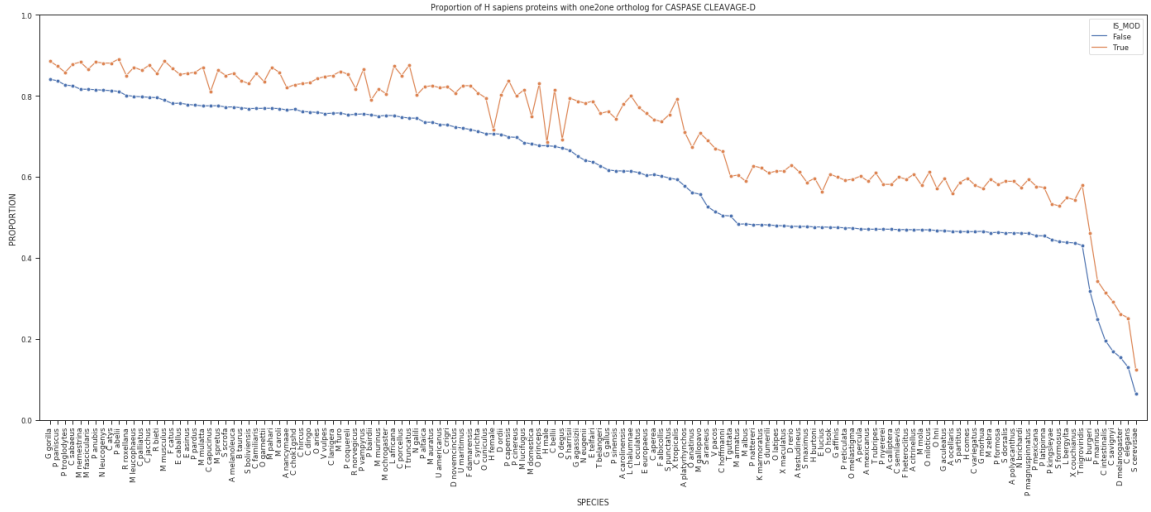
i)


Proportion of H sapiens proteins with one2one ortholog for O-GlcNAc-T
IS_MOD
False
True
PROPORTION
SPECIES

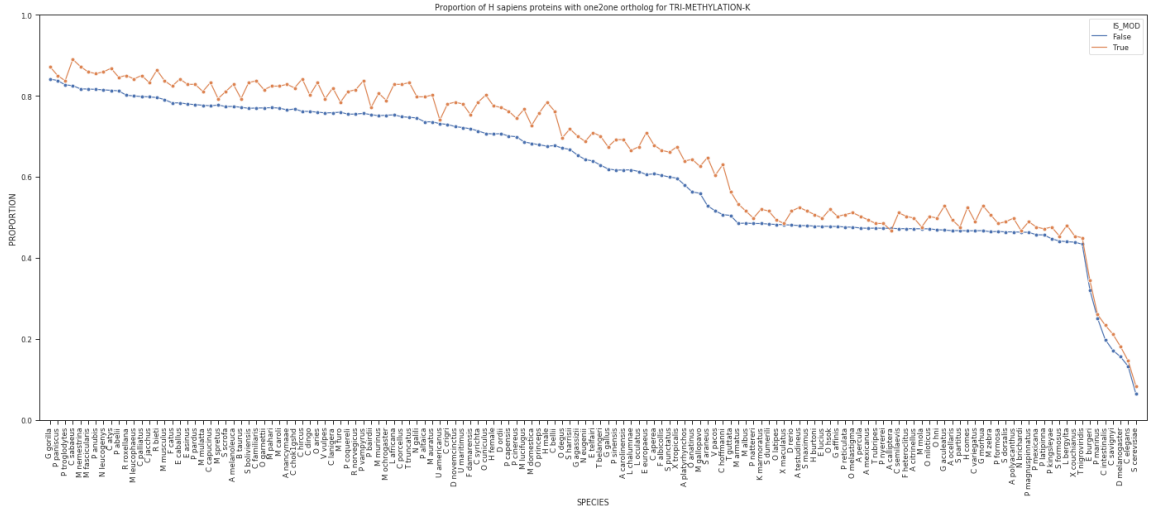
j)


Proportion of H sapiens proteins with one2one ortholog for DI-METHYLATION-K
IS_MOD
False
True
PROPORTION
SPECIES

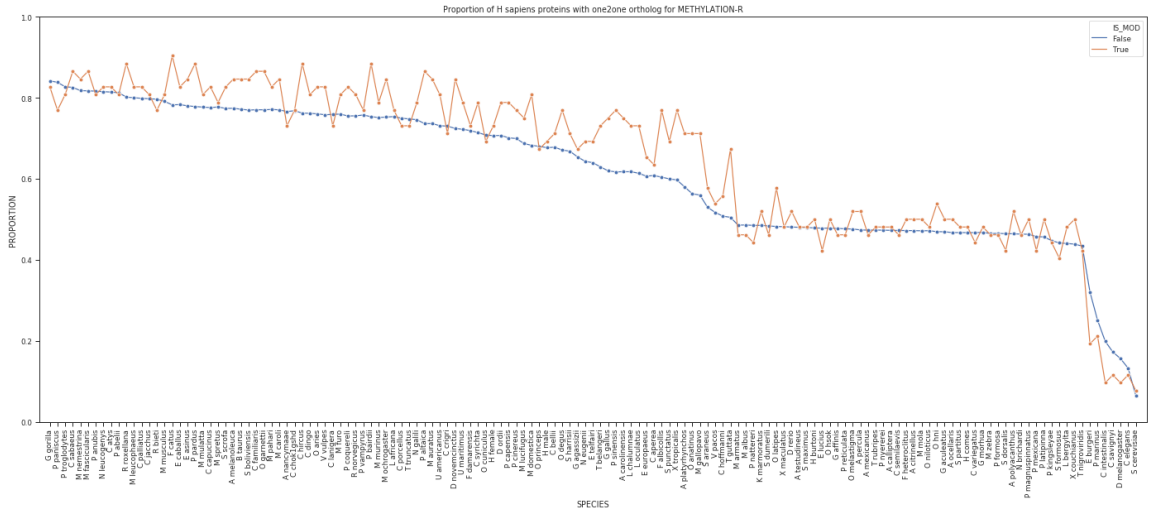
k)


Proportion of H sapiens proteins with one2one ortholog for CASPASE CLEAVAGE-D
IS_MOD
False
True
PROPORTION
SPECIES

