pLM-BLAST – distant homology detection based on direct comparison of sequence representations from protein language models

Kamil Kaminski[a,b], Jan Ludwiczak[a,†], Vikram Alva[c], and Stanislaw Dunin-Horkawicz[a,b,c,*]

[a] Institute of Evolutionary Biology, Faculty of Biology, Biological and Chemical Research Centre, University of Warsaw, Żwirki i Wigury 101, 02-089, Warsaw, Poland

[b] Laboratory of Structural Bioinformatics, Centre of New Technologies, University of Warsaw, Banacha 2c, 02-097 Warsaw, Poland

[c] Department of Protein Evolution, Max Planck Institute for Biology Tübingen, Max-Planck-Ring 5, 72076 Tübingen, Germany

[†] Present address: Prescient Design, Genentech Research & Early Development, Roche Group, Grenzacherstrasse 124, 4070 Basel, Switzerland.

[*] To whom correspondence should be addressed
E-mail: stanislaw.dunin-horkawicz@tuebingen.mpg.de

Abstract

Summary: Homology detection by sequence comparison is a typical first step in the study of protein function and evolution. Here, we describe a new homology detection tool, pLM-BLAST, that uses a modified Smith-Waterman algorithm for unsupervised comparison of single-sequence representations obtained from a protein language model (such as ProtT5) trained on millions of sequences. In our benchmarks, pLM-BLAST has shown the ability to detect homology between highly divergent proteins, demonstrating its applicability to tasks such as protein classification, domain annotation, and function prediction. *Availability and Implementation:* pLM-BLAST is available as a web server in the MPI Bioinformatics Toolkit (https://toolkit.tuebingen.mpg.de/tools/pLM-BLAST), where it can be used to search precomputed databases. It is also available as a standalone tool to build custom databases and run batch searches (https://github.com/labstructbioinf/pLM-BLAST).

Introduction

Homology, i.e., descent from a common ancestor, and its inference are fundamental to comparative, evolutionary, and molecular biology. In the case of proteins, statistically significant local or global sequence similarity is accepted as the primary marker for inferring homology. When the similarity between the protein sequences being compared is high (>=30%), homology can be readily detected using methods based on sequence-to-sequence and sequence-to-profile comparisons such as BLAST or PSI-BLAST (Altschul *et al.*, 1997; Eddy, 2011). However, when similarity is low (<30%), methods based on profile HMMs, such as HMMER and HHsearch, are currently our best tools to infer homology (Steinegger *et al.*, 2019).

In the case of highly distant evolutionary relationships, sequences may have diverged to the point where we can no longer detect their relatedness. Because structures diverge much more slowly than sequences, their similarity is often used to infer homology in such cases. In fact, owing to the recent revolution in structure prediction (Lin *et al.*, 2022; Jumper *et al.*, 2021; Li *et al.*, 2022), predicted structures are available for a large proportion of known proteins, and structural similarity is increasingly being used to infer homology. However, similar structures, particularly at the domain and sub-domain levels, may have evolved convergently due to the limited number of structural solutions available to a folded polypeptide chain, and thus structural similarity is often not conclusive evidence of common

ancestry. To address this issue, several deep learning-based methods have sought to detect weak sequence signals between highly divergent proteins in recent years (Zheng *et al.*, 2019; Gao and Skolnick, 2021; Li *et al.*, 2017). The most promising of these methods, such as TM-Vec (Hamamsy *et al.*, 2022) and DEDAL (Llinares-López *et al.*, 2022), rely on single-sequence representations obtained from protein language models (pLMs) (Bepler and Berger, 2021). pLMs are neural networks trained on a large number of natural protein sequences for tasks such as guessing masked residues based on contextual information (Elnaggar *et al.*, 2021; Lin *et al.*, 2022). Once trained, pLMs can be used to rapidly compute the aforementioned representations (also referred to as embeddings), i.e., numerical descriptors of protein sequences that place them in the context of the total knowledge collected by the network. Because of their high information content, sequence embeddings have been successfully applied to many other tasks, including prediction of tertiary structures (Lin *et al.*, 2022), transmembrane segments (Bernhofer and Rost, 2022), and signal peptides (Teufel *et al.*, 2022).

