BIOLOGICAL DATA PROJECT REPORT

THIOREDOXIN DOMAIN CHARACTERIZATION

SOFIIA TATOSH - 2072547, IVAN PIACERE - 2058071
Department of Engineering
Department of Mathematics
University of Padova
February 2023

1 Models building

The first thing we did was to collect homologous proteins performing a Blast search on the SwissProt database. To do that we used the NCBI BLAST+ website changing just the following parameters from the default values:

- database: UniProtKB/Swiss-Prot
- alignment-views: BLASTXML
- exp. thr.: 10
- scores: 500
- alignments: 500

We downloaded the result through

"Tool output" –> Right click "Download" –> "Save link as..." –> "blast_P07237.xml"

We then selected the proteins matched with an E-value lower than 0.001, and collected the corresponding sequences using the `https://www.ebi.ac.uk/proteins/api/proteinsAPI`.
Next we used Clustal Omega website to build a multiple sequence alignment of the homologous proteins. In this case we just specified the parameter

- output format: Pearson/FASTA.

At this point we refined the MSA using the JalView Software. We produced 10 different versions:

1. Removed redundancy above 100% threshold, recomputed alignment with ClustalO with Default parameters, then removed empty columns, and then recomputed the alignment again.
2. No refinement.
3. Removed before 1256 and after 1457, then removed redundancy above 100% threshold, and then realigned with ClustalO with Default parameters. Here the selected region covers almost exactly the Thioredoxin domain position in our initial protein.
4. Selected just the most conserved region, removing up to 1314 included, and from 1337 included on. Then removed redundancy as before and realigned.
5. Removed up to 328 included, and from 2599 included on. Then removed redundancy and realigned.
6. Removed rows which were inconsistent with the majority in the center and hence caused gaps. Then removed redundancy and realigned.
7. Removed poorly occupied columns (< 15) in the center and removed redundancy.
8. Same as 7 but realigned.
9. Removed poorly occupied columns (< 15) everywhere and removed redundancy.
10. Same as 9 but realigned.

The final step of this part consisted in building PSSM and HMM models from the MSAs. To do that we obtained the psiblast and the hmmbuild binaries. The first one was downloaded from `https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ncbi-blast-2.13.0+-x64-linux.tar.gz`, while the second one was installed with the command

```
sudo apt install hmmer
```

To build the models we used the commands

```
psiblast
-subject irrelevant.fasta
-in_msa refined_clustal_blast_P07237.fasta
-out_ascii_pssm models/P07237_ascii_pssm
                                .pssm_ascii
-out_pssm models/P07237_pssm.pssm

hmmbuild
P07237_hmm.hmm
refined_clustal_blast_P07237.fasta
```

The first command requires a "subject" sequence which we put in the file "irrelevant.fasta". Its content has to conform to the fasta format, but apart from that it doesn't affect the resulting model.

2 Models evaluation

After obtaining ten different variations of HMM and PSI-BLAST models, we obtained the predictions against SwissProt. We performed HMM-SEARCH and PSI-BLAST search with default parameters setting e-value threshold to 0.03. For HMM-SEARCH we obtained the hits and the alignments of our models to all of the hits in the form of

```
hmm_hits_{m}.tsv || hmm_ali_{m}.xml
```

And PSI-BLAST output as

```
blast_hits_{m}.csv || blast_ali_{m}.xml
```

where m indicates a model number from section 1. To replicate PSI-BLAST search, we followed these steps:

- Insert initial domain sequence into 'Enter Query Sequence' window
- In 'Choose Search Set' select SwissProt database
- 'Program Selection' Algorithm - PSI-BLAST
- Under 'Algorithm parameters' upload PSSM model, set e-value threshold at 0.03, and leave other parameters default

After creating the set of the models to evaluate (twenty in total), we defined our ground truth, which was a set of significant hits against SwissProt annotated with the given Pfam domain PF00085. The Pfam data was retrieved performing HMM-SEARCH against SwissProt directly, so as a result, we got the list of matched proteins as well as the alignments. To perform this search, we followed the steps. Go to HMM-SEARCH and select 'Accession search' → select 'PF00085' as a Pfam domain and SwissProt as a database → parameters are default → download .tsv file with matched hits.

