Functional domain annotation by structural similarity

Poorya Mirzavand Borujeni[1] and Reza Salavati[1,2,*]
[1] Institute of Parasitology, McGill University
[2] Department of Biochemistry, McGill University

Abstract

The most widely used tools for electronic functional annotation rely on domain annotation. Pfam, one of the most used domain annotation tools, has a list of well-known protein domain sequences called seed sequences from which a domain's sequence signature is drawn and is used for annotating new domains in other proteins. From a biological point of view, structures are more conserved than sequences. Therefore, we expect to annotate more domains in phylogenetically remote proteins if we consider the structural similarity. However, it was not possible on a large scale until recently, as the structure of a limited number of proteins had been resolved experimentally, and computational approaches were not sufficiently accurate. Besides, there were no efficient tools for finding structurally similar proteins quickly.

In this study, we took advantage of two state-of-the-art programs, Alphafold2, the pioneer program for predicting protein structures with an accuracy comparable to experiments, and Foldseek, the ultra-fast program for finding structurally similar proteins, to annotate new Pfam domains in the proteome of *Trypanosoma brucei* based on structural similarity to the Pfam Seed structures. We did this by training a model for estimating the confidence score based on alignment characteristics and greedily selecting the minimally overlapping most confident Pfam predictions. As a result, we predicted over 1100 new Pfam domains in *T. brucei*, which increases the number of annotated Pfam domains in *T. brucei* up to 20%. The scripts used in this work are freely available at https://github.com/Pooryamb/FDASS.

Keywords: Domain, Pfam, Structure, functional similarity

[*] To whom correspondence should be addressed.

Introduction

In functional annotation, the endeavor is to describe the identity of a protein in three different aspects: its final location, its function, and the process in which it is involved. Experimental functional annotation is time-consuming and costly, which explains why electronic functional annotations that do not need a manual inspection are highly important. Currently, most electronic annotation tools rely on the sequence similarity to annotated proteins. Annotation based on 1:1 orthology and InterPro2GO are the two most widely used electronic annotation tools(*Gene Ontology Consortium Website*, n.d.). In annotation based on 1:1 orthology, sequence search tools like blast are mainly used for inferring the orthology relationship. InterPro2GO, the most widely used electronic annotation tool, looks for different sequence signatures in the query protein by InterProScan (Paysan-Lafosse et al., 2022). Some identified sequence signatures will later be translated into GO terms using the InterPro2GO table, a curated table showing the associations between sequence signatures and the functional annotation of the proteins. Protein structure is three to ten times more conserved than its sequence (Illergård et al., 2009). Therefore, we expect to annotate more proteins if we use protein structure in the annotation process. Some studies have successfully used protein structures to annotate the domains of protein that couldn't be annotated by sequence alone (Bartas et al., 2022; Li et al., 2018; Zarembinski et al., 1998). However, these approaches have not been widely adopted because it is difficult to experimentally determine the structure of a protein with an unknown function. Additionally, the accuracy of structure prediction tools in the past has been limited, making it difficult to predict the structure of an unannotated protein. Despite their inaccuracy, they have been shown to help in functional annotation (Gligorijević et al., 2021; C. Zhang et al., 2017).

Structure-based functional annotation hasn't been the center of attention for large-size annotations because the structure of a limited number of proteins had been resolved experimentally, and computational predictions were not accurate enough. However, with the introduction of Alphafold2, the structure of proteins could be predicted with an accuracy comparable to experimental approaches(Jumper et al., 2021; Varadi et al., 2022).

By having the structures predicted, we need to find structurally similar proteins for functional annotation. Even though sophisticated tools like DALI(Holm, 2022) and TM-align(Y. Zhang & Skolnick, 2005) are available, these tools can be computationally intensive when searching against comprehensive structure databases. RUPEE (Ayoub & Lee, 2019) is a fast tool, yet it only relies on structural similarity and does not consider sequence similarity. Foldseek, the newly developed structure search tool, can find

structurally similar proteins orders of magnitude faster than other tools (van Kempen et al., 2022). The combination of Alphafold2 and Foldseek could help us annotate more proteins. For instance, when comparing the proteomes of two organisms, in some cases, the potential homolog of a gene can be identified by the best reciprocal structural hits to infer the function of a protein when sequence-based best reciprocal hits fail (Monzon et al., 2022).

