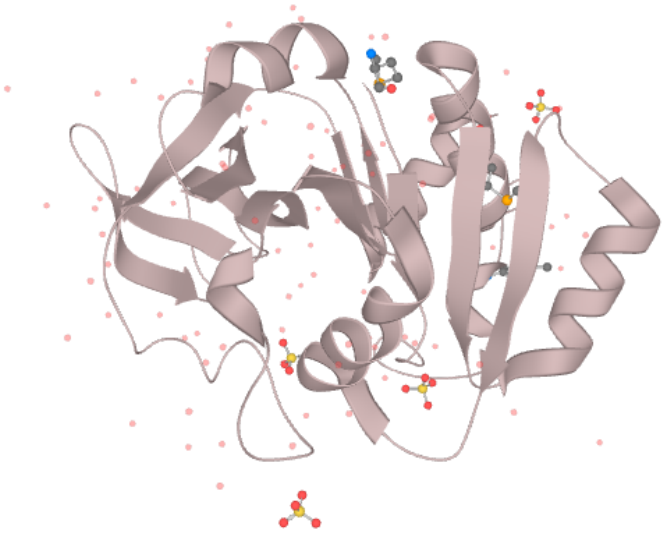
BIOLOGICAL DATA 2020/21, Data Science, University of Padova

Characterizing Nucleotidyl Transferase

Sandra Andovska Danail Krzhalovski Alfredo Petrella Marco Francesco Sommaruga



Introduction

The aim of this project is to characterize a single domain, particularly Nucleotidyl Transferase from the organism Yersinia pseudotuberculosis - a bacterial species that most commonly causes foodborne illness. We have been assigned a domain sequence from which we build a sequence model to provide structural and functional characterization of the domain family. All the results represented in this report can be reconstructed by using this code available on Github.

Domain Model Definition

To find a good model that represents our assigned domain, we developed an automated process for building HMM and PSSM models based on an input multiple sequence alignment. Prior to running this automated procedure we took two steps. Firstly, using the InterPro API we gathered all reviewed proteins belonging to SwissProt along with the domain position within them and defined them as our ground truth. Then, we performed a BLAST search against UniProt and its clusters to retrieve homologous proteins in order to perform a multiple sequence alignment to detect conserved regions. We used the MSA to build our models upon it. We chose to work with UniRef50 afterwards because we obtained a lot of significant hits ($e \leq 10^{-20}$) and the UniRef cluster allowed for easier programmatic access.

Building and evaluating the models

Once we obtained the homologous proteins, we wanted to have a broader choice for our models so we developed an easy way of building them using Python. Prior to building them, we performed MSA using T-Coffee, Muscle and ClustalOmega. The alignments produced were not very noisy but using Jalview we removed redundant sequences to gather more alignments. Having obtained sixteen alignments, we built an HMM and a PSSM model for each and we evaluated them with respect to our ground truth ([Table 1], [Table 2] in the Appendix). We realized that overall the alignments built by T-Coffee performed best, while varying redundancy thresholds do not produce better results.

The final model we chose to work with is an HMM model built upon an MSA produced by T-Coffee. We selected it since it produced the best results metrics-wise and overall HMMs are a generalization of the profile concept and take the positions of indels into account when modeling the relationship of the sequences as a "family", to represent the most probable reality.

Domain Family Characterization

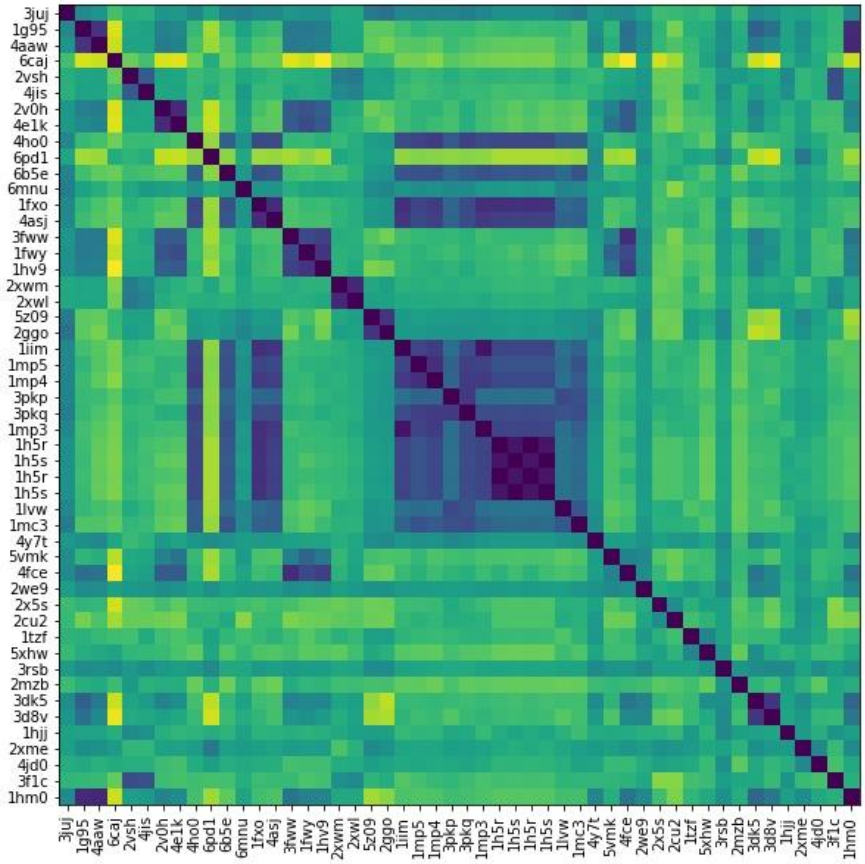
Having defined a model, we examine the structural and functional characteristics of the protein family. By identifying two datasets, one used for the structural and the other for the functional characterization, we were able to gain some insights regarding the family. The datasets constructed are:

1. **family_structures**, obtained by using the HMM created previously to search sequence databases for homologous sequences on HMMER. Only the PDB structures with a minimum overlap of 80% were kept to build the dataset.
2. **family_sequences**, which comprises all UniRef90 sequences matching the HMM. This was retrieved in .xml format.

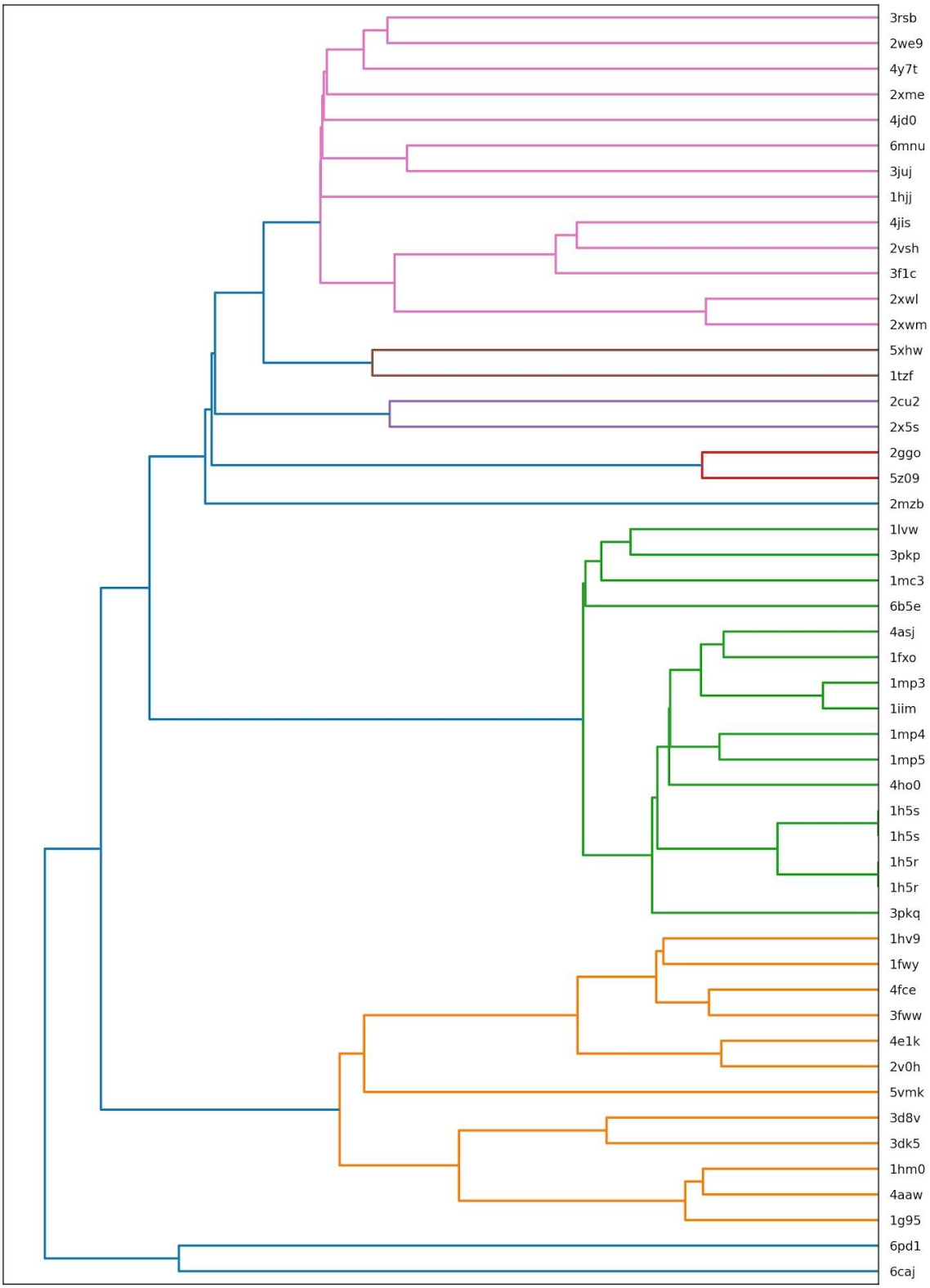
Structural Characterization

Homology modeling is based on the observation that related protein sequences adopt similar three-dimensional structure. It is the structure that determines the molecular function of the protein.

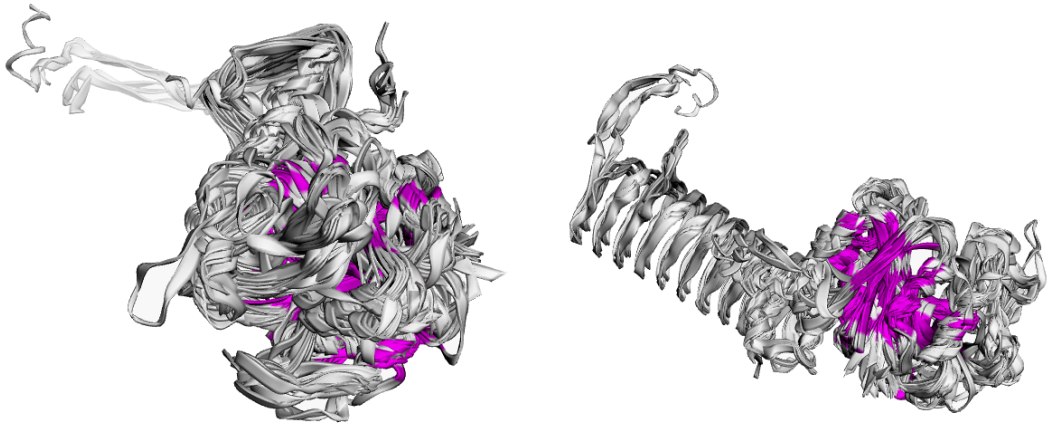
The first task after acquiring the family structures was to perform a pairwise structural alignment using TM-align, which uses a simple approach with