Now, having the ground truth and models to evaluate, we performed two types of statistical analysis - at the protein level and at the residual level and then combined them together to decide which model performs best overall.

On a protein level, to gather the important information, we calculated True Positives (TP) as number of proteins occurring in Pfam and our model, False Positives (FP) as proteins predicted by our model, but were not predicted by Pfam, and False Negatives as proteins predicted by Pfam, but not catched by our model. Having these values, we then calculated precision, recall and F-score statistics.

Precision and recall are the measures that use the same concept as accuracy, but those metrics work well on imbalanced datasets.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

Precision shows the proportion of positive identifications that were actually correct. Recall depicts the proportion of actual positives that were identified correctly.

F-1 score measures test accuracy calculated from the precision and recall. The score represents a harmonic mean of precision and recall.

$$F\_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{3}$$

As for the statistics on residual level, the figure below explains the logic behind labelling TPs, TNs, etc.

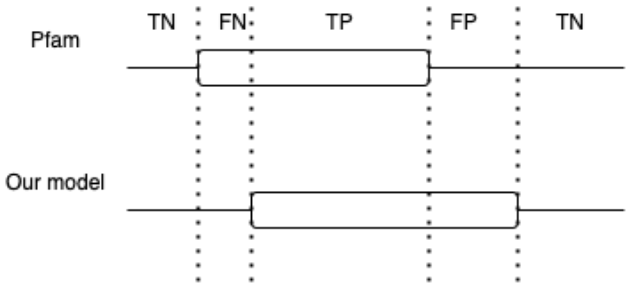

Figure 1: Labelling of TP, TN, FP and FN regions on a residual level.

As additional statistics, we added balanced accuracy and Matthew's Correlation Coefficient (MCC).

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{4}$$

$$Bal\_acc = 0.5 \cdot \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP} \right) \tag{5}$$

The detailed results of the models performance can be seen in Appendix (refer to tables 1 3 2 4). The column **Average** in tables 3 and 4 are the average value between F-score calculated on a protein level and MCC as these values contain the most information about the reliability and goodness of a model. Based on the results, we identified that the best performing model is model number 3 obtained by keeping just the range 1256-1457 in MSA.

From the models' performances we observed that: (1) Realigning reduced the performance. Probably because it changes the relative positions of many amminoacids in different sequences. (2) Selecting the specific domain region was the best form of refinement and would have probably given even better results without a realignment. (3) Removing poorly occupied columns decreased the performance. Probably also in this case the reason is a change from the original structure of the MSA, which instead is not affected by removal of columns at the start or end of the alignment. Further sections cover the analysis performed on Model 3 only.

3 Taxonomy

We wanted to display the taxonomic tree of our protein family. To do that we first extracted the UniProt names of all the proteins matched by the best model. We needed to collect data about the proteins from UniProt but to do that we needed UniProt accessions instead of names. So we used the UniProt mapping service.
After that we downladed data about proteins using the UniProt api, and then extracted the lineages from it.

At this point we wanted to build the taxonomic tree, so we iterated through the list of proteins' lineages and for each of them iterated through the taxa. During this iteration we populated a list of directed connections between the taxa, and also counted the occurrences of each taxon.
The last step was to display our tree, in a hierarchical fashion and with node sizes proportional to the cardinality of each taxon. We obtained this with the networkx and EoN (Epidemics on Network) python packages. The figure with the complete taxonomic tree can be found in the corresponding notebook and also in the image hierarchy.jpg. Since it is too large to fit here, we report below a reduced version in which we keep only nodes which include at least 5% of our proteins.