As discussed above, in sequence-based domain annotation, Interpro2GO relies on InterPro, which includes multiple member databases, each of which stores a specific kind of sequence signature. For instance, Pfam, one of the most popular members, stores the sequence signature of functional domains of proteins (Mistry et al., 2021). The sequence signatures are drawn from multiple sequence alignments of well-known domains. Pfam uses Hidden Markov Model (HMM) for modeling the signature. Later, proteins of interest can be queried against the signatures by hmmscan (Potter et al., 2018). There are some tools for improving the sensitivity of Pfam annotation. For instance, more hits can be found by making a profile of query proteins and aligning against Pfam profiles, yet the hit orders might not reflect their relevance (Söding et al., 2005). A recent study found that by using Convolutional Neural Networks (CNNs) for storing the sequence signature, over 9.5% more new domains would be annotated (Bileschi et al., 2022), referred to as Pfam-N annotations.

This study aimed to annotate the functional domains by finding Pfam Seeds structurally similar to the *T. brucei* proteins. Even though there are some general rules for inferring functional similarity from sequence similarity to an annotated protein (Pearson, 2013), to our knowledge, there is no equivalent for inferring functional similarity from structural similarity. TM scores above 0.5 show a considerable chance of sharing the same folds (Xu & Zhang, 2010), yet structurally similar proteins might have different functions (Nagano et al., 2002). In this regard, a model has been trained for estimating the confidence score of having the same Pfam based on structural alignment characteristics to a Pfam seed. Next, the most confident minimally overlapping Pfams are predicted for *T. brucei* proteome. Interestingly, even though the training has been done in *T. brucei*, the trained model has low dependency on the training set. It could predict the correct Pfam even in phylogenetically distant organisms from three domains of life. As a case study, we looked deeper into the predicted Pfams for proteins involved in the mitochondrial RNA (mtRNA) editing process in *T. brucei* by manually checking the predictions.

Materials and Methods

Making a database of chopped domains

The Full Length (FL) structure of the Pfam Seeds (version 35.0) was retrieved from the Alphafold database (version 3). Alphafold and Pfam use different versions of sequences. If the protein sequence lengths differed in the two databases, we crossed them from the list. We cut the FL structure of proteins used in Pfam Seeds on their Pfam borders and stored the chopped domains in PDB format. For this task, we modified the source codes of the PDB-tools package (Rodrigues et al., 2018) to do chopping and format conversion to PDB simultaneously. Foldseek describes the state of each amino acid in a 3D structure with a 3Di state, and the state of each amino acid depends on its interactions with the rest of the amino acids of the tertiary structure. Foldseek converts target and query databases to a 3Di sequence and uses 3Di + amino acid sequence alignment for prefiltering the target database. Due to the dependency of the state of each amino acid on its interactors, in some cases, the state of the amino acid had a change in the Chopped structure compared to the FL structure. For correction, we also made a database of the FL proteins of Pfam Seeds and cut the 3Di sequences of the FL proteins on their Pfam borders. 3Di sequences of the chopped database were substituted with the chopped 3Di sequences. The Pfam Seed structure database contained over 1.1 million structures.

Structure alignment

We queried the structure of *T. brucei* proteome against the structure of Pfam Seeds. To report the Pfams whose significant portion is aligned with the target protein, we ran the Foldseek with the "--cov-mode 1 -c 0.8" option to just report Pfams whose more than 80 percent are aligned with the target structure. The number of sequences passing the pre-filtering step was set to a high number by the "--max-seqs 1e9" option to get all possible alignments.

Labeling the alignments

We labeled the alignment of the Pfam Seed structures to the *T. brucei* proteins based on the Pfam Annotation of the part they aligned with. If the aligned part on the query covered more than 50 percent of a Pfam equal to the structurally aligned target Pfam, it was labeled as 1. If another Pfam covered more than 25 percent of the query alignment, it was labeled 0. It was labeled as -1 otherwise. 60, 20, and 20 percent of the data labeled 0 or 1 were used for training, cross-validation, and testing, respectively. We used Seaborn to visualize the relationship between the alignment characteristics and the alignment label (Waskom, 2021).

Training and cross-validation

We used sklearn (Pedregosa et al., 2011) to train a Random Forest (RF) (Ho, 1995) to predict the alignment label based on the fraction of identical amino acids (fident), alignment length(alnlen), the normalized number of mismatches (nmismatch) calculated as "(the number of mismatches)/alnlen * 100", normalized number of gap openings (ngapopen) calculated as" (the number of gap openings)/alnlen * 100", Pfam instance length (pflen), bit score, TM-score, lddt, -log of e-value(p_evalue), and the fraction of the Pfam instance covered in the alignment (pfcov). The number of trees in the model was set to 20, 100, 500, or 1000. The max depth was set to 5,6,7,8,9, or unlimited. Gini, log_loss, and entropy criteria were examined for splitting at each node. The best value of the hyperparameters was inferred from the cross-validation set based on the area under the receiver operating characteristic (ROC) curve.

To see the dependence of the trained model on the training size, we selected a Down Sample Training (DST) by training a model on 0.01 of the entire training set. Yet, it was tested on the former test set.