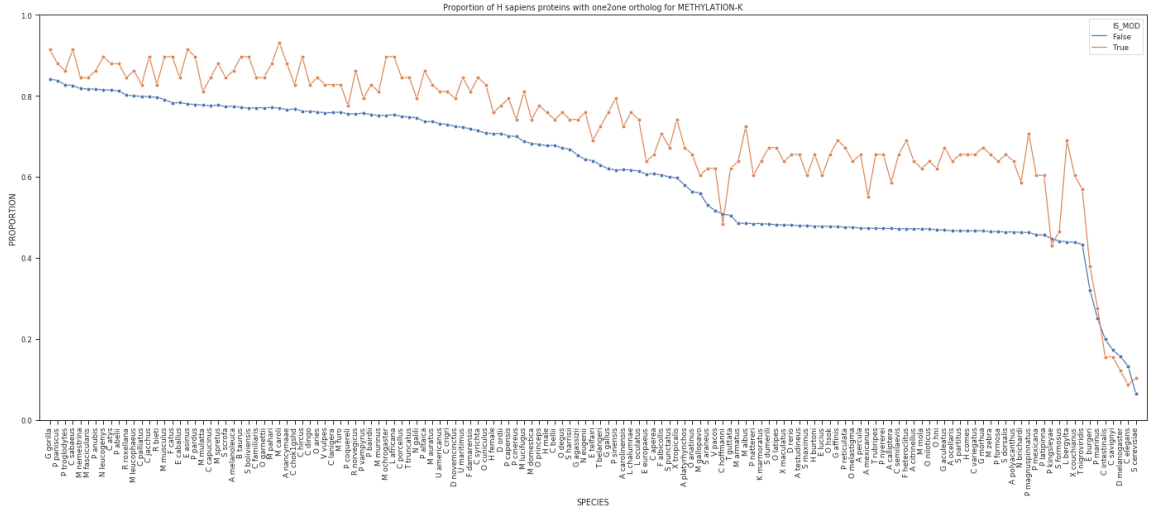
l)


Proportion of H sapiens proteins with one2one ortholog for TRI-METHYLATION-K
IS_MOD
False
True
PROPORTION
SPECIES

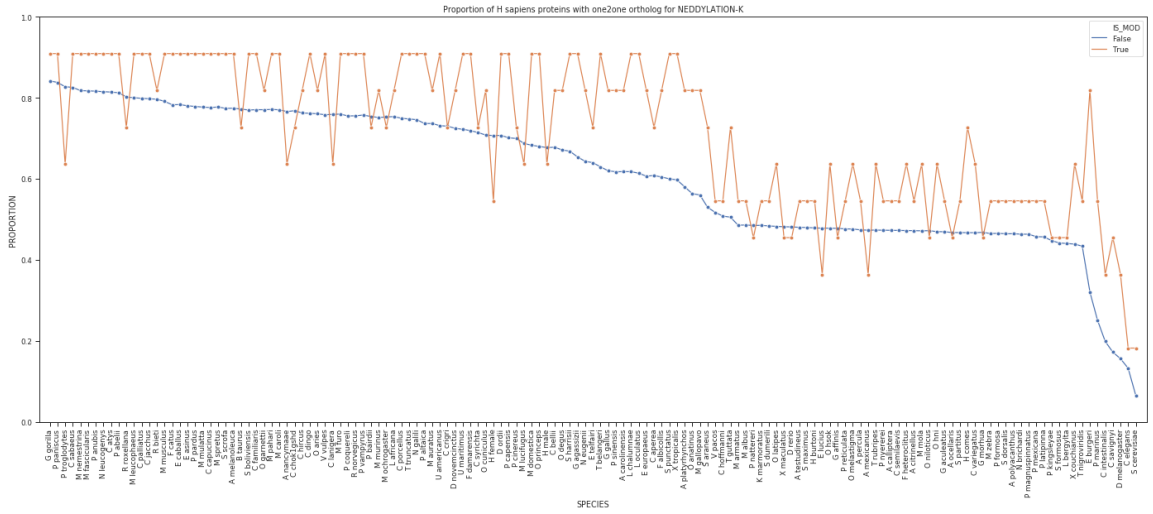
m)


Proportion of H sapiens proteins with one2one ortholog for METHYLATION-R
IS_MOD
False
True
PROPORTION
SPECIES

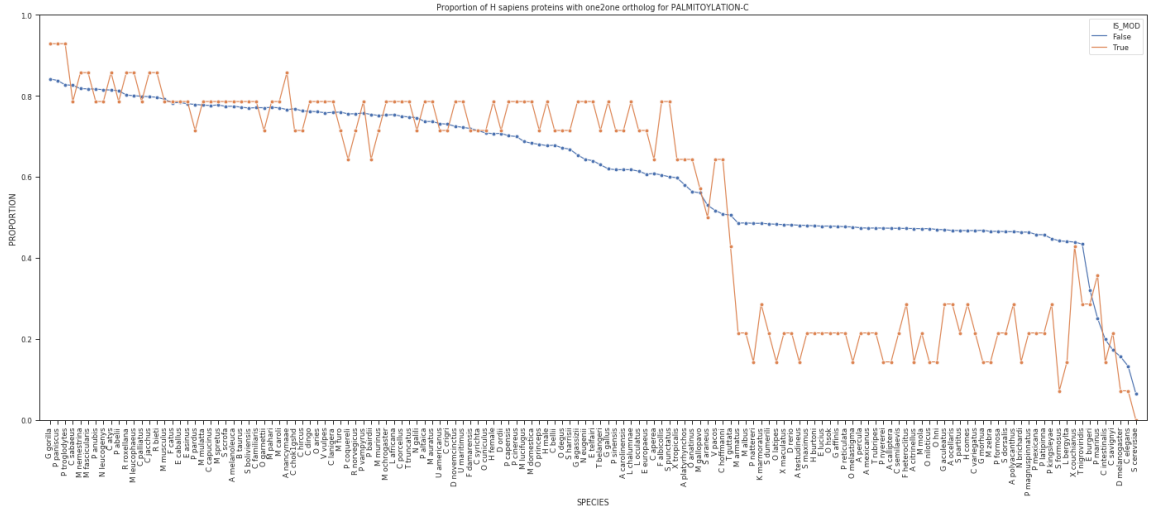
n)


Proportion of H sapiens proteins with one2one ortholog for METHYLATION-K
IS_MOD
False
True
PROPORTION
SPECIES

o)



p)