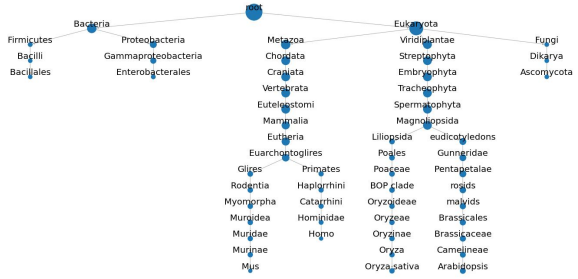

Figure 2: Hierarchical view of the taxonomic tree of our protein family, limited to nodes which contain at least 5% of our proteins

Our protein family spans across all the taxonomic domains, but is most present in the Eukaryota one, in

which it is mainly distributed in Metazoa (Animals) and Viridiplantae (Plants). In particular we can see that the paths with the most abundant presence of our family are the ones leading to mice, human, rice and Arabidopsis.

4 Function

To perform the analysis of gene ontology, we first retrieved all of the GO annotations associated with our set of proteins predicted by HMM model. To do this, we used the UniProt API to map the proteins to UniProt IDs, and parsed XML files of each protein retrieved from UniProtDB.

After we gathered GO terms related to our proteins, we calculated the terms enrichment comparing our set of proteins to the entire SwissProt DB. To get the SwissProt DB GO terms, we worked with SwissProt XML file. The next step was to create matrices of size (proteins x terms) for both our set of proteins and the rest of SwissProt proteins, where the entry $M_{i,j}$ is 1 if the i-th protein is associated with j-th GO term.

To calculate enrichment of the terms, we first performed ancestors search for our proteins, so that we encounter the higher levels of ontology as well. For this we used topological sorting and propagated through all the existing terms related to our dataset. Then, we used Fisher exact test to get the enrichment "score" of each term.

Since we are working on looking for the enrichment level of terms in our set of proteins and we want to figure out which terms are 'the most' enriched, we would like to refer to the variable that is Odds ratio. It is calculated within Fisher's exact test and it indicates the 'odds' of the protein having a certain GO term to be associated with a certain set of proteins (in our case it is 'our' proteins and 'the rest' of the human proteins). The null hypothesis is that there is no significant association between the variables (term presence and group belonging). If the odds ratio is greater than 1, it indicates that the GO term is enriched in my set, and vice versa when it is less than 1. The alternative hypothesis tested should be that the odds ratio is greater than 1, since we want to find the enriched terms in our protein set, which is exactly what is depicted by the right-sided alternative hypothesis. Usually, with p-value < 0.05 we can reject the null hypothesis and accept the hypothesis that the odds ratio is greater than 1, meaning we are obtaining an enriched term with those two conditions.

In our case, since the amount of GO terms was high, we decided to take different thresholds of p-value for each sub-ontology to properly depict them on a word cloud. The vast majority of terms were from "biological process" sub-ontology, so the threshold for "biological process" was $p < 4e{-}100$. For "molecular function" - $p < 4e{-}10$ and for "cellular component" - $p < 4e{-}20$.

The top 10 enriched terms for each sub-ontology are depicted in the below figures.

	Right-tail	Left-tail	Two-sided	GO ID	Namespace	GO name
0	0.0	1.0	0.0	GO:0008150	biological_process	biological_process
1	0.0	1.0	0.0	GO:0009987	biological_process	cellular process
2	0.0	1.0	0.0	GO:0016043	biological_process	cellular component organization
3	0.0	1.0	0.0	GO:0044085	biological_process	cellular component biogenesis
4	0.0	1.0	0.0	GO:0050789	biological_process	regulation of biological process
5	0.0	1.0	0.0	GO:0050794	biological_process	regulation of cellular process
6	0.0	1.0	0.0	GO:0065007	biological_process	biological regulation
7	0.0	1.0	0.0	GO:0071840	biological_process	cellular component organization or biogenesis
8	0.0	1.0	0.0	GO:0022607	biological_process	cellular component assembly
9	0.0	1.0	0.0	GO:0050896	biological_process	response to stimulus

Figure 3: Fisher exact test results for "biological processes" sub-ontology.