To see generalizability to other organisms, the model trained on *T. brucei* (using the whole training set) was tested on three phylogenetically distant organisms from the domain of life, *Saccharomyces cerevisiae,* a Eukaryotic model organism, *Escherichia coli,* a Prokaryotic model organism, and *Methanocaldococcus jannaschii*, an Archaean model organism.

Pfam Prediction

The trained model was used to predict the label greedily, and the probability score for the proteins whose aligned part against a Pfam seed didn't have a Pfam annotation. Then, for each protein, the alignments of the Pfam instances against the target proteome were sorted by the probability score in descending order.

Starting with the most probable prediction, the Pfam of a query was attributed to a protein if it overlapped less than 10 amino acids with each of the previously attributed domains. For benchmarking the performance of "best hit attribution," we started with an empty list of Pfam annotations for each protein. However, for reporting the new Pfam domains, we started with the domains already annotated by Pfam.

Sequence-based Pfam annotation

Even though Pfam predicts domains based on sequence, it models the sequence signature with HMMs that do not have an equivalent for structure searching at the time of writing this paper. To investigate

the added value of structural similarity for the trained model performance, we did the same procedure as explained above by searching the sequences of *T. brucei* proteome against Pfam instances by Mmseqs2 (Steinegger & Söding, 2017), the same program used by Foldseek under the hood. We specified the same sensitivity setting used for Foldseek by the "-s 9.5" parameter. The label of the alignments was imbalanced, and we used SMOT to balance the training set (Chawla et al., 2002) before training.

Critical residue conservation

As a case study, we focused on the protein involved in mtRNA editing; refer to (Aphasizheva et al., 2020) for the summary. Alphafold uses the Uniprot ID of proteins, while the list of proteins is based on the locus ID found in TriTrypDB (Amos et al., 2022), the organism-specific database. To map the IDs, the proteome of the *T. brucei* in TriTrypDB was blasted against the *T. brucei* in UniProt. The closest hit for each protein with more than 99 percent identity was kept.

In each Pfam signature, we looked for conserved positions. After locating the conserved site of the Pfam profile on the Pfam instance, we identified its corresponding position on the query protein based on FoldSeek structural alignment.

GO term prediction

GO terms were predicted by merging the table showing the Pfam domains of a protein and Pfam2GO, the related part of Interpro2GO. A GO term was attributed only if neither itself nor its children in the GO graph had been attributed to the query protein in the GAF file of *T. brucei TREU927* (downloaded from TriTrypDB version 60). GOATOOLS was used for working with GO terms (Klopfenstein et al., 2018).

Results and discussions

Primary inspection of the alignments

Table 1 shows the total number of alignments and the percentage of each label. Compared to the model organisms, a higher percentage of alignments against *T. brucei* are labeled -1, meaning there is a higher list of unannotated protein regions that are structurally similar to the Pfam Seeds and can be annotated by structural similarity. On the other hand, more than 97 percent of the alignments against *E. coli* have already been labeled, showing that only a few new domains can be annotated for *E. coli* by structural similarity to Pfam Seeds using this paper's approach.

The probability density functions of the alignment characteristics do not show sufficient separation between the 0 and 1 labels (supplementary figure 1). Yet alignment characteristics, such as lddt and fident have a more distinctive distribution for each label. The same plot shows that by having stringent cut-offs on fident, only hits labeled as one are retained, which aligns with earlier reports stating that at least 40% sequence identity is needed for confident, functional similarity inference through sequence similarity (Pearson, 2013). Yet, high thresholds on a single alignment characteristic will miss many hits.

Even though the distribution of the lddt and TM-score for labels 0 and 1 shows some separation, it is hard to select even a stringent threshold for separating positive labels. Proteins with similar structures but different functions align with this idea (Keskin & Nussinov, 2005). The scatter plots of the binary combination of alignment labels colored by the alignment label show that the combination of two alignment characteristics shows a higher separation of the labels (supplementary figure 1).

Table 1. The total number of alignments and the percentage of alignments labeled as 0 or 1 in studied organisms.

Organism	Total number of alignments	Percent of alignments labeled 1	Percent of alignments labeled 0
T. brucei	1.04 M	39.3%	35.4%
S. cerevisiae	1.13 M	43.9%	42.4%
E. coli	1.31 M	47.3%	51.0%
M. jannaschii	0.45 M	41.9%	52.7%

Performance of classifier

The cross-validation data showed that when the depth of the tree is unlimited, the trained model has the highest performance. The criterion for the split didn't have a significant effect on the performance. Models with a higher number of trees performed better, but the difference is insignificant for models with more than 100 trees. In the end, the model trained by 100 trees with the Gini split criterion and unlimited tree depth was used for the rest of the study.