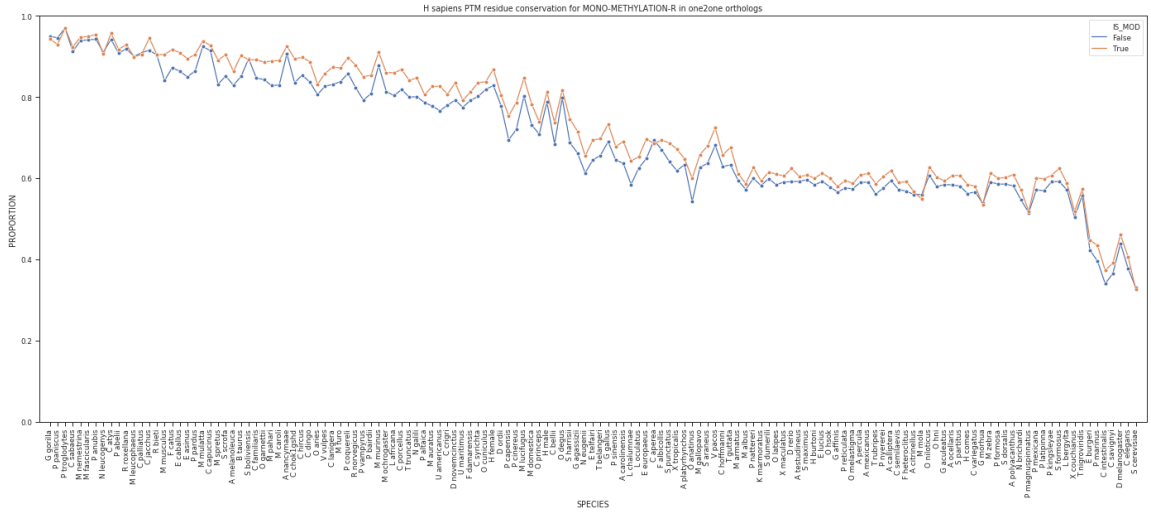


Supplementary Figure 1. Conservation of post-translationally modified proteins

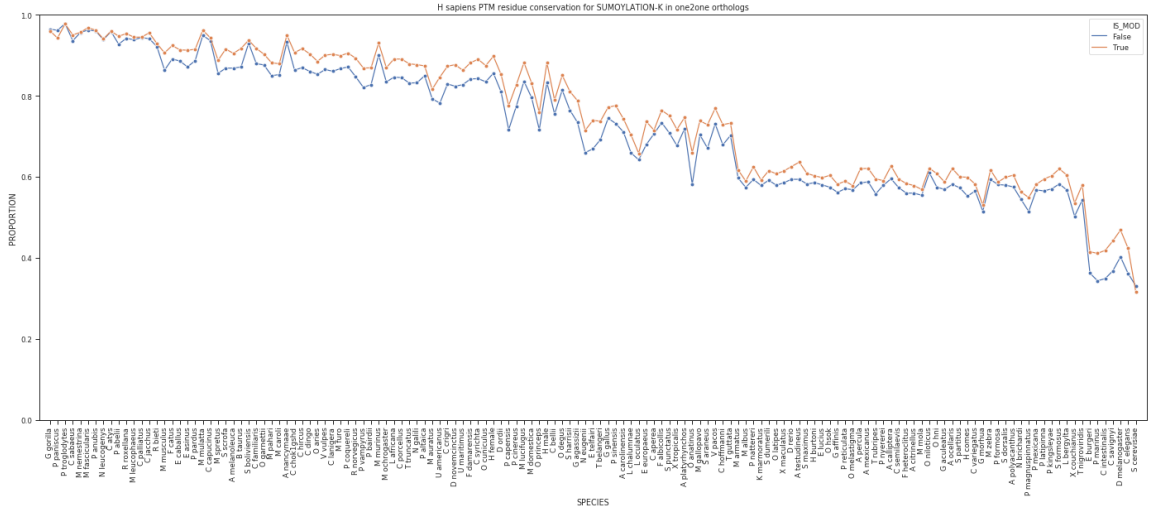
The proportion of human proteins with a one to one ortholog observed in a given species. The species are sorted in descending order along the x-axis by proportion of proteins, both modified an unmodified, that have a one to one ortholog within the given species. Proteins with at least one or more R-m1 (a), K-sm (b), K-m1 (c), K-sc (d), R-m2 (e), O-

GalNAc-T (f), O-GlcNA-S (g), O-GalNAc-S (h), O-GlcNAc-T (i), K-m2 (j), D-ca (k), K-m3 (l), R-m (m), K-m (n), K-ne (o), and C-pa (p) site are shown in orange. The unmodified proteins for each group are shown in blue.

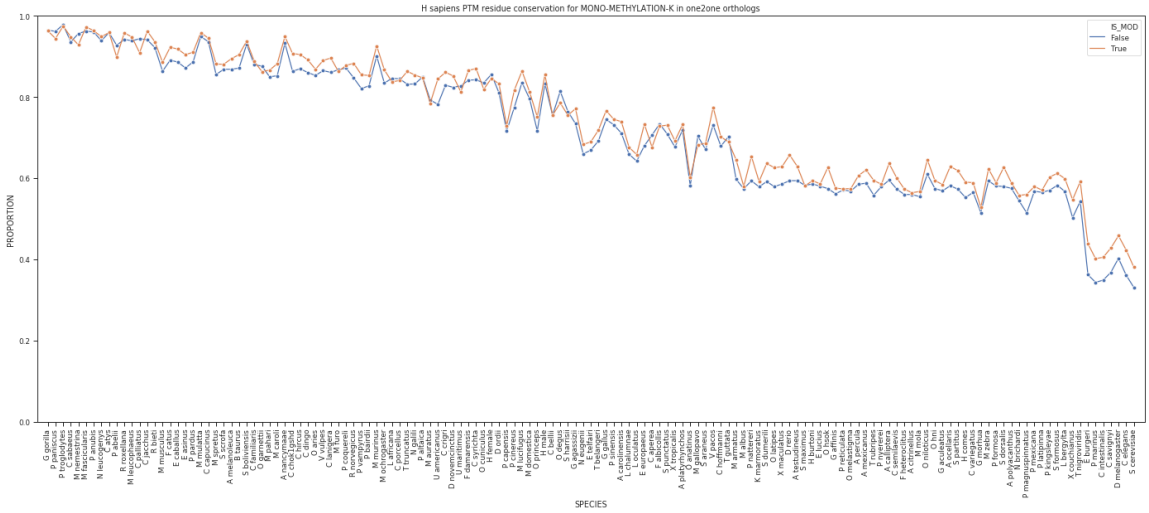
a)


H sapiens PTM residue conservation for MONO-METHYLATION-R in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

