*Bioinformatics*, 33(21), 2017, 3473–3476
doi: 10.1093/bioinformatics/btx429
Advance Access Publication Date: 3 July 2017
Applications Note

Sequence analysis

# ProtDec-LTR2.0: an improved method for protein remote homology detection by combining pseudo protein and supervised Learning to Rank

**Junjie Chen, Mingyue Guo, Shumin Li and Bin Liu***

School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, China

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on April 8, 2017; revised on June 1, 2017; editorial decision on June 12, 2017; accepted on June 29, 2017

## Abstract

**Summary**: As one of the most important tasks in protein sequence analysis, protein remote homology detection is critical for both basic research and practical applications. Here, we present an effective web server for protein remote homology detection called *ProtDec-LTR2.0* by combining ProtDec-Learning to Rank (LTR) and pseudo protein representation. Experimental results showed that the detection performance is obviously improved. The web server provides a user-friendly interface to explore the sequence and structure information of candidate proteins and find their conserved domains by launching a multiple sequence alignment tool.

**Availability and implementation:** The web server is free and open to all users with no login requirement at http://bioinformatics.hitsz.edu.cn/ProtDec-LTR2.0/.

**Contact:** bliu@hit.edu.cn

## 1 Introduction

Protein remote homology detection is critical for both basic research (such as protein attribute prediction) and practical application (such as modeling the 3D structures of target proteins for drug development).

A great variety of computational approaches have been developed to detect protein remote homology (Chen *et al*., 2016a; Wei and Zou, 2016). The key to improve the performance of homology detection is to find a suitable approach to incorporate the evolutionary information into the predictors. The profile (Altschul *et al*., 1997) and HMM-based profile model (Remmert *et al*., 2012) are two rich protein representations, but they require a time consuming process to compute the similarity score by using alignment algorithm, and redundant information is often in these profiles. In contrast, the pseudo protein approach (Liu *et al*., 2014a,b), a language model of protein sequences, is an efficient protein representation with less noisy information and lower computational cost. It has been successfully employed by many computational predictors, such as disPseAAC (Liu *et al*., 2015b), SVM-Ensemble (Chen *et al*., 2016b), PDC-Ensemble (Liu *et al*., 2016), and dRHP-PseRA (Chen *et al*., 2016c). For more information of protein remote homology detection, please refer to a recent review paper (Chen *et al*., 2016a), which comprehensively compared the state-of-the-art methods in this field.

ProtDec-LTR is the first approach to combine different ranking predictors by using Learning to Rank (LTR) in a supervised manner (Liu *et al*., 2015a). It takes the advantages of different basic ranking predictors to achieve sensitive and stable performance. However, ProtDec-LTR is only based on the protein sequences without using the evolutionary information in profiles, and its web server is not available, preventing its applications.

Here, we are to propose a new web server for protein remote homology detection, called *ProtDec-LTR2.0*, which improves the previous ProtDec-LTR (Liu *et al*., 2015a) by incorporating the evolutionary information from pseudo protein representations. Furthermore, *ProtDec-LTR2.0* is trained with an updated benchmark dataset SCOPe with more samples and protein families, making it more robust and useful. *ProtDec-LTR2.0* web server provides graphical interface, facilitating to explore the sequence and structure information of detected proteins and their conserved regions.

## 2 Implementation

As demonstrated in previous studies (Chen *et al.*, 2016b,c; Liu *et al.*, 2008, 2014a,b; Wei *et al.*, 2015), pseudo protein representation is useful for protein remote homology detection. For achieving highly sensitive performance, the pseudo protein approach is employed, which can extract the evolutionary information from the profiles. The detailed description of pseudo protein can be found in (Liu *et al.*, 2008).