In this report, we describe pLM-BLAST, a new tool for detecting local homology between protein sequences that combines pLM representations with a local similarity detection algorithm inspired by BLAST (Altschul *et al.*, 1997). In contrast to TM-vec and DEDAL, pLM-BLAST is based on an unsupervised approach that does not require training of a specialized deep-learning model or defining positive labels based on structurally similar or homologous protein pairs. To this end, we implemented a modified Smith-Waterman algorithm in which the substitution matrix is generated directly from the raw pLM representations of the two sequences to be compared. In this work, we used the pLM ProtT5 (Elnaggar *et al.*, 2021), but this approach can be used in conjunction with other pLMs, such as ESM-2 (Lin *et al.*, 2022), to generate local alignments. In a benchmark using domain pairs from the ECOD protein classification database (Cheng *et al.*, 2014), pLM-BLAST demonstrated its suitability for fast pairwise comparisons and database searches to detect distant homology.

Methods

pLM-BLAST extends the concept of BLAST by replacing invariant substitution matrices, such as BLOSUM62 (Henikoff and Henikoff, 1992), with per-residue similarities between ProtT5

embeddings. Consequently, the similarity between a given pair of residues is entirely context-dependent. Such sequence context information has been shown to significantly improve the sensitivity of sequence search methods and enable the detection of nontrivial conserved patterns (Biegert and Söding, 2009; Remmert *et al.*, 2011).

Substitution matrix

The embedding of a sequence seq_l is given by matrix $\mathbf{E}_l \in \{e_{ij}\}$ of size $n \times m$, where n and m denote the number of residues and the size of the embedding, respectively. To obtain standardized and comparable results, the values in each row are scaled by dividing them by the Euclidean norm of the row

$$e_{ij}^{\dagger} = \frac{1}{\sqrt{\sum_{j=m}^{j=0} e_{ij}^2}} e_{ij}$$

where e_{ij} denotes the $\mathbf{E}_l$ matrix element in the i-th row and j-th column, and † is the normalization operator. In pLM-BLAST, a substitution matrix $\mathbf{S}_{lk}$ for two sequences seq_l and seq_k is calculated as

$$\mathbf{S}_{lk} = \mathbf{E}_l^{\dagger T} \mathbf{E}_k^{\dagger}$$

The element s_{ij} of the matrix S_{lk} is in the interval $(-1,1)$ and denotes the embedding similarity of the i-th residue of seq_L and the j-th residue of seq_K. In other words, the above equations are equivalent to calculating the cosine similarity of each possible pair of residues from seq_L and seq_K and placing them in a two-dimensional array.

Scoring matrix

A scheme adapted from the Smith-Waterman algorithm is used to create a scoring matrix $H_{LK} \ni \{h_{ij}\}$ of size $(|L| + 1) \times (|K| + 1)$, where the first row and column are filled with zeros and the rest of the matrix is filled with the following procedure

$$h_{ij} = \max \begin{cases} h_{i-1,j-1} + s_{ij} \\ \max_p\{h_{i-p,j}\} \\ \max_s\{h_{i,-s}\} \end{cases}$$

Unlike the original Smith-Waterman algorithm, pLM-BLAST does not use gap penalties because the s_{ij} values for dissimilar regions are negative due to the use of the cosine similarity property. Gap penalties were implemented, but in our tests they did not provide any significant benefits, so they are turned off by default.

Traceback procedure

To extract matching regions for a given pair of sequences, a traceback procedure is used. For this purpose, n possible paths P are constructed, starting from the right and bottom edges of the matrix $\mathbf{H}$. The path p_n is a set of coordinates from the matrix H_{LK}, where $n \in N$ and N equals $|L| + |K|$.