gapless threading and secondary structure similarity. To interpret the results, we used RMSD to tell us how close two structures are. This value equals 0 for identical structures, and is very high if the two are significantly different. The results are plotted in a heatmap, purple indicating closer structures and yellow the opposite. We can clearly observe a cluster in the middle of the plot, indicating the similarity.



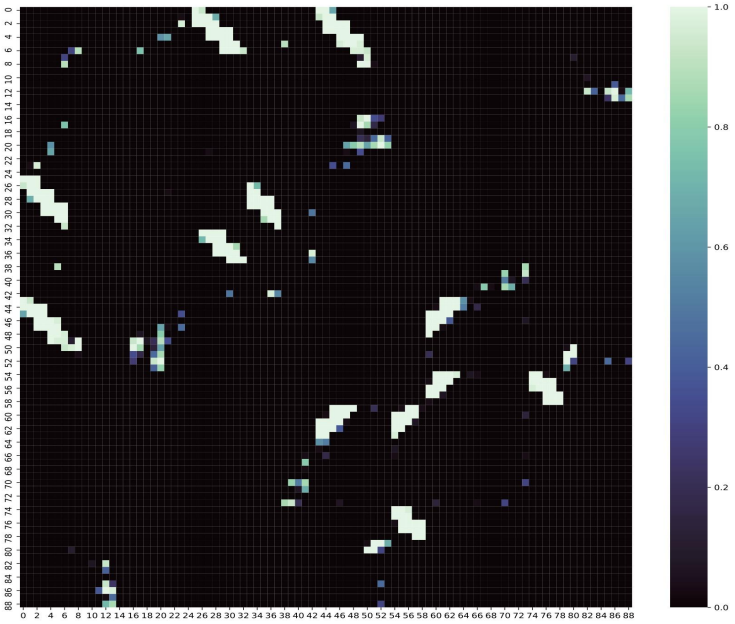
We also constructed a dendrogram to represent the hierarchical clustering of the RMSD values obtained. Analyzing it, we assume that the far down sequences are outliers that have a very high structural difference. By doing a box plot analysis, we confirmed that one of these should be removed from further investigation (6caj).