The ROC curve shows that the trained model can separate the 0 and 1 labels in *T. brucei* with an Area Under the Curve (AUC) equal to 0.986 Figure 1. Given such a high AUC on a model trained on 10 parameters, we suspected there are universal rules for labeling the alignments, and they do not have a high dependency on the source of the training set. To test this, first, we redid the training on a smaller subsample of the training set. Even when just 0.01 of the whole training set is used in training, AUC is

0.95 (shown as DST in Figure 1). Testing the model trained on *T. brucei* on three model organisms from the domains of life again showed high AUCs, too (Figure 1).

Pfam domains have a heterogenous frequency; some are frequent in an organism, and some are abundant. Due to the small data size compared to the high number of Pfams, we couldn't stratify the model performance based on the annotated Pfam. Yet, to make sure the abundance of some Pfams does not overshadow the classifier's performance on the rare ones, we calculated the frequency of the Pfams of the query proteins in the training set and considered those whose frequency was above the median as frequent Pfams. Finally, we plotted the ROC by Excluding Frequent Pfams (EFP), as shown in Figure 1.

To see the classifier's performance on the query proteins whose Pfam is different from those used in training, we plotted the ROC for proteins whose Pfam was not among the Pfams of the query proteins in the training set (shown as Diff in Figure 1). We also plotted the ROC of the Diff datasets after excluding the frequent Pfams (shown as Diff, EFP in Figure 1). All these show general rules for inferring domain similarity from structural alignment characteristics.

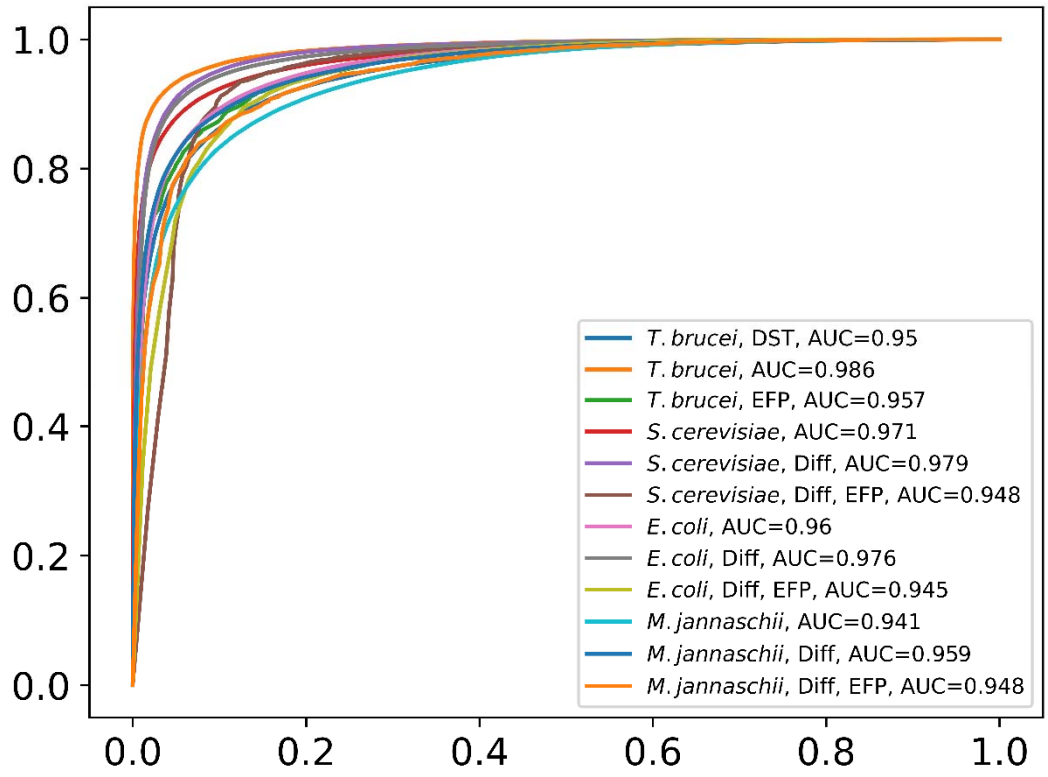

Figure 1. The ROC curve for classifying the structural alignment of Pfam instances to proteins of interest.

Figure 2 shows the model trained to classify the Pfam instance sequence alignments against proteins of interest. Compared to structure-based alignment (shown in Figure 1), it shows much lower generalizability concerning the training size, the query organism, Pfam frequency, and Pfam similarity to the training set.