Input: H

Output: P

For $n \in N$ **do**

    Initialize path p_n
    Set pair (i', j') to n border coordinates
    While $i \geq 1$ and $j \geq 1$ **do**
        $(i', j') = argmax_{i,j}\{H_{i-1,j-1}, H_{i,j-1}, H_{i-1,j}\}$
    If $\boldsymbol{H}_{i',j'} \leq \mathbf{0}$ **then**
        Break
    Else
        Add (i', j') to p_n
        Decrement by one i, j

Local alignment

Each path is scanned for the presence of high-scoring sub-paths (local alignments) using the substitution matrix. A moving average is applied to each path p_n and all continuous subpaths where the average is greater than a specified threshold are treated as local alignments (the threshold is set by default to 1 standard deviation of the substitution matrix). Each local alignment is assigned a score calculated as the mean of substitution matrix values at coordinates defined by its subpath. In the case of database searches, an optional procedure may be applied in which local alignments to a single database entry are merged. Our tests showed that the standard deviation of the substitution matrix tends to assume a value of around 0.05, whereas the path score for meaningful alignments is an

order of magnitude higher. However, increasing (>1) or decreasing (<1) the standard deviation cut-off will make the algorithm more strict or permissive, respectively.

Benchmark

The sequences of all ECOD domains (version 20220912), excluding nested domains, with a length of 50-600 amino acids were obtained with *localpdb* ECOD plugin (Ludwiczak *et al.*, 2022), and clustered using MMSeqs2 at a sequence identity cut-off of 30% (Steinegger and Söding, 2017). The longest sequences were selected as clusters' representatives and those belonging to H-groups with less than five members were removed. The final benchmark set was created by randomly selecting five domains from each Homology(H)-group, an ECOD level that groups together domains with common ancestry. This procedure was designed to select domains from different Topology(T)-groups; a T-group is a sublevel that groups together homologous domains within an H-group that are topologically similar. For each ECOD domain, an HMM profile and sequence embedding were calculated using three iterations of HHblits over the UniRef30 database with default settings (Steinegger *et al.*, 2019) and ProtT5 (prot_t5_xl_half_uniref50-enc) (Elnaggar *et al.*, 2021), respectively (note that the usage of the ProtT5 model limits the maximal sequence length to 1000). The embeddings were compared all-against-all with pLM-BLAST and TM-Vec, whereas the HMM profiles were compared with HHsearch. To speed up the embedding comparison, a pre-filtering procedure was used. All the embeddings were individually flattened to a constant size by applying an arithmetic mean and then efficiently compared to each other with a cosine similarity metric. Only pairs that scored 0.4 or higher were passed to the actual pLM-BLAST (the 0.4 cut-off corresponds to the 90th percentile of all scores). The same procedure was used to improve the performance of database searches: first, the database of flattened embeddings is searched with the cosine similarity, which is much faster than the comparison of per-residue embeddings, and then the actual pLM-BLAST comparisons are performed only for matches above the user-provided cosine similarity cut-off (95th percentile by default).

Results

To assess the sensitivity of pLM-BLAST in detecting distant homologs, we evaluated its ability to correctly detect homology between ECOD domains in a benchmark set (see

Methods) and compared its sensitivity in this task to that of BLAST, TM-Vec, and HHsearch. While matches between domains from the same H-group were considered true positives, matches between different H-groups were considered false positives. In our benchmark, HHsearch performed best, pLM-BLAST and TM-Vec performed equally well, and BLAST was the least sensitive. This result was quite unexpected since TM-Vec uses sequence representations from ProtT5 as an input to a specialized deep learning model, trained end-to-end to produce structural alignments, while pLM-BLAST is based only on unsupervised comparisons of these representations. This suggests that the information contained in the raw embeddings is sufficient for homology detection (Figure 1A).