3rsb
2we9
4y7t
2xme
4jd0
6mnu
3juj
1hjj
4jis
2vsh
3f1c
2xwl
2xwm
5xhw
1tzf
2cu2
2x5s
2ggo
5z09
2mzb
1lvw
3pkp
1mc3
6b5e
4asj
1fxo
1mp3
1iim
1mp4
1mp5
4ho0
1h5s
1h5s
1h5r
1h5r
3pkq
1hv9
1fwy
4fce
3fww
4e1k
2v0h
5vmk
3d8v
3dk5
1hm0
4aaw
1g95
6pd1
6caj

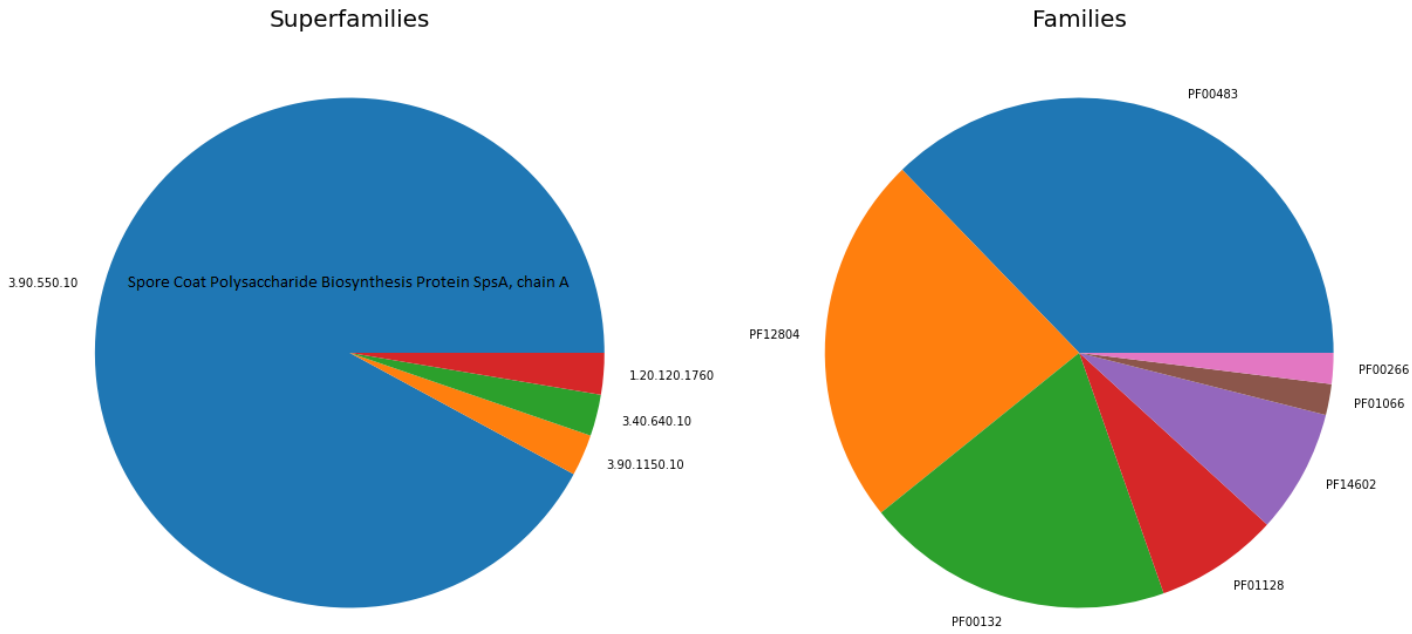
To identify the most conserved positions, we carried out a multiple structure alignment making use of mTM-align server and the PDB structures excluding outliers. We worked with the common core alignment (highlighted in magenta colour in the following figure) to spot long range conserved contacts.



Essentially, we built a distance matrix for every domain structure by setting a threshold of 12 sequence positions for the separation. Likewise, we produced a contact map examining only the areas separated no more than 8Å.



Finally, by using the PDB IDs from the family structures in the "Retrieve/ID mapping" service from UniProt and cross-referencing Gene3D and Pfam, we derived the CATH superfamilies as well as the Pfam families concerning the proteins we worked with. As a result, four superfamilies and seven families were identified. The most dominant superfamily is 3.90.550.10 and each of the structures was identified with it, whereas there are three families with a significant influence: PF00483 - which is the family of the input sequence assigned (Y. pseudotuberculosis); PF12804; and PF00132.