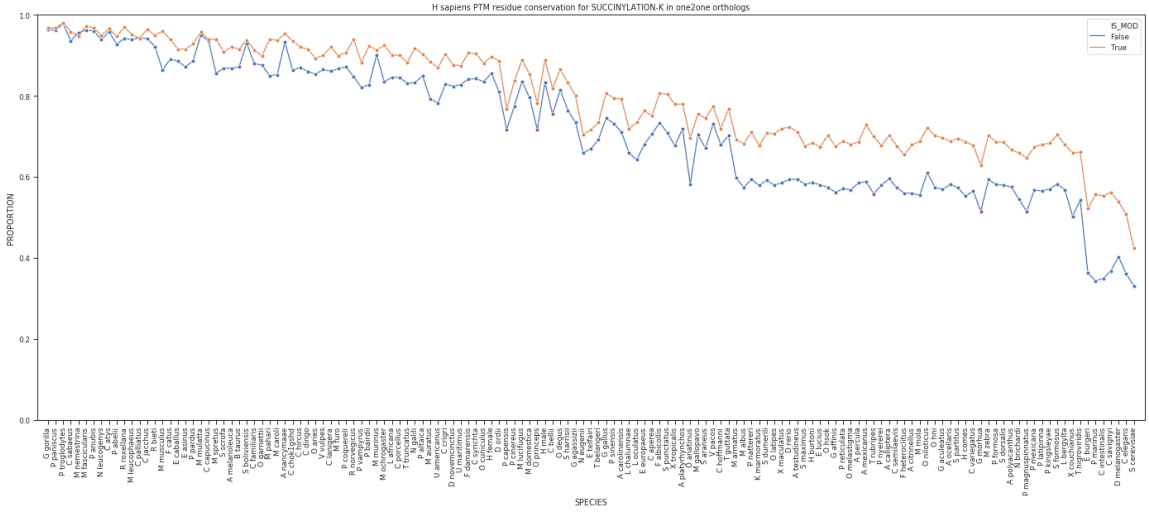
b)


H sapiens PTM residue conservation for SUMOYLATION-K in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

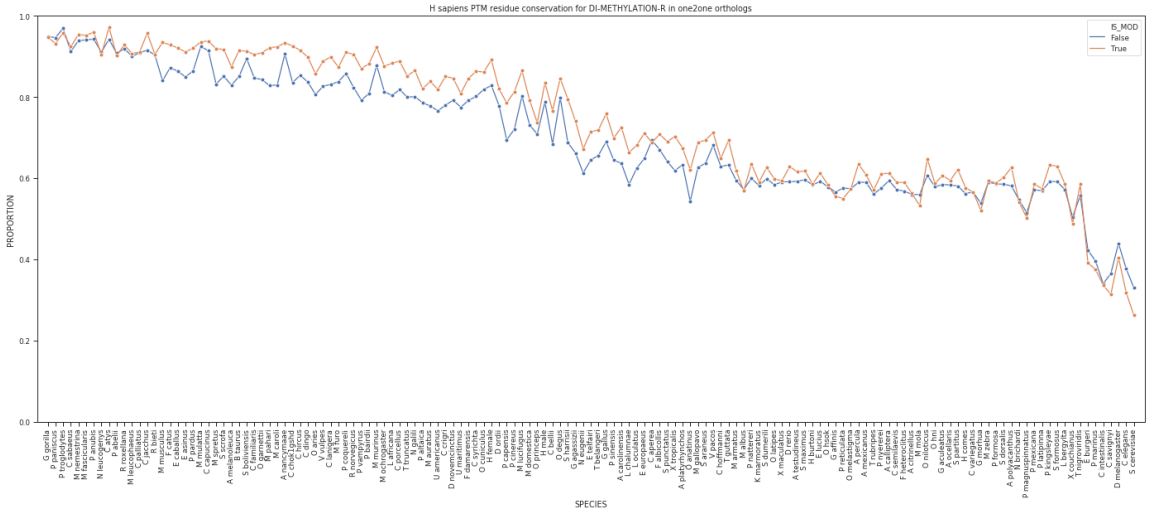
c)


H sapiens PTM residue conservation for MONO-METHYLATION-K in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

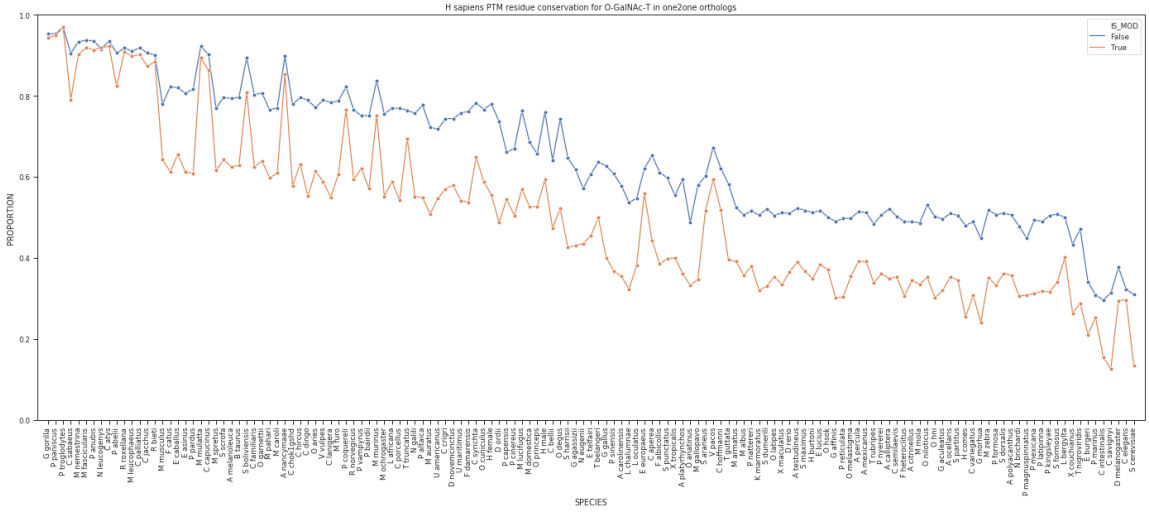
d)


H sapiens PTM residue conservation for SUCCINYLATION-K in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