We conjectured that the significantly higher performance of HHsearch in our benchmark may result from it detecting a higher number of true positives and/or fewer false positives. To investigate this systematically, we used US-align (Zhang *et al.*, 2022) to compare the structures corresponding to the true positive and false positive pairs returned by the three tools (Figure 1D). In the case of HHsearch, we found only a few structurally similar false positive pairs, which is to be expected given that ECOD H-groups were defined using HHsearch (Cheng *et al.*, 2014). In contrast to HHsearch, both pLM-BLAST and TM-Vec yielded a number of structurally similar false positive pairs (Supplementary Table 1; Figure 1D), many of which belong to the same X-group, an ECOD level that groups together domains that may be homologous. In addition, both TM-Vec and pLM-BLAST also detected connections between domains of different X-groups that were often hinged on the presence of a subdomain-sized fragment (Alva *et al.*, 2015; Kolodny, 2021). For example, pLM-BLAST found a fragment conserved between domains of two different sandwich folds, namely cupredoxin (ECOD H-group 3156.1) and immunoglobulin (11.1, Figure 1B). A homologous relationship between these two domains has been discussed previously (Gough and Chothia, 2004; Stevens, 2008), and a pLM-BLAST scan of the entire ECOD database with a cupredoxin domain as a query revealed additional connections to Ig-like domains (Figure 1C). We also tested other pLM models such as ESM2 (Lin *et al.*, 2022) and SeqVec (Heinzinger *et al.*, 2019); however, ProtT5 performed best in our tests.

In summary, pLM-BLAST is a sensitive tool for remote homology detection based on comparison of sequence representations obtained from the pLM ProtT5. pLM-BLAST is

available both as a stand-alone package and as an easy-to-use web server within the MPI Bioinformatics Toolkit (Zimmermann *et al.*, 2018), where it can be used to search several precomputed databases (e.g., ECOD, InterPro, and PDB). pLM-BLAST is also quite fast: while the computation of ProtT5 representations for ECOD30, which included 29,330 domains, took 30 minutes on a single NVIDIA RTX A4500 card, a search over it took on average 2 mins on a 20-core CPU (including the steps of loading the database to the memory and formatting the results). Currently, the main limitations of pLM-BLAST are the maximal size of the target database (searching huge, redundant databases containing millions of sequences would be computationally very expensive) and the fact that it typically produces rather short, local alignments (this, however, may be attributed to the used pLM model rather the pLM-BLAST algorithm itself). Nevertheless, we anticipate that pLM-BLAST will prove useful for studying protein function and evolution, also in the cases of singletons/orphan sequences and taxonomically-restricted genes (Barrera-Redondo *et al.*, 2022) for which deep HMM profiles cannot be calculated.

Acknowledgments

This work was supported by the First TEAM program of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund (grant POIR.04.04.00-00-5CF1/18-00). V.A. and S.D-H. were supported by institutional funds from the Max Planck Society. The authors would like to thank Andrei Lupas, Hajk-Georg Drost, and Kaiyu Qiu for their constructive comments and critical evaluation of the manuscript.

Figure Legends

Figure 1. (A) Benchmark on the task of reconstructing H-groups defined in the ECOD database. False positives are shown in the logarithmic scale. **(B)** A substitution matrix depicting the similarity between cupredoxin and immunoglobulin folds. Axes numbering denote the consecutive residues. The brighter the color the more similar the sequences in a given area. The highest-scoring local alignment is indicated with a thin red line. **(C)** (top) pLM-BLAST scan of ECOD database with the sequence of cupredoxin domain. (bottom) Structures of cupredoxin and immunoglobulin folds where homologous regions detected by pLM-BLAST are shown in red. **(D)** Sequence and structure similarity of false positive pairs.

Sequence similarity was calculated with HHsearch (left panel), pLM-BLAST (middle), and TM-Vec (right).

Bibliography

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–402.

Alva,V. *et al.* (2015) A vocabulary of ancient peptides at the origin of folded proteins. *Elife*, **4**, e09410.