	Right-tail	Left-tail	Two-sided	GO ID	Namespace	GO name
0	0.0	1.0	0.0	GO:0003674	molecular_function	molecular_function
1	0.0	1.0	0.0	GO:0003824	molecular_function	catalytic activity
2	0.0	1.0	0.0	GO:0015035	molecular_function	protein-disulfide reductase activity
3	0.0	1.0	0.0	GO:0015036	molecular_function	disulfide oxidoreductase activity
4	0.0	1.0	0.0	GO:0016491	molecular_function	oxidoreductase activity
5	0.0	1.0	0.0	GO:0016667	molecular_function	oxidoreductase activity, acting on a sulfur gr...
6	0.0	1.0	0.0	GO:0016860	molecular_function	intramolecular oxidoreductase activity
7	0.0	1.0	0.0	GO:0016864	molecular_function	intramolecular oxidoreductase activity, transp...
8	0.0	1.0	0.0	GO:0140096	molecular_function	catalytic activity, acting on a protein
9	0.0	1.0	0.0	GO:0003756	molecular_function	protein disulfide isomerase activity

Figure 4: Fisher exact test results for "molecular function" sub-ontology.

	Right-tail	Left-tail	Two-sided	GO ID	Namespace	GO name
0	0.0	1.0	0.0	GO:0005575	cellular_component	cellular_component
1	0.0	1.0	0.0	GO:0005622	cellular_component	intracellular anatomical structure
2	0.0	1.0	0.0	GO:0012505	cellular_component	endomembrane system
3	0.0	1.0	0.0	GO:0031974	cellular_component	membrane-enclosed lumen
4	0.0	1.0	0.0	GO:0043226	cellular_component	organelle
5	0.0	1.0	0.0	GO:0043227	cellular_component	membrane-bounded organelle
6	0.0	1.0	0.0	GO:0043229	cellular_component	intracellular organelle
7	0.0	1.0	0.0	GO:0043231	cellular_component	intracellular membrane-bounded organelle
8	0.0	1.0	0.0	GO:0043233	cellular_component	organelle lumen
9	0.0	1.0	0.0	GO:0070013	cellular_component	intracellular organelle lumen

Figure 5: Fisher exact test results for "cellular component" sub-ontology.

Based on the information we obtained from Fisher exact test, we sorted our enriched terms by the number of occurrences in our set of proteins and visualized it on a word cloud. From each set of terms we removed the roots such as "biological_process", "molecular_function" and "cellular_component", since they take the majority of place on the word cloud and it is obvious that they are present in significant GO terms. The word clouds are presented below.

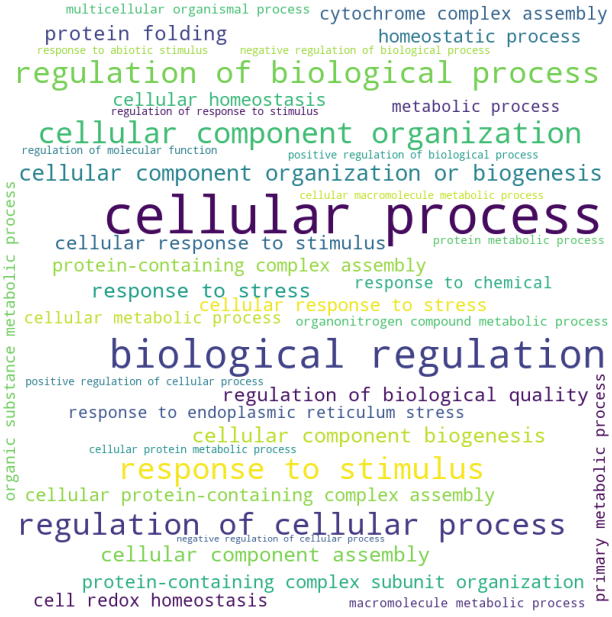

Figure 6: Word cloud for "biological processes" sub-ontology.