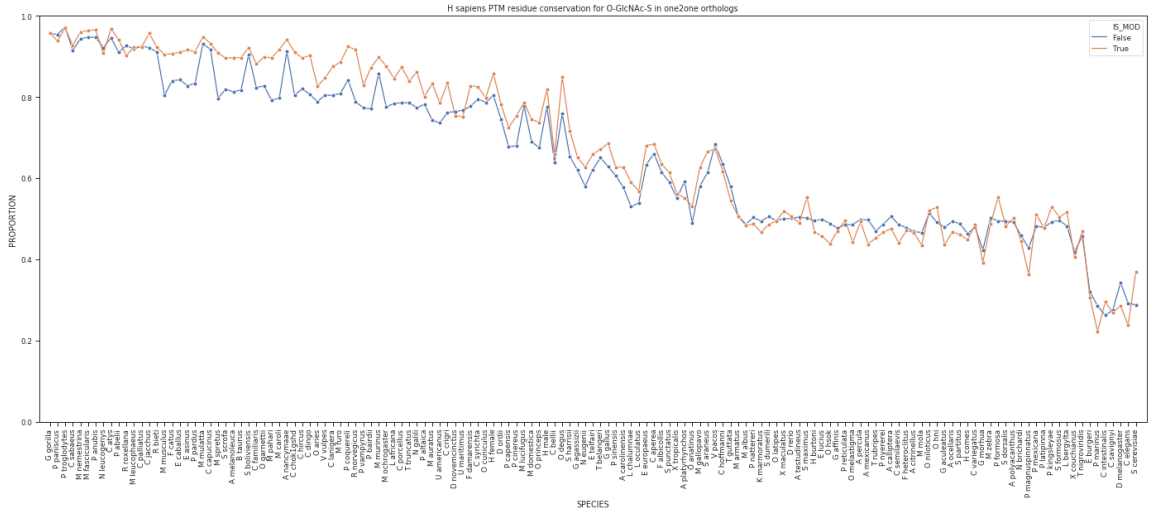
e)


H sapiens PTM residue conservation for DI-METHYLATION-R in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

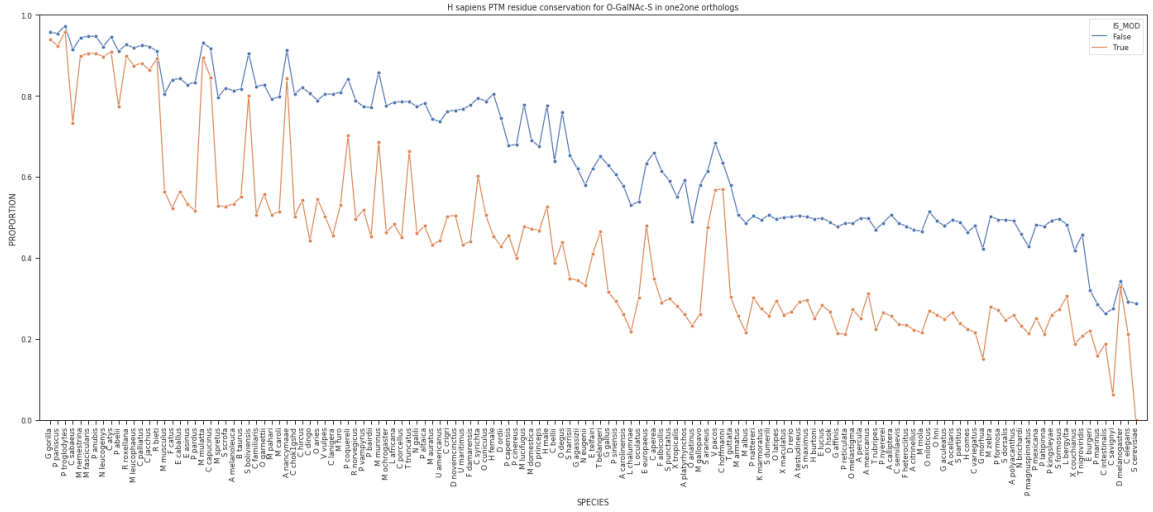
f)


H sapiens PTM residue conservation for O-GalNAc-T in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

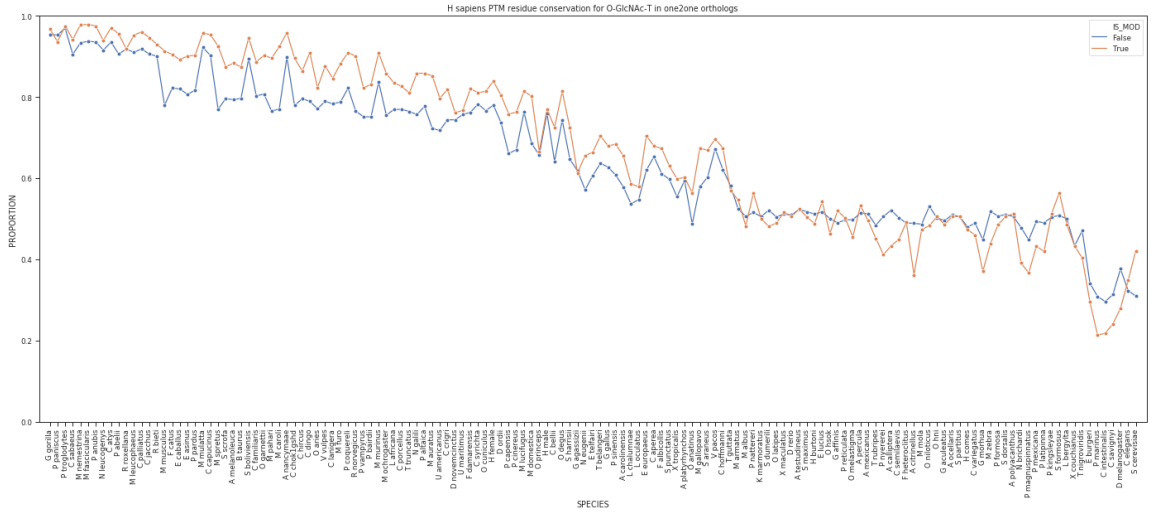
g)


H sapiens PTM residue conservation for O-GlcNAc-S in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