Barrera-Redondo,J. *et al.* (2022) Uncovering gene-family founder events during major evolutionary transitions in animals, plants and fungi using GenEra. *bioRxiv*, 2022.07.07.498977.

Bepler,T. and Berger,B. (2021) Learning the protein language: Evolution, structure, and function. *Cell Syst.*, **12**, 654-669.e3.

Bernhofer,M. and Rost,B. (2022) TMbed: transmembrane proteins predicted through language model embeddings. *BMC Bioinformatics*, **23**, 326.

Biegert,A. and Söding,J. (2009) Sequence context-specific profiles for homology searching. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 3770–5.

Cheng,H. *et al.* (2014) ECOD: an evolutionary classification of protein domains. *PLoS Comput. Biol.*, **10**, e1003926.

Eddy,S.R. (2011) Accelerated Profile HMM Searches. *PLoS Comput. Biol.*, **7**, e1002195.

Elnaggar,A. *et al.* (2021) ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1–1.

Gao,M. and Skolnick,J. (2021) A novel sequence alignment algorithm based on deep learning of the protein folding code. *Bioinformatics*, **37**, 490–496.

Gough,J. and Chothia,C. (2004) The Linked Conservation of Structure and Function in a Family of High Diversity. *Structure*, **12**, 917–925.

Hamamsy,T. *et al.* (2022) TM-Vec: template modeling vectors for fast homology detection and alignment. *bioRxiv*, 2022.07.25.501437.

Heinzinger,M. *et al.* (2019) Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, **20**, 723.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks.

Proc. Natl. Acad. Sci., **89**, 10915–10919.

Jumper,J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.

Kolodny,R. (2021) Searching protein space for ancient sub-domain segments. *Curr. Opin. Struct. Biol.*, **68**, 105–112.

Li,S. *et al.* (2017) Protein remote homology detection based on bidirectional long short-term memory. *BMC Bioinformatics*, **18**, 443.

Li,Z. *et al.* (2022) Uni-Fold: An Open-Source Platform for Developing Protein Folding Models beyond AlphaFold. *bioRxiv*, 2022.08.04.502811.

Lin,Z. *et al.* (2022) Evolutionary-scale prediction of atomic level protein structure with a language model. *bioRxiv*, 2022.07.20.500902.

Llinares-López,F. *et al.* (2022) Deep embedding and alignment of protein sequences. *bioRxiv*, 2021.11.15.468653.

Ludwiczak,J. *et al.* (2022) Localpdb- a Python package to manage protein structures and their annotations. *Bioinformatics*.

Remmert,M. *et al.* (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Steinegger,M. *et al.* (2019) HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, **20**, 473.

Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.