The detailed process of *ProtDec-LTR2.0* is shown in the followings. First, the query and all the proteins in the benchmark database are transformed into pseudo proteins by using PSI-BLAST v2.2.30+ with parameters (-num_iteratives 3 -evalue 0.001) and the NCBI's nrdb90 database, and then three state-of-the-art tools [PSI-BLAST v2.2.30+ (Altschul *et al.*, 1997), HHblits v2.0.15 (Remmert *et al.*, 2012) and Hmmer v3.1b2 (Mistry *et al.*, 2013)] are used as basic ranking predictors to search the query protein against the benchmark database with their default parameters. For each query, it can be represented as a feature matrix Π:

$$\mathbf{\Pi} = \begin{bmatrix} s_1(q,p_1) & s_2(q,p_1) & \cdots & s_8(q,p_1) \\ s_1(q,p_2) & s_2(q,p_2) & \cdots & s_8(q,p_2) \\ \vdots & \vdots & \vdots & \vdots \\ s_1(q,p_n) & s_2(q,p_n) & \cdots & s_8(q,p_n) \end{bmatrix} \tag{1}$$

where $p_i$ $(1 \leq i \leq n)$ represents the *i*-th protein in the results, which is a potential homologous protein related with $q$ retrieved by three basic ranking methods; $s_1(q,p_i)$ and $s_2(q,p_i)$ represent the bitscore and E-vaule output by PSI-BLAST, respectively; $s_3(q,p_i)$ and $s_4(q,p_i)$ represent the probability and E-value output by HHblits, respectively; $s_5(q,p_i)$ is the *E*-value output by Hmmer, and the last three elements $s_6(q,p_i)$, $s_7(q,p_i)$ and $s_8(q,p_i)$ represent the reciprocal of ranking position in three ranking results. And then each element $s_j(q,p_i)$ in feature matrix Π (Equation 1) is normalized by using the following equation:

$$s'_j(q,p_i) = \frac{s_j(q,p_i)}{\max\limits_{i=1}^{n}\{s_j(q,p_i)\}} \tag{2}$$

Finally, this feature matrix is fed into the trained LTR model to reranking by using LTR algorithm. For more detailed description of LTR algorithm, please refer to (Liu *et al.*, 2015a).

The pseudo protein representation is critical for improving the sensitivity of protein remote homology detection, because it can consider the evolutionary events, such as mutations, deletions and insertions of residues in protein sequences, and find the conserved domains in structures.

## 3 Datasets

Two benchmark datasets were used to evaluate the performance of predictors: SCOP (Murzin *et al.*, 1995) and SCOPe (Fox *et al.*, 2014).

The SCOP benchmark dataset was constructed based on SCOP v1.59, which contains 7329 proteins with <95% sequence identity. It is a widely used dataset, and it can provide good comparability with other related methods (Chen *et al.*, 2016c; Liu *et al.*, 2015a; Melvin *et al.*, 2011) . There are 1073 superfamilies and 1827 families in this dataset.

The SCOPe benchmark dataset was constructed based on the SCOPe version v2.06 (Chandonia *et al.*, 2017) released on 6 April 2017 (the latest version), containing 28010 proteins with <95% sequence identity with 2008 superfamilies and 4851 families.

## 4 Performance comparison with related methods

Two performance measures were employed to evaluate the performance of each method, including ROC1 score and ROC50 score (Gribskov and Robinson, 1996). ROC1 and ROC50 scores represent the area under ROC curve up to the first false positive and the 50th false positives, respectively. A score of 1 means perfect prediction, whereas a score of 0 means that none of the proteins is correctly identified. In this study, if the detected proteins and the query protein are in the same SCOP superfamily, the detected proteins are considered as true positives, otherwise they are false positives. The jackknife validation is employed to evaluate the performance of methods, because it is deemed the most objective cross-validation approach (Chou, 2011).

Table 1 shows the performance of various methods on SCOP v1.59, from which we can see that the performance of the three predictors (PSI-BLAST, HHblits and Hmmer) can be improved by using the pseudo protein approach. *ProtDec-LTR2.0* obviously outperforms ProtDec-LTR in term of ROC1, and is highly comparable with ProtDec-LTR in term of ROC50.

In order to further evaluate its performance, *ProtDec-LTR2.0* is evaluated on the updated benchmark dataset SCOPe v2.06, and the results are shown in Figure 1, from which we can see that *ProtDec-LTR2.0* obviously outperforms the basic predictors in terms of ROC1and ROC 50.