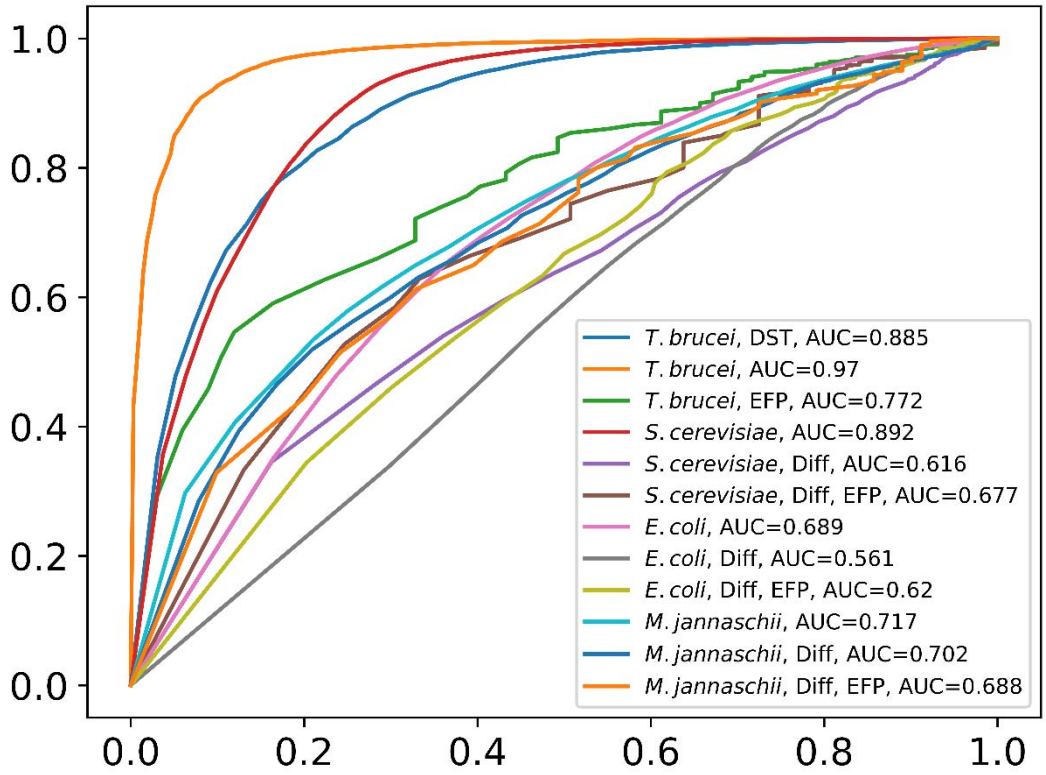

Figure 2. The ROC curve for classifying the sequence alignment of Pfam instances to proteins of interest.

The feature importance plot (Figure 3) shows that lddt is the most important feature in predicting the labels, signifying the importance of structural similarity in the classifier performance.

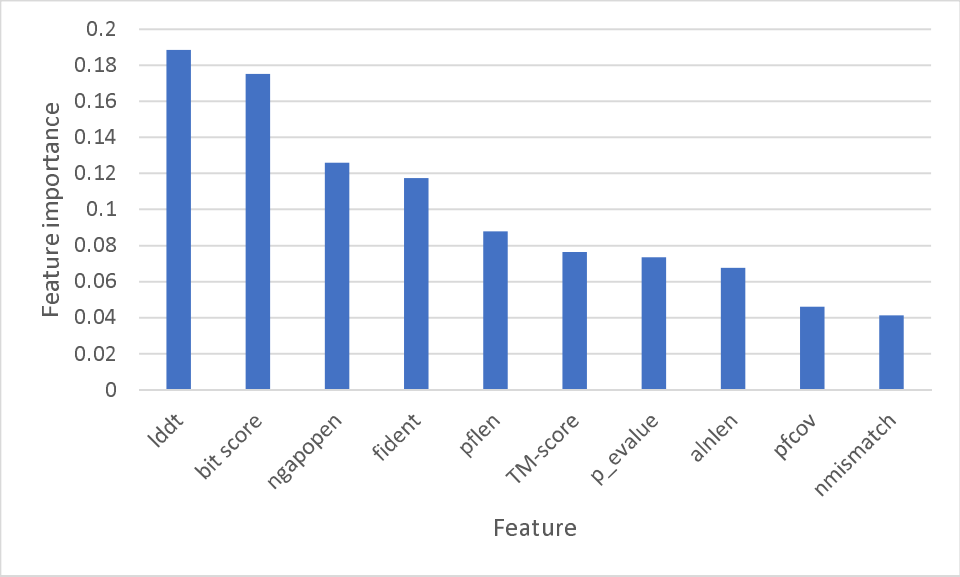

Figure 3. Feature importance plot

Greedy Pfam Annotation

Considering the formerly annotated domains of *T. brucei*, the current approach has a sensitivity of 50%, meaning that 50% of the formerly annotated Pfams were correctly retrieved using our approach. The precision for the predictions is 98.5%. It is worth mentioning that if no Pfam instance was structurally similar to previously Pfam annotated regions on a target protein, it wasn't considered a false negative on the data shown in Figure 1. In contrast, it is considered a false negative when the sensitivity of retrieving former Pfam annotations is calculated.