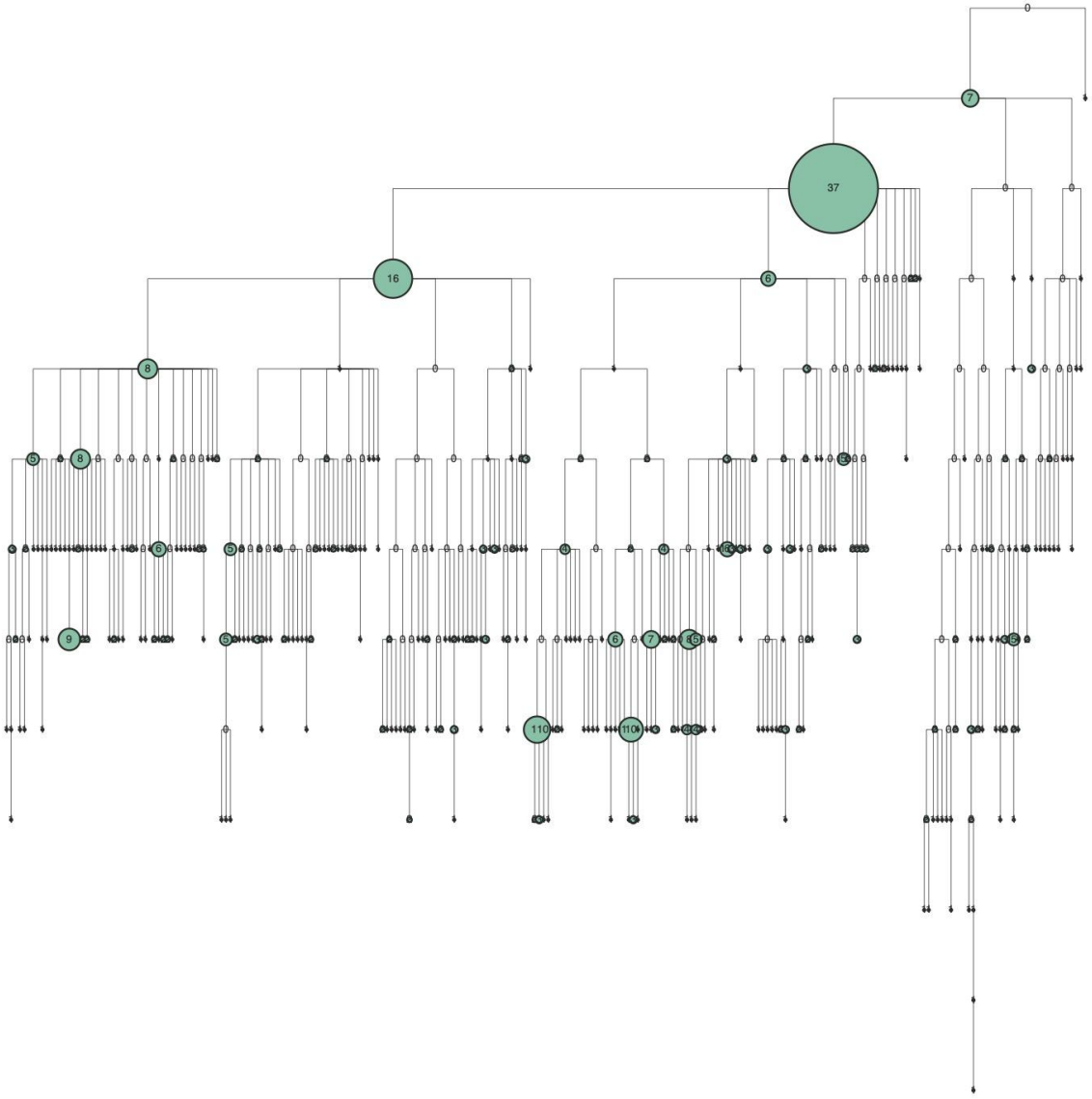
Taxonomy

In this section, we address the evolutionary relationships among sequences. We make use of the family_sequences database, comprising 664 clusters containing a total of 1179 protein sequences.

Each cluster is characterized by a taxonomic ID (i.e. the cluster identification), a representative member ID (that is the ID of the representative member of the cluster) and a taxonomic ID for each protein contained in the cluster.

In order to find a representative tree branch, we decided to consider the cluster ID and download the NCBI taxonomy database using the ete3 library. Once we found the lineage of all the proteins, we counted the abundance of each ID in the lineage and set the dimension of the respective node accordingly.