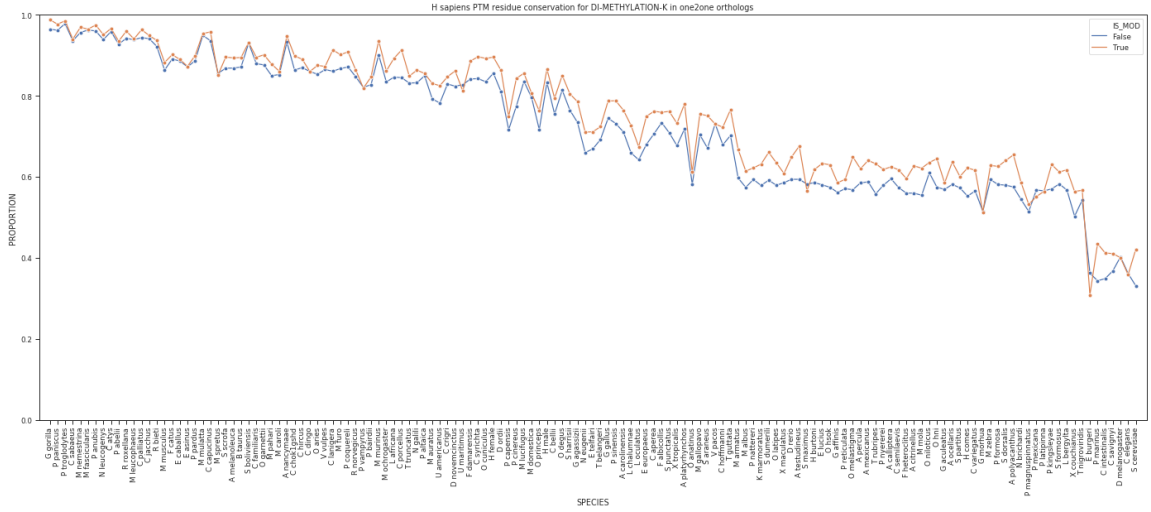
h)


H sapiens PTM residue conservation for O-GalNAc-S in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

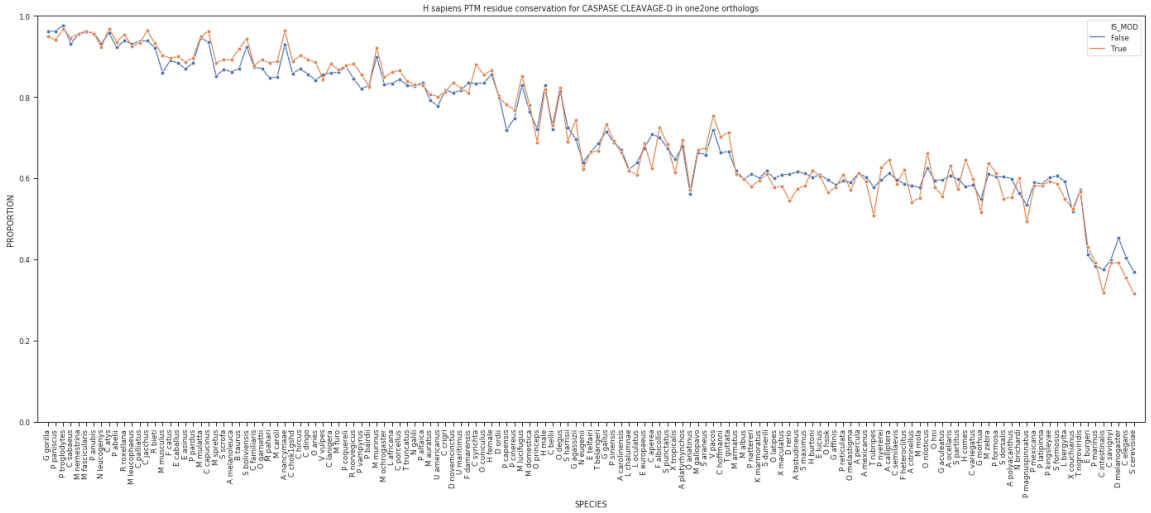
i)


H sapiens PTM residue conservation for O-GlcNAc-T in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

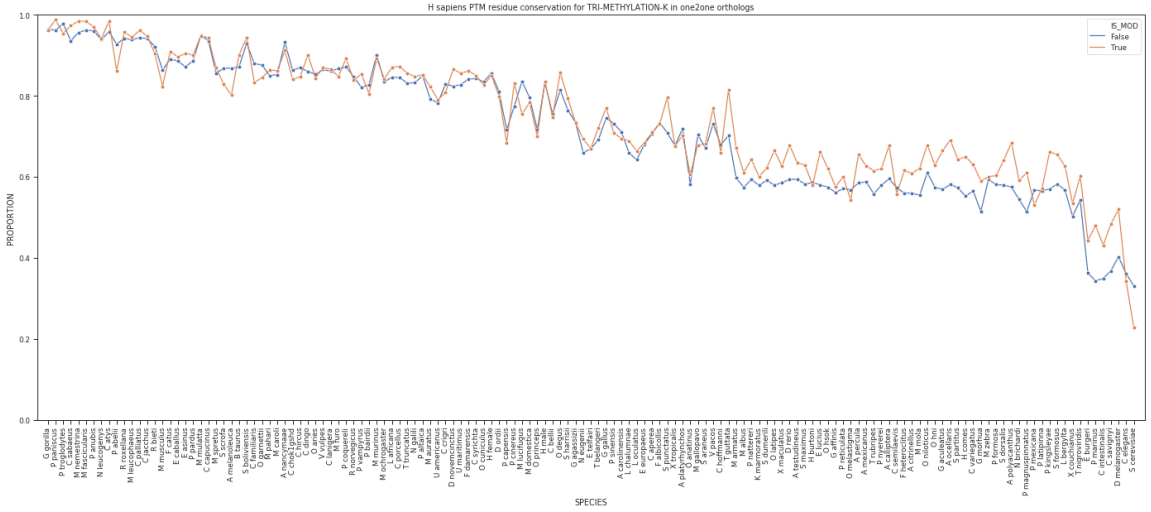
j)


H sapiens PTM residue conservation for DI-METHYLATION-K in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

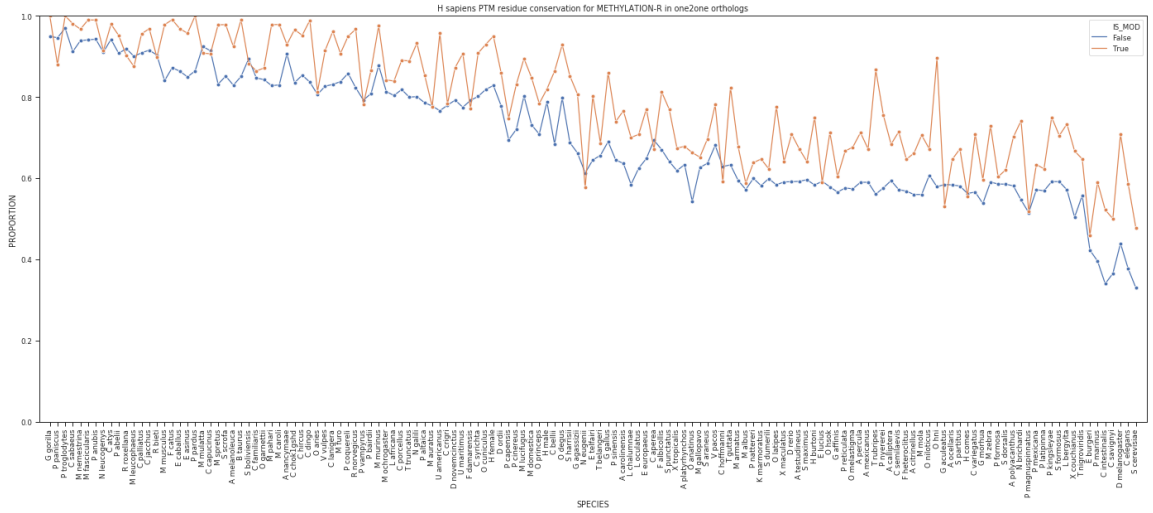
k)