We expect to improve the sensitivity of the searching step using more sensitive tools. Recently, Foldseek has introduced iterative searching that is expected to have a higher sensitivity. Yet, the evaluation of iterative searching for the functional domain annotation is beyond the scope of this manuscript.

New Pfam predictions

A total of 1146 new Pfam domains are predicted for *T. brucei* (supplementary table 1). PF13458 and PF00913 are the two most frequently attributed Pfams; they have been attributed 79 and 31 times, respectively. All the proteins annotated with PF13458, "Leucine-binding protein domain," had adenylate cyclase in their description in TriTrypDB. Such a domain had been earlier suggested by comparing the profile of adenylate cyclases in *T. brucei* against Leucine binding proteins (Emes & Yang, 2008). All the proteins annotated with PF00913 already have been annotated as Variant Surface Glycoprotein in TriTrypDB, yet they didn't have the domain annotation in the reported area.

Case study

We looked deeper into the predictions for the proteins involved in mtRNA editing in *T. brucei* and found 18 new domains annotated in these proteins (Table 2). KRET1, MEAT1, and KPAP2 are annotated with a new Nucleotidyltransferase domain (PF01909), which Pfam-N has also predicted. The signature of the PF01909 shows that positions 21, 22, 33, and 35 of PF01909 are conserved, and they contain Glycine, Serine, Aspartic acid, and Aspartic acid, respectively. In KRET1, MEAT1, and KPAP2, all corresponding positions have the same amino acid. For all covered cases and the following proteins, mapping between conserved positions of the Pfam signature and the query proteins is available in supplementary table 2, and supplementary table 3 shows domains predicted by other databases for the area a Pfam domain has been annotated.

Table 2. Domain predictions for the proteins involved in mtRNA editing in T. brucei and their coordinates (based on Uniprot genome annotation)

Name	UniProt ID	Predicted Pfam	Start	End	Confidence score	Pfam Description
KRET1	Q57UM5	PF01909	278	513	0.8	Nucleotidyltransferase domain
MEAT1	Q4GZ86	PF01909	34	173	0.94	Nucleotidyltransferase domain
KPAP2	Q38CM2	PF01909	25	188	0.73	Nucleotidyltransferase domain
MPSS5	Q38F66	PF00436	178	281	0.97	Single-strand binding protein family
MPSS6	Q584T7	PF00436	170	279	0.81	Single-strand binding protein family
KREPA1	Q586L9	PF00436	627	761	0.94	Single-strand binding protein family
KREPA2	Q38AE3	PF00436	490	585	0.7	Single-strand binding protein family
KREPA3	D6XMN0	PF00436	290	393	0.99	Single-strand binding protein family
KREPA4	Q38B91	PF00436	101	211	0.89	Single-strand binding protein family
KREPA5	Q57YG2	PF00436	14	146	1	Single-strand binding protein family
KREX1	Q57WU3	PF03159	20	208	0.73	XRN 5'-3' exonuclease N-terminus
KREX2	Q38BP2	PF03159	13	230	0.58	XRN 5'-3' exonuclease N-terminus
MERS1	Q381M2	PF00293	193	351	0.61	NUDIX domain
KPAF3	Q38DD7	PF14559	155	222	1	Tetratricopeptide repeat
KREN1	Q4GZ50	PF00636	271	363	1	Ribonuclease III domain
KREX1	Q57WU3	PF03372	621	882	0.93	Endonuclease/Exonuclease/phosphatase family
RESC5	Q389F5	PF19420	32	281	0.7	N, N dimethylarginine dimethylhydrolase, eukaryotic
KH2F2	Q584U2	PF01557	65	266	0.93	Fumarylacetoacetate (FAA) hydrolase family

MPSS5, MPSS6, KREPA1, KREPA2, KREPA3, KREPA4, and KREPA5 are predicted to have PF00436, Single-strand binding domains (SSB). Pfam had predicted PF00436 for KREPA6. KREPA4 and KREPA6 are

adjacent to each other on the genome, and they are the closest non-self structural hits when each is searched against the structure of all organism proteins, suggesting that they might be paralogs of each other. It is expected that paralogous proteins contain similar domains. Pfam-N predicts the same domain for KREPA3 and KREPA5. For KREPA2, an OB-fold has been predicted by Pfam-N (PF01336). PF01336 and PF00436 belong to the same Pfam clan, meaning their sequence signatures are close to each other.