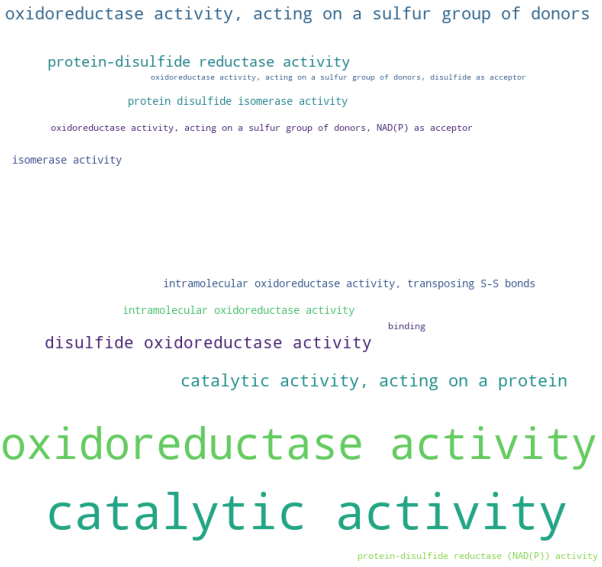

Figure 7: Word cloud for "molecular function" sub-ontology.

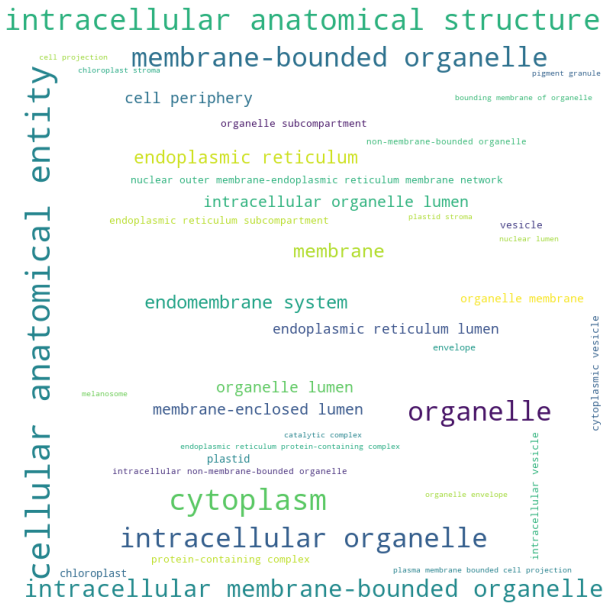

Figure 8: Word cloud for "cellular component" sub-ontology.

To analyze the most enriched GO terms' branches of our protein family, we followed three main steps. Obtain

the children of three roots (sub-ontologies) on the top layer of the tree. For each child in children perform depth first search algorithm traversing the child's children up to the leaves, and count the number of enriched terms. As soon as we reach 10+ enriched terms on a child's branch, we consider this branch an enriched one with regards to an entire sub-ontology tree. The enriched branches for biological process sub-ontology are included in Appendix (Tables 5, 6, 7)

In this part we decided to also take into consideration the biological meaning behind the data we retrieved. Thioredoxin is known to be involved in cell-to-cell communication. The primary function of Thioredoxin is the reduction of oxidized cysteine residues and the cleavage of disulfide bonds. It also acts as electron donor to peroxidases and ribonucleotide reductase. It is also known as oxireductase protein. The protein that the domain was extracted from, P07237, is acting as a reductase that cleaves disulfide bonds of proteins attached to the cell. It also catalyzes the rearrangement of -S-S- bonds in proteins (Taken from UniProtDB). What is interesting to observe is that Figures 6, 7 and 8 depict the main functions and characteristics of the family derived from this domain and it captures the ones described above from a biological standpoint. For instance, on Figure 7 we can see that the big part of the cloud is oxireductase activity as well as more specific one - disulfide oxireductase activity. On Figure 6 we can observe that cellular process plays a big role in our family, that can be pointed back to already mentioned cell-to-cell communication. By this analysis we can conclude, that our model is able to derive a significant and relevant information about the family of proteins which were initially derived from Thioredoxin domain.