Finally, we were able to find the taxonomy tree of our family sequences. In particular, we chose to exploit the toytree library. There are 242 leaves and all the other nodes are internal. The most common nodes can be seen in [Table 3] in the Appendix. The taxonomy tree is shown in the following figure:



Functional Characterization

In order to provide a functional characterization of our domain, we finally analyzed the GO annotations of each retrieved sequence. For clarification, Gene Ontology is a structured, controlled vocabulary for the classification of gene function at the molecular and cellular level, divided in three separate sub-ontologies or GO types: biological process, molecular function and cellular component. The needed information was already contained in the family sequences dataset, so we simply extracted the annotations (entity/dbReference) as explained for the taxonomy lineage. For the same purposes, we needed the annotations from the SwissProt dataset.

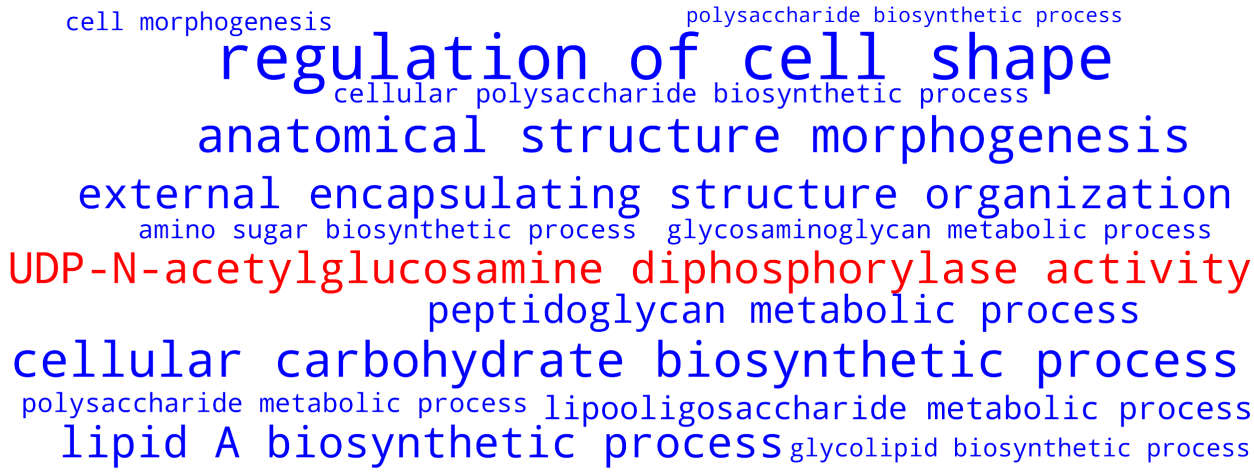
It is important at this point to recall that, according to the true path rule, a gene annotated to a term is also implicitly annotated to each ancestor of that term in the GO graph, so the following statistics will be performed, for each sequence, on the set of all the annotations relative to the sequence itself and to its ancestors, in order to obtain more reliable results.

The next step was to perform a functional enrichment analysis: it consists of applying statistical tests to verify if genes of interest are more often associated with certain biological functions than what would be expected in a random set of genes. We chose Fisher's exact test as a reference, given that, especially in the case of strongly unbalanced datasets like ours, it usually outperforms the common alternatives such as fold increase, computed aside and clearly greater than 1, even if not in agreement with the F-test about the order of the first most significant annotations. The results are shown below, in an absolute scale and grouped by sub-ontology.

Note that, given both the huge dimension of the entire SwissProt with respect to the number our matching sequences and the good performances of our model, the resulting right-tailed p-values are extremely small, indicating that our model performs very well and that the found terms are strongly likely to be correct.