## 5 Description of *ProtDec-LTR2.0* web server

*ProtDec-LTR2.0* web server is constructed based on SCOPe v2.06, which is compatible with most major browsers, and the parallel speed-up is implemented. As shown in the Figure 2, this web server only requires protein sequences in FASTA format as inputs. A LTR score is used to rank the detecting results, and its default threshold is 0.0, which is a tradeoff between the number false positives and the number of hits. After clicking the *Submit* button, the job will be assigned an ID and added into the processing queue. The processing

**Table 1.** The performance comparison of various methods on SCOP benchmark dataset via jackknife validation

| Methods | ROC1 | ROC50 |
|---|---|---|
| **ProtDec-LTR 2.0** | **0.8911** | 0.8955 |
| ProtDec-LTR | 0.8510 | **0.8969** |
| Pse-PSI-BLAST | 0.7893 | 0.8235 |
| Pse-HHblits | 0.8195 | 0.8885 |
| Pse-Hmmer | 0.8164 | 0.8409 |
| PSI-BLAST | 0.7718 | 0.7794 |
| HHblits | 0.8187 | 0.8669 |
| Hmmer | 0.7796 | 0.7830 |
| Coma | 0.6989 | 0.7785 |
| ProtEmbed | 0.8136 | 0.8897 |
| dRHP-PseRA | 0.8314 | 0.8924 |

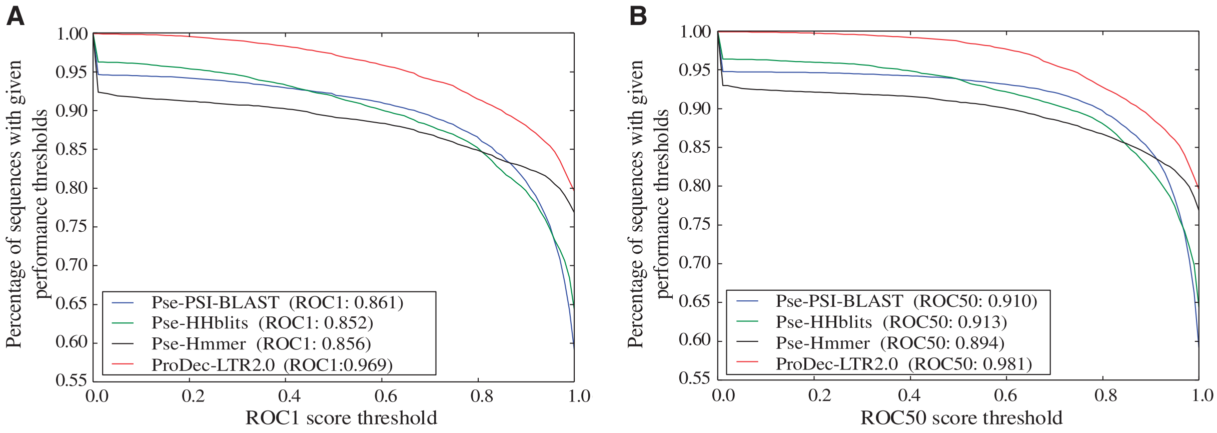
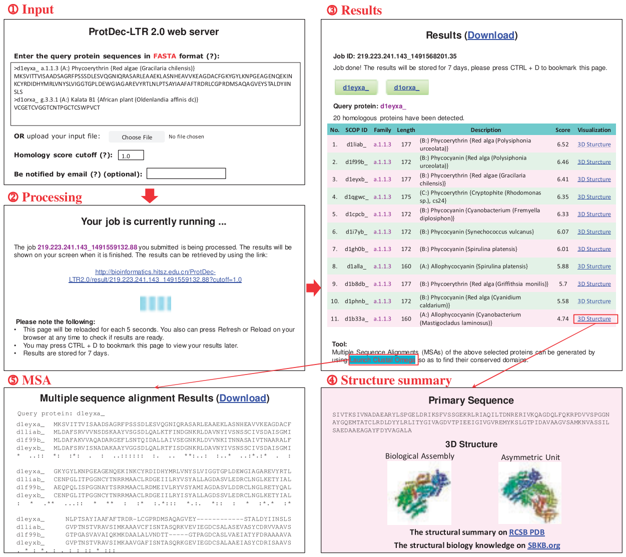