5 Motifs

As in the Taxonomy step, we first obtained the names of the proteins matched by the best model, mapped them to UniProt accessions, and used the latter to download UniProt data for each of the elements. In this case we extracted protein sequences.
At this point we downloaded files with the definitions of ELM and ProSite patterns, from `http://elm.eu.org/elms/elms_index.tsv` and `https://ftp.expasy.org/databases/prosite/prosite.dat` respectively. While ELM patterns have the same syntax as python's regex, ProSite have a different one and we had to translate them.
Now we downloaded the MobiDB list of disordered regions for SwissProt.
Next we counted the occurrences of each pattern in each protein, limiting the search to the disordered regions.
In the last step we computed the z-scores of the counts according to the formula

$$z_s = \frac{N_s - E(X)}{\sigma(X)}$$

where N_s is the number of occurrences of a pattern in our protein family, $E(X)$ is the mean number of occurrences among all patterns, and $\sigma(X)$ is their standard deviation. We applied the formula to each of the patterns, but considered ELM and ProSite patterns separately, meaning that we computed their E and σ independently. Then we calculated the p-values considering the right tail of the distribution.

In tables 8 and 9 we show the significant (p-value <0.001) ELM and ProSite patterns. Using the two pattern sets we found just one significant element in common: CK2 phosphorylation site.
Many of the significant motifs (5 out of 12) are involved in phosphorylation, which is a chemical process involved in cell-signaling and metabolism in a wide variety of organisms and which operates on a big proportion of proteins.
Our domain, Thioredoxin, is known to have a characteristic CXXC motif which allows it to reduce other proteins. This motif is not included in the list of significant pat-

terns (8, 9) and that's because the datasets we used don't contain this motif. However it has 696 occurrences in our protein family, meaning it is more significantly conserved than all the found patterns.

6 Appendix

Table 1: PSI-BLAST models statistical results on the protein level.

Model	TP	FP	FN	Precision	Recall	F-score
blast_hits_2.csv	492	8	119	0.98	0.81	0.89
blast_hits_3.csv	490	10	121	0.98	0.80	0.88
blast_hits_6.csv	491	9	120	0.98	0.80	0.88
blast_hits_9.csv	492	8	119	0.98	0.81	0.89
blast_hits_5.csv	492	8	119	0.98	0.81	0.89
blast_hits_7.csv	490	10	121	0.98	0.80	0.88
blast_hits_10.csv	490	10	121	0.98	0.80	0.88
blast_hits_1.csv	491	9	120	0.98	0.80	0.88
blast_hits_8.csv	491	9	120	0.98	0.80	0.88
blast_hits_4.csv	370	4	241	0.99	0.61	0.75

Table 2: HMM models statistical results on the protein level.

Model	TP	FP	FN	Precision	Recall	F-score
hmm_hits_3.csv	594	12	22	0.98	0.96	0.97
hmm_hits_2.csv	558	13	58	0.98	0.91	0.94
hmm_hits_9.csv	542	15	74	0.97	0.88	0.92
hmm_hits_7.csv	530	11	86	0.98	0.86	0.92
hmm_hits_4.csv	467	0	149	1	0.76	0.86
hmm_hits_10.csv	535	13	81	0.98	0.87	0.92
hmm_hits_6.csv	548	21	68	0.96	0.89	0.92
hmm_hits_5.csv	538	20	78	0.96	0.87	0.92
hmm_hits_1.csv	528	24	88	0.96	0.86	0.90
hmm_hits_8.csv	522	18	94	0.97	0.85	0.90

Table 3: PSI-BLAST models statistical results on the residual level.

Model	Precision	Recall	F-score	Balanced Accuracy	MCC	Average
blast_hits_2.csv	0.88	0.80	0.84	0.90	0.79	**0.840**
blast_hits_3.csv	0.85	0.80	0.83	0.89	0.77	**0.825**
blast_hits_6.csv	0.82	0.80	0.81	0.88	0.75	**0.815**
blast_hits_9.csv	0.78	0.80	0.79	0.86	0.72	**0.805**
blast_hits_5.csv	0.78	0.80	0.79	0.85	0.71	**0.800**
blast_hits_7.csv	0.77	0.80	0.78	0.85	0.71	**0.795**
blast_hits_10.csv	0.78	0.80	0.79	0.85	0.71	**0.795**
blast_hits_1.csv	0.76	0.79	0.78	0.84	0.70	**0.790**
blast_hits_8.csv	0.77	0.80	0.78	0.85	0.70	**0.790**
blast_hits_4.csv	0.99	0.15	0.26	0.88	0.33	**0.540**

Table 4: HMM models statistical results on the residual level.