H sapiens PTM residue conservation for CASPASE CLEAVAGE-D in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

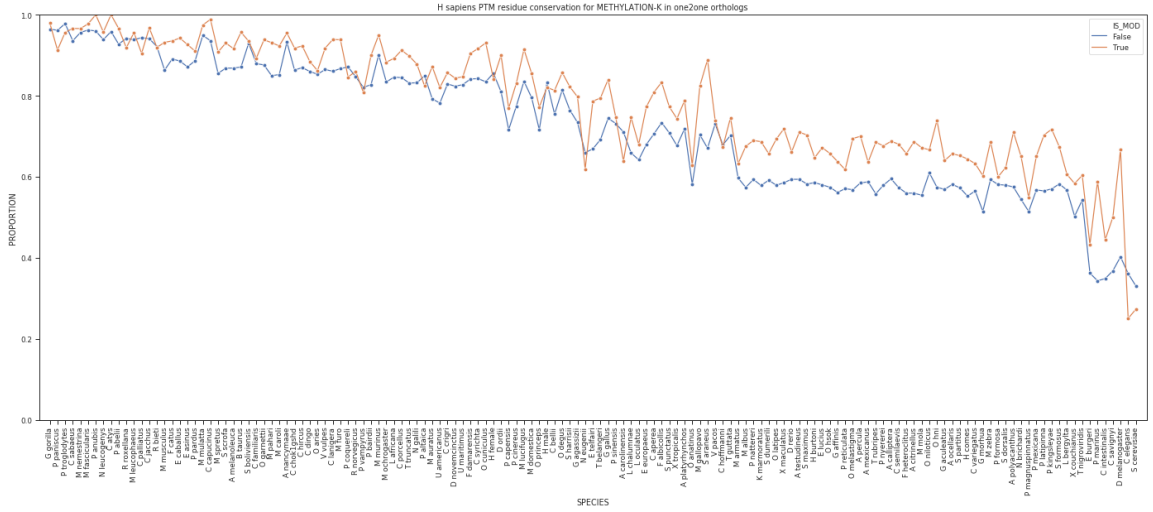
l)


H sapiens PTM residue conservation for TRI-METHYLATION-K in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

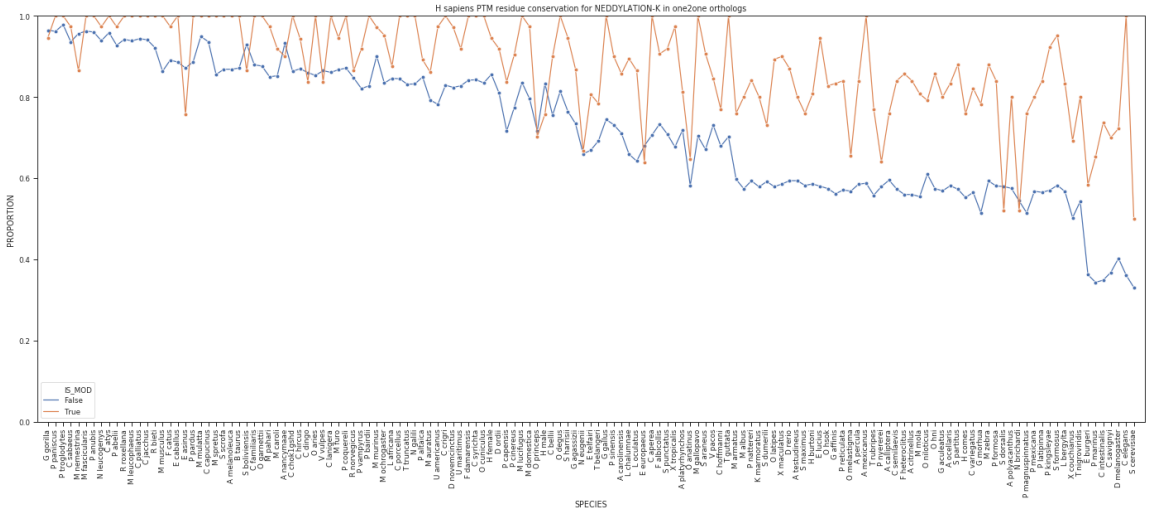
m)


H sapiens PTM residue conservation for METHYLATION-R in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

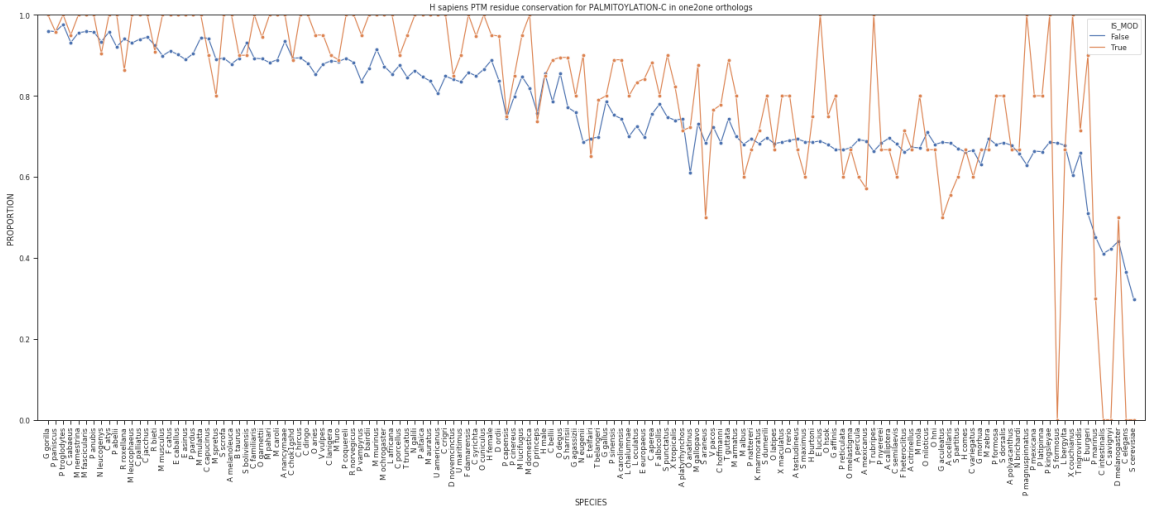
n)