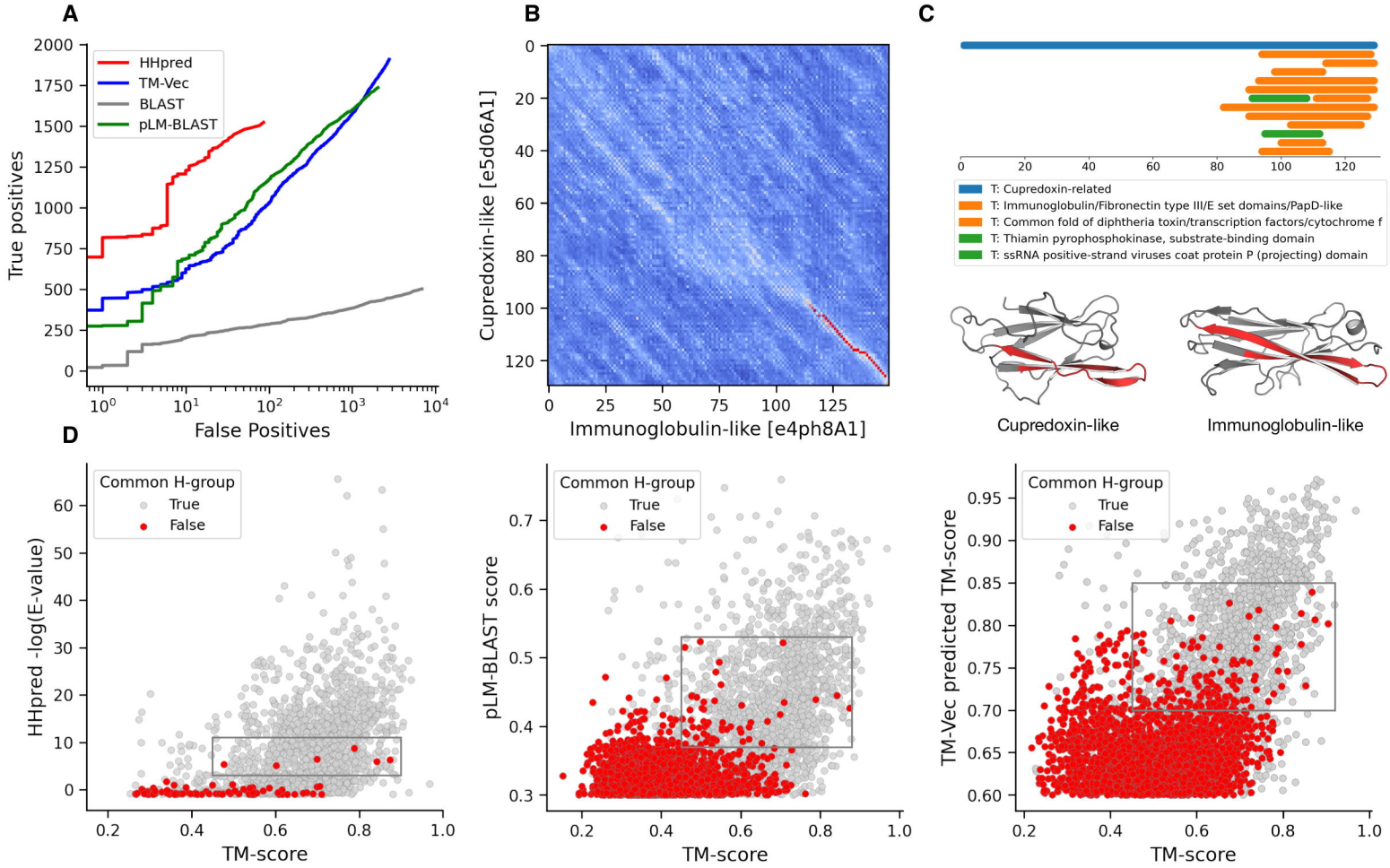
Stevens,F.J. (2008) Homology versus analogy: possible evolutionary relationship of immunoglobulins, cupredoxins, and Cu,Zn-superoxide dismutase. *J. Mol. Recognit.*, **21**, 20–29.

Teufel,F. *et al.* (2022) SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.*, **40**, 1023–1025.

Zhang,C. *et al.* (2022) US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat. Methods*, **19**, 1109–1115.

Zheng,W. *et al.* (2019) Detecting distant-homology protein structures by aligning deep neural-network based contact maps. *PLOS Comput. Biol.*, **15**, 1–27.

Zimmermann,L. *et al.* (2018) A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. *J. Mol. Biol.*, **430**, 2237–2243.


A
HHpred
TM-Vec
BLAST
pLM-BLAST
True positives
False Positives
B
Cupredoxin-like [e5d06A1]
Immunoglobulin-like [e4ph8A1]
C
T: Cupredoxin-related
T: Immunoglobulin/Fibronectin type III/E set domains/PapD-like
T: Common fold of diphtheria toxin/transcription factors/cytochrome f
T: Thiamin pyrophosphokinase, substrate-binding domain
T: ssRNA positive-strand viruses coat protein P (projecting) domain
Cupredoxin-like
Immunoglobulin-like
D
Common H-group
True
False
HHpred -log(E-value)
pLM-BLAST score
TM-Vec predicted TM-score
TM-score