KREX1 and KREX2 are predicted to have PF03159, XRN 5'-3' exonuclease N-terminus domain. Pfam-N had predicted the same domains for them. An earlier study hypothesized such a function for the same proteins (MP100 and MP99 in the paper) based on amino acid conservation. Yet, the overall homology to the known XRN domains of the time had been low in attributing the domain confidently (Worthey et al., 2003), which now can be justified by structural similarity.

For KPAF3, a Tetratricopeptide repeat (PF14559) is predicted to be in the N-terminal region. The same protein has been annotated with multiple PF01535, Pentatricopeptide repeat (PPRs) in its C-terminal. Interpro Domain Architecture plots show it is common to see multiple PPRs in tandem on a protein. Furthermore, PF01535 and PF14559 belong to the same Pfam clan, raising the possibility of attributing a wrong Pfam. However, such a role has been inferred by the similarity of KPAF3 to position 227 to 294 of Q4CQ55, the ortholog of KPAF3 in *Trypanosoma cruzi*, a phylogenetically close organism with over 66% sequence identity and TM-score = 0.99. Given such a high similarity, PF14559 seems to be a correct prediction.

For the proteins in the case study, if the new Pfam annotation coincided with the domain annotation by Pfam-N, they both annotated the same domain, or the predictions belonged to the same Pfam clan. However, in some cases, even though the current approach and Pfam-N predicted the same domain for a region, Pfam-N predicted two fragmented domains of the same ID in the region. For instance, our model predicted a PF00436 in KREPA3 between positions 290 and 393, while Pfam-N predicted two domains between 313 and 341 and the other one between 348 and 387. Comparing the structures encompassing these regions to the structure of proteins annotated with the same domain (such as 1kaw in PDB), it turns out these are fragmented predictions.

GO term annotations

In total, 494 new GO terms, including 275 molecular function terms, 169 biological process terms, and 50 cellular component terms, are attributed to *T. brucei* proteome. The discrepancy between the

number of annotated domains and the number of GO terms can be because some Pfam domains still need to be associated with GO terms. Besides, in some cases, GO terms related to the newly annotated domains had been previously attributed to the same protein, and here, we only reported the new GO terms.

Future directions

As explained earlier, Foldseek converts the 3D structure of a protein to a linear sequence called the 3Di sequence, which is later used for searching structures. Currently, Foldseek uses fixed rewards/penalties for the alignment. As profile-based searches increase the sensitivity of sequence search tools, we expect to see a higher sensitivity here if they are used for searching the 3Di sequences. We expect a higher performance with the same logic if Convolutional Neural Networks are used for domain annotation like Pfam-N (Bileschi et al., 2022).

From an organism-specific point of view, we now have some hints about the function of genes in *T. brucei,* which can be tested experimentally for validation. We will validate some functions in our lab soon.

Conclusion

In this study, a simple model was trained to predict the domains based on the overall structural similarity of the query protein to the annotated domains. Against its simpleness, it could predict the correct Pfams in most cases for the previously annotated domains and could predict new domains in the proteome of *T. brucei*. Results show there are some universal rules for inferring domain similarity based on structural alignment characteristics, and these rules can be learned by a few training samples and have quite a high generalizability, while this is not the case for models trained on sequence similarity. Manual inspection of some results supports the predictions. The study shows that more domains can be annotated by considering structural similarity.

Finally, even though we have used the output of this study for functional annotation, accurate domain predictions by a simple RF indirectly show that the structures have been predicted very accurately by Alphafold2, and Foldseek correctly finds relevant structures.

Acknowledgements

This project has been supported by a research grant 252733 form CIHR to RS. Robert Harpur foundation is also acknowledged for providing financial support to PMB.

References

Amos, B., Aurrecoechea, C., Barba, M., Barreto, A., Basenko, E. Y., Belnap, R., Blevins, A. S., Böhme, U., Brestelli, J., Brunk, B. P., & others. (2022). VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Research*, *50*(D1), D898–D911.

Aphasizheva, I., Alfonzo, J., Carnes, J., Cestari, I., Cruz-Reyes, J., Göringer, H. U., Hajduk, S., Lukeš, J., Madison-Antenucci, S., Maslov, D. A., & others. (2020). Lexis and grammar of mitochondrial RNA processing in trypanosomes. *Trends in Parasitology*, *36*(4), 337–355.

Ayoub, R., & Lee, Y. (2019). RUPEE: A fast and accurate purely geometric protein structure search. *PLoS One*, *14*(3), e0213712.