**Fig. 1.** Performance comparison of various methods on SCOPe benchmark dataset via jackknife validation. The graph plots the percentage of sequences, for which the method exceeds a given performance threshold. The higher curve means the method performs better. ROC1 and ROC50 are used as the performance measures for **(A)** and **(B)**, respectively. *ProtDec-LTR2.0* achieves the best performance with a ROC1 score of 0.969 and a ROC50 score of 0.981, obviously outperforming other methods



ProtDec-LTR 2.0 web server

Enter the query protein sequences in FASTA format (?):

```
>d1eyxa_ a.1.1.3 (A:) Phycoerythrin {Red algae (Gracilaria chilensis)}
MKSVITTVISAADSAGRFPSSSDLESVQGNIQRASARLEAAEKLASNHEAVVKEAGDACFGKYGYLKNPGEAGENQEKIN
KCYRDIDHYMRLVNYSLVIGGTGPLDEWGIAGAREVYRTLNLPTSAYIAAFAFTRDRLCGPRDMSAQAGVEYSTALDYIIN
SLS
>d1orxa_ g.3.3.1 (A:) Kalata B1 {African plant (Oldenlandia affinis dc)}
VCGETCVGGTCNTPGCTCSWPVCT
```

OR upload your input file: Choose File No file chosen

Homology score cutoff (?): 1.0

Be notified by email (?) (optional):

Your job is currently running ...

The job 219.223.241.143_1491559132.88 you submitted is being processed. The results will be shown on your screen when it is finished. The results can be retrieved by using the link:

http://bioinformatics.hitsz.edu.cn/ProtDec-LTR2.0/result/219.223.241.143_1491559132.88?cutoff=1.0

Please note the following:

- This page will be reloaded for each 5 seconds. You also can press Refresh or Reload on your browser at any time to check if results are ready.
- You may press CTRL + D to bookmark this page to view your results later.
- Results are stored for 7 days.

Results (Download)

Job ID: 219.223.241.143_1491568201.35

Job done! The results will be stored for 7 days, please press CTRL + D to bookmark this page.

d1eyxa_ d1orxa_

Query protein: d1eyxa_

20 homologous proteins have been detected.

| No. | SCOP ID | Family | Length | Description | Score | Visualization |
|---|---|---|---|---|---|---|
| 1. | d1liab_ | a.1.1.3 | 177 | (B:) Phycoerythrin {Red alga (Polysiphonia urceolata)} | 6.52 | 3D Sturcture |
| 2. | d1f99b_ | a.1.1.3 | 172 | (B:) Phycocyanin {Red alga (Polysiphonia urceolata)} | 6.46 | 3D Sturcture |
| 3. | d1eyxb_ | a.1.1.3 | 177 | (B:) Phycoerythrin {Red algae (Gracilaria chilensis)} | 6.41 | 3D Sturcture |
| 4. | d1qgwc_ | a.1.1.3 | 175 | (C:) Phycoerythrin {Cryptophite (Rhodomonas sp.), cs24} | 6.35 | 3D Sturcture |
| 5. | d1cpcb_ | a.1.1.3 | 172 | (B:) Phycocyanin {Cyanobacterium (Fremyella diplosiphon)} | 6.33 | 3D Sturcture |
| 6. | d1i7yb_ | a.1.1.3 | 172 | (B:) Phycocyanin {Synechococcus vulcanus} | 6.07 | 3D Sturcture |
| 7. | d1gh0b_ | a.1.1.3 | 172 | (B:) Phycocyanin {Spirulina platensis} | 6.01 | 3D Sturcture |
| 8. | d1alla_ | a.1.1.3 | 160 | (A:) Allophycocyanin {Spirulina platensis} | 5.88 | 3D Sturcture |
| 9. | d1b8db_ | a.1.1.3 | 177 | (B:) Phycoerythrin {Red alga (Griffithsia monilis)} | 5.7 | 3D Sturcture |
| 10. | d1phnb_ | a.1.1.3 | 172 | (B:) Phycocyanin {Red alga (Cyanidium caldarium)} | 5.58 | 3D Sturcture |
| 11. | d1b33a_ | a.1.1.3 | 160 | (A:) Allophycocyanin {Cyanobacterium (Mastigocladus laminosus)} | 4.74 | 3D Sturcture |

Tool:
Multiple Sequence Alignments (MSAs) of the above selected proteins can be generated by using Launch Clustal Omega so as to find their conserved domains.

Primary Sequence

```