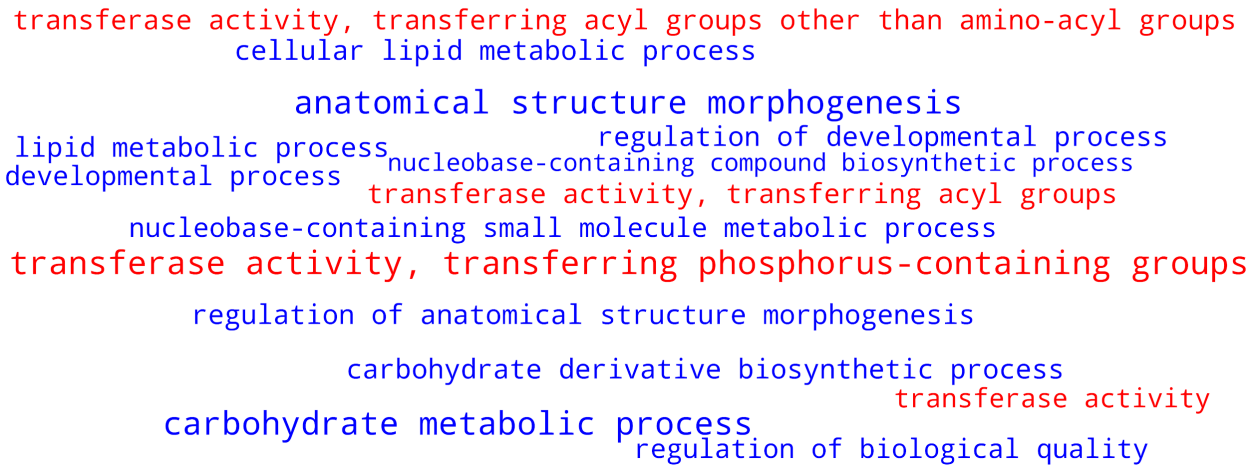
GOi	l_p	r_p	two_p	fold_inc	n_children	def
GO:0016772	1.0	0.000000e+00	0.000004	14.005525	597.0	transferase activity, transferring phosphorus-...
GO:0005975	1.0	0.000000e+00	0.000003	14.843562	539.0	carbohydrate metabolic process
GO:0009653	1.0	0.000000e+00	0.000002	37.104949	704.0	anatomical structure morphogenesis
GO:0016747	1.0	1.362233e-316	0.000002	27.466950	412.0	transferase activity, transferring acyl groups...
GO:0022603	1.0	5.894371e-309	0.000002	26.136851	393.0	regulation of anatomical structure morphogenesis
GO:0016746	1.0	8.738476e-294	0.000001	23.247296	458.0	transferase activity, transferring acyl groups
GO:0050793	1.0	5.754133e-240	0.000002	15.653365	1713.0	regulation of developmental process
GO:0044255	1.0	4.762986e-232	0.000002	12.966483	557.0	cellular lipid metabolic process
GO:0006629	1.0	1.883860e-217	0.000003	11.679714	672.0	lipid metabolic process
GO:1901137	1.0	1.589272e-199	0.000003	8.951599	319.0	carbohydrate derivative biosynthetic process
GO:0055086	1.0	4.438470e-194	0.000003	8.558009	642.0	nucleobase-containing small molecule metabolic...
GO:0065008	1.0	7.156973e-184	0.000002	10.133582	1506.0	regulation of biological quality
GO:0032502	1.0	1.637212e-178	0.000003	9.637514	3212.0	developmental process
GO:0016740	1.0	2.402559e-175	0.000006	5.134042	2478.0	transferase activity
GO:0034654	1.0	1.882027e-171	0.000004	7.784914	457.0	nucleobase-containing compound biosynthetic pr...

The word cloud in the figure below contains the 15 most enriched GO terms descriptions, independently from their sub-ontology, which is nevertheless represented by their score-proportionally vanishing colours: in red the children of the molecular_function node and in blue the ones under the biological_process class.



It is easy to note that no green term shows up, which would be relative to the cellular_component branch, for which the lower p-value is lower with respect to the other terms, but still significant in the first cases. Moreover, it is interesting that only one molecular_function term shows up, revealing that with high probability (p-value < $1e^{308}$) our domain can be annotated with depth of 6 *UDP-N-acetylglucosamine diphosphorylase activity* term, GO:0003977, and so with all its ancestors.

Note that, with the p-value threshold alone, it may be hard to tell which are the most interesting hits, so a condition about the number of children of each term was added, in order to prefer, among the top-ranked results, the terms which are higher up in the main branches, as an assurance of the stability of the results. In this case the resulting word cloud is slightly more general and shows higher level terms of the two previously mentioned sub-ontologies, and in general most of them have as children the nodes in the previous word cloud, confirming we are on the right path.



Finally, the most significant enriched branches are shown below, coherently with the previous results.

GOi	Sub-Ontology	def	depth
GO:0008360	biological_process	regulation of cell shape	3
GO:0009653	biological_process	anatomical structure morphogenesis	2
GO:0000902	biological_process	cell morphogenesis	3
GO:0071555	biological_process	cell wall organization	3
GO:0016772	cellular_component	transferase activity, transferring phosphorus-...	3
GO:0044262	biological_process	cellular carbohydrate metabolic process	3
GO:0005975	biological_process	carbohydrate metabolic process	3
GO:0071554	biological_process	cell wall organization or biogenesis	2
GO:0016746	cellular_component	transferase activity, transferring acyl groups	3
GO:0050793	biological_process	regulation of developmental process	3
GO:0044255	biological_process	cellular lipid metabolic process	3
GO:0006629	biological_process	lipid metabolic process	3
GO:0055086	biological_process	nucleobase-containing small molecule metabolic...	3
GO:0065008	biological_process	regulation of biological quality	2

Summary

To summarize, using an automated process and by exploiting some of the available Multiple Sequence Alignment tools like T-Coffee, ClustalOmega and Muscle we developed a model that represents our domain precisely considering the important metrics. We then constructed the family structures and sequences databases to be able to provide additional insights into the structural and functional characteristics of our domain. After performing pairwise and multiple sequence structural alignments, we obtained the CATH superfamily 3.90.550.10 and several families of which the most dominant was our starting sequence's family - PF00483. For the functional part, we found out that it only covers two of the three sub-ontologies namely the biological processes (biosynthetic and metabolic processes) and molecular function (transferase activity).

Appendix

Model Names: /models/model_type/UR50_Alignment-Software_Redundancy-Threhsold

1. Sequence Metrics

accuracy	precision	sensitivity	specificity	MMC	f1_score	model_name
0.998423	0.391760	0.965577	0.998456	0.614588	0.557377	/models/hmm/UR50_CO_R96
0.998338	0.380667	0.982788	0.998354	0.611158	0.548775	/models/pssm/UR50_CO_R96
0.995792	0.192281	0.969019	0.995820	0.430801	0.320889	/models/hmm/UR50_CO_R98
0.998320	0.377333	0.974182	0.998345	0.605810	0.543969	/models/pssm/UR50_CO_R98
0.996165	0.207290	0.969019	0.996193	0.447377	0.341523	/models/hmm/msa_modified_clustalomega
0.998341	0.381333	0.984509	0.998356	0.612227	0.549736	/models/pssm/msa_modified_clustalomega
0.998412	0.390278	0.967298	0.998444	0.613966	0.556160	/models/hmm/UR50_MU_R92
0.998341	0.381333	0.984509	0.998356	0.612227	0.549736	/models/pssm/UR50_MU_R92
0.998483	0.399563	0.944923	0.998538	0.614042	0.561637	/models/hmm/UR50_MU_R95
0.998310	0.375333	0.969019	0.998340	0.602601	0.541086	/models/pssm/UR50_MU_R95
0.998904	0.484007	0.989673	0.998913	0.691732	0.650085	/models/hmm/UR50_TC_R96
0.998324	0.378000	0.975904	0.998347	0.606879	0.544930	/models/pssm/UR50_TC_R96
0.998912	0.486005	0.986231	0.998926	0.691958	0.651136	/models/hmm/UR50_TC
0.998310	0.375333	0.969019	0.998340	0.602601	0.541086	/models/pssm/UR50_TC
0.998736	0.447012	0.965577	0.998770	0.656594	0.611111	/models/hmm/UR50_MU_R93
0.998352	0.383333	0.989673	0.998361	0.615436	0.552619	/models/pssm/UR50_MU_R93

2. Position Metrics

precision	sensitivity	specificity	MMC	f1_score	model_name
0.634587	0.780827	0.999623	0.703797	0.700152	/models/hmm/UR50_CO_R96
0.698593	0.996851	0.999703	0.834378	0.821488	/models/pssm/UR50_CO_R96
0.615631	0.710022	0.999586	0.661030	0.659466	/models/hmm/UR50_CO_R98
0.698619	0.997639	0.999706	0.834724	0.821773	/models/pssm/UR50_CO_R98
0.614502	0.525077	0.999593	0.567959	0.566281	/models/hmm/msa_modified_clustalomega
0.704494	0.996303	0.999711	0.837667	0.825365	/models/pssm/msa_modified_clustalomega
0.626689	0.397448	0.999669	0.499033	0.486412	/models/hmm/UR50_MU_R92
0.708340	0.997348	0.999721	0.840395	0.828359	/models/pssm/UR50_MU_R92
0.600556	0.395862	0.999645	0.487562	0.477184	/models/hmm/UR50_MU_R95
0.937367	0.994358	0.999956	0.965420	0.965022	/models/pssm/UR50_MU_R95
0.649828	0.882171	0.999627	0.757001	0.748381	/models/hmm/UR50_TC_R96
0.725858	0.994664	0.999742	0.849587	0.839262	/models/pssm/UR50_TC_R96
0.808590	0.871012	0.999837	0.839130	0.838641	/models/hmm/UR50_TC
0.983009	0.994448	0.999989	0.988705	0.988695	/models/pssm/UR50_TC
0.615642	0.399782	0.999650	0.496073	0.484768	/models/hmm/UR50_MU_R93
0.933524	0.993567	0.999952	0.963053	0.962610	/models/pssm/UR50_MU_R93

3. Nodes' Abundance

Organism	Abundance
Bacteria	37
Proteobacteria	16
Clostridium	11
Streptococcus	10
Shewanella	9
Gammaproteobacteria	8
Mycobacteriaceae	8
Pasteurellaceae	8
Bacillaceae	7
cellularorganisms	7
Streptomyces	6
Lactobacillaceae	6
Terrabacteriagroup	6
Vibrionaceae	6