Bartas, M., Slychko, K., Brázda, V., Červeň, J., Beaudoin, C. A., Blundell, T. L., & Pečinka, P. (2022). Searching for New Z-DNA/Z-RNA Binding Proteins Based on Structural Similarity to Experimentally Validated Zα Domain. *International Journal of Molecular Sciences*, *23*(2). https://doi.org/10.3390/ijms23020768

Bileschi, M. L., Belanger, D., Bryant, D. H., Sanderson, T., Carter, B., Sculley, D., Bateman, A., DePristo, M. A., & Colwell, L. J. (2022). Using deep learning to annotate the protein universe. *Nature Biotechnology*, 1–6.

Chawla, N. v, Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Emes, R. D., & Yang, Z. (2008). Duplicated paralogous genes subject to positive selection in the genome of Trypanosoma brucei. *PloS One*, *3*(5), e2295.

Gene Ontology Consortium Website. (n.d.). Retrieved September 20, 2022, from http://geneontology.org/docs/guide-go-evidence-codes/

Gligorijević, V., Renfrew, P. D., Kosciolek, T., Leman, J. K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B. C., Fisk, I. M., Vlamakis, H., & others. (2021). Structure-based protein function prediction using graph convolutional networks. *Nature Communications*, *12*(1), 1–14.

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282.

Holm, L. (2022). Dali server: structural unification of protein families. *Nucleic Acids Research*.

Illergård, K., Ardell, D. H., & Elofsson, A. (2009). Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. *Proteins: Structure, Function, and Bioinformatics*, *77*(3), 499–508.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Ž'idek, A., Potapenko, A., & others. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589.

Keskin, O., & Nussinov, R. (2005). Favorable scaffolds: proteins with different sequence, structure and function may associate in similar ways. *Protein Engineering Design and Selection*, *18*(1), 11–24.

Klopfenstein, D. v, Zhang, L., Pedersen, B. S., Ram\'\irez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C. J., Yunes, J. M., Botvinnik, O., Weigel, M., & others. (2018). GOATOOLS: A Python library for Gene Ontology analyses. *Scientific Reports*, *8*(1), 1–17.

Li, G., Huang, J., Yang, J., He, D., Wang, C., Qi, X., Taylor, I. A., Liu, J., & Peng, Y.-L. (2018). Structure based function-annotation of hypothetical protein MGG_01005 from Magnaporthe oryzae reveals it is the dynein light chain orthologue of dynlt1/3. *Scientific Reports*, *8*(1), 1–11.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., & others. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419.

Monzon, V. A., Paysan-Lafosse, T., Wood, V., & Bateman, A. (2022). Reciprocal best structure hits: Using AlphaFold models to discover distant homologues. *BioRxiv*.

Nagano, N., Orengo, C. A., & Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *Journal of Molecular Biology*, *321*(5), 741–765.

Paysan-Lafosse, T., Blum, M., Chuguransky, S., Grego, T., Pinto, B. L., Salazar, G. A., Bileschi, M. L., Bork, P., Bridge, A., Colwell, L., & others. (2022). InterPro in 2022. *Nucleic Acids Research*.

Pearson, W. R. (2013). An introduction to sequence similarity ("homology") searching. *Current Protocols in Bioinformatics*, *42*(1), 1–3.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel V. and Thirion, B., Grisel, O., Blondel, M., Prettenhofer P. and Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., & Finn, R. D. (2018). HMMER web server: 2018 update. *Nucleic Acids Research*, *46*(W1), W200–W204.

Söding, J., Biegert, A., & Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research*, *33*(suppl_2), W244–W248.

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, *35*(11), 1026–1028.

van Kempen, M., Kim, S., Tumescheit, C., Mirdita, M., Söding, J., & Steinegger, M. (2022). Foldseek: fast and accurate protein structure search. *BioRxiv*.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., & others. (2022). AlphaFold Protein Structure Database: massively expanding the

structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, *50*(D1), D439–D444.

Waskom, M. L. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Worthey, E. A., Schnaufer, A., Mian, I. S., Stuart, K., & Salavati, R. (2003). Comparative analysis of editosome proteins in trypanosomatids. *Nucleic Acids Research*, *31*(22), 6392–6408.

Xu, J., & Zhang, Y. (2010). How significant is a protein structure similarity with TM-score= 0.5? *Bioinformatics*, *26*(7), 889–895.

Zarembinski, T. I., Hung, L.-W., Mueller-Dieckmann, H.-J., Kim, K.-K., Yokota, H., Kim, R., & Kim, S.-H. (1998). Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proceedings of the National Academy of Sciences*, *95*(26), 15189–15193.

Zhang, C., Freddolino, P. L., & Zhang, Y. (2017). COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. *Nucleic Acids Research*, *45*(W1), W291–W299.

Zhang, Y., & Skolnick, J. (2005). TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research*, *33*(7), 2302–2309.