Model	Precision	Recall	F-score	Balanced Accuracy	MCC	Average
hmm_hits_3.csv	0.86	0.91	0.89	0.92	0.85	**0.910**
hmm_hits_2.csv	0.43	0.75	0.55	0.65	0.35	**0.645**
hmm_hits_9.csv	0.44	0.74	0.55	0.66	0.36	**0.640**
hmm_hits_7.csv	0.44	0.73	0.55	0.65	0.34	**0.630**
hmm_hits_4.csv	1.00	0.18	0.30	0.89	0.37	**0.615**
hmm_hits_10.csv	0.41	0.72	0.53	0.64	0.31	**0.615**
hmm_hits_6.csv	0.38	0.74	0.51	0.62	0.27	**0.595**
hmm_hits_5.csv	0.37	0.71	0.49	0.61	0.26	**0.590**
hmm_hits_1.csv	0.41	0.68	0.51	0.62	0.27	**0.585**
hmm_hits_8.csv	0.40	0.66	0.49	0.61	0.25	**0.575**

Table 5: Enriched branches in "biological process" sub-ontology.

Enriched branch ID	Enriched branch name
GO:0000003	reproduction
GO:0008152	metabolic process
GO:0009987	cellular process
GO:0022414	reproductive process
GO:0032501	multicellular organismal process
GO:0032502	developmental process
GO:0050896	response to stimulus
GO:0051179	localization
GO:0065007	biological regulation

Table 6: Enriched branches in "molecular function" sub-ontology.

Enriched branch ID	Enriched branch name
GO:0003824	catalytic activity
GO:0005488	binding

Table 7: Enriched branches in "cellular component" sub-ontology.

Enriched branch ID	Enriched branch name
GO:0032991	protein-containing complex
GO:0110165	cellular anatomical entity

Table 8: ELM significant patterns

Name	Description	Count	Z-Score	p-Value
GSK3 phosphorylation site	GSK3 phosphorylation recognition site	171	7.484382	3.59E-14
Casein kinase 1 (CK1) Phosphorylation site	CK1 phosphorylation site	141	6.112947	4.89E-10
Glycosaminoglycan attachment site	Glycosaminoglycan attachment site	129	5.564374	1.32E-08
WDR5 WD40 repeat (blade 5,6)-binding ligand	Fungi-specific variant of the WDR5-binding motif that binds to a cleft between blades 5 and 6 of the WD40 repeat domain of WDR5, opposite of the Win motif-binding site, to mediate assembly of histone modification complexes.	128	5.518659	1.71E-08
Casein kinase 2 (CK2) Phosphorylation site	Casein kinase 2 (CK2) phosphorylation site	111	4.741513	1.06E-06
USP7 binding motif	The USP7 MATH domain binding motif variant based on the MDM2 and p53 interactions.	109	4.650084	1.66E-06
WW domain ligands	The Class IV WW domain interaction motif is recognised primarily by the Pin1 phosphorylation-dependent prolyl isomerase.	99	4.192939	1.38E-05
MAPK Phosphorylation Site	Proline-Directed Kinase (e.g. MAPK) phosphorylation site in higher eukaryotes.	90	3.781508	7.79E-05
TRAF2 binding site	Major TRAF2-binding consensus motif. Members of the tumor necrosis factor receptor (TNFR) superfamily initiate intracellular signaling by recruiting the C-domain of the TNFR-associated factors (TRAFs) through their cytoplasmic tails.	80	3.324364	4.43E-04
SH3 ligand	This is the motif recognized by those SH3 domains with a non-canonical class I recognition specificity	79	3.278649	5.22E-04

Table 9: ProSite significant patterns

Prosite Id	Description	Count	Z-Score	p-Value
CK2_PHOSPHO_SITE	Casein kinase II phosphorylation site.	214	9.450105	1.69E-21
PKC_PHOSPHO_SITE	Protein kinase C phosphorylation site.	153	6.661521	1.36E-11
MYRISTYL	N-myristoylation site.	82	3.415793	3.18E-04