H sapiens PTM residue conservation for METHYLATION-K in one2one orthologs
IS_MOD
False
True
PROPORTION
SPECIES

o)



p)



Supplementary Figure 2. Conservation of PTM sites

The proportion of modified (orange) and unmodified (blue) residues conserved for proteins with one or more R-m1 (a), K-sm (b), K-m1 (c), K-sc (d), R-m2 (e), O-GalNAc-T

(f), O-GlcNA-S (g), O-GalNAc-S (h), O-GlcNAc-T (i), K-m2 (j), D-ca (k), K-m3 (l), R-m (m), K-m (n), K-ne (o), and C-pa (p) site.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402.

Bateman, A., Martin, M. J., O'Donovan, C., Magrane, M., Alpi, E., Antunes, R., … Zhang, J. (2017). UniProt: the universal protein knowledgebase. *Nucleic Acids Research*, *45*(D1), D158–D169. https://doi.org/10.1093/nar/gkw1099

Bylund, D., & Henriksson, A. E. (2015). Proteomic approaches to identify circulating biomarkers in patients with abdominal aortic aneurysm. *American Journal of Cardiovascular Disease*, *5*(3), 140–145.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications., BLAST+: architecture and applications. *BMC Bioinformatics*, *BMC Bioinformatics*, *10*, *10*, 421, 421–421. https://doi.org/10.1186/1471-2105-10-421, 10.1186/1471-2105-10-421

celery: Distributed Task Queue (development branch) [Python]. (2018). Retrieved from https://github.com/celery/celery (Original work published 2009)

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., … de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https:// doi.org/10.1093/bioinformatics/btp163

Davies, H., Bignell, G. R., Cox, C., & Stephens, P. (2002). Mutations of the BRAF gene in human cancer. *Nature*, *417*(6892), 949–954.

Django (Version 2.0). (2018). Retrieved from https://djangoproject.com

Docker CE. Contribute to docker/docker-ce development by creating an account on GitHub [Go]. (2019). Retrieved from https://github.com/docker/docker-ce (Original work published 2017)

Dunn, O. J. (1959). Estimation of the Medians for Dependent Variables. *The Annals of Mathematical Statistics*, *30*(1), 192–197. https://doi.org/10.1214/aoms/1177706374

Gnad, F., Forner, F., Zielinska, D. F., Birney, E., Gunawardena, J., & Mann, M. (2010). Evolutionary Constraints of Phosphorylation in Eukaryotes, Prokaryotes, and Mitochondria. *Molecular & Cellular Proteomics : MCP*, *9*(12), 2642–2653. https://doi.org/10.1074/mcp.M110.001594

Gnad, F., Ren, S., Cox, J., Olsen, J. V., Macek, B., Oroshi, M., & Mann, M. (2007). PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biology*, *8*(11), R250. https://doi.org/10.1186/gb-2007-8-11-r250

Heffernan, R., Yang, Y., Paliwal, K., & Zhou, Y. (2017). Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics*, *33*(18), 2842–2849. https:// doi.org/10.1093/bioinformatics/btx218

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hornbeck, P. V., Kornhauser, J. M., Latham, V., Murray, B., Nandhikonda, V., Nord, A., … Gnad, F. (2019). 15 years of PhosphoSitePlus®: integrating post-translationally modified sites, disease variants and isoforms. *Nucleic Acids Research*, *47*(Database issue), D433–D441. https://doi.org/10.1093/nar/gky1159

Katoh, K., Misawa, K., Kuma, K., & Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066.

Khoury, G. A., Baliban, R. C., & Floudas, C. A. (2011). Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Scientific Reports*, *1*. https://doi.org/10.1038/srep00090

Kinsella, R. J., Kähäri, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., … Flicek, P. (2011). Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database: The Journal of Biological Databases and Curation*, *2011*. https:// doi.org/10.1093/database/bar030

Li, H. (n.d.). *softwares for phylogenetic trees / Code / [r304] /branches/lh3*. Retrieved from https://sourceforge.net/p/treesoft/code/HEAD/tree/branches/lh3/

Lyons, J., Dehzangi, A., Heffernan, R., Sharma, A., Paliwal, K., Sattar, A., … Yang, Y. (2014). Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *Journal of Computational Chemistry*, *35*(28), 2040–2046. https://doi.org/10.1002/jcc.23718

McKinney, W. (2013). *Python for data analysis*. Beijing: O'Reilly.

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Pearson, K. (1900). X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *F.R.S, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *50*(302), 157–175. https://doi.org/10.1080/14786440009463897

Production-Grade Container Scheduling and Management: kubernetes/kubernetes [Go]. (2019). Retrieved from https://github.com/kubernetes/kubernetes (Original work published 2014)

Reimand, J., & Bader, G. D. (2013). Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Molecular Systems Biology*, *9*, 637. https://doi.org/10.1038/msb.2012.68

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2012). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, *9*(2), 173–175. https://doi.org/10.1038/nmeth.1818

Schuster, M., & Paliwal, K. K. (1997). Bidirectional Recurrent Neural Networks. *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, *45*(11), 9.
Smith, K. D., Jewett, J. J., Montanaro, S., & Baxter, A. (2003, June 5). PEP 318 -- Decorators for Functions and Methods. Retrieved June 7, 2018, from Python.org website: https://www.python.org/dev/peps/pep-0318/

Wallace, I. M., O'Sullivan, O., Higgins, D. G., & Notredame, C. (2006). M-Coffee: combining multiple sequence alignment methods with T-Coffee., M-Coffee: combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Research*, *Nucleic Acids Research*, *34*, *34*(6, 6), 1692, 1692–1699. https://doi.org/10.1093/nar/gkl091, 10.1093/nar/gkl091

Walsh, C. T., Garneau-Tsodikova, S., & Gatto, G. J. (2005). Protein Posttranslational Modifications: The Chemistry of Proteome Diversifications. *Angewandte Chemie International Edition*, *44*(45), 7342–7372. https://doi.org/10.1002/anie.200501023

Welch, B. L. (1947). The Generalization of `Student's' Problem when Several Different Population Variances are Involved. *Biometrika*, *34*(1/2), 28. https://doi.org/10.2307/2332510

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., … Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. https://doi.org/10.1093/nar/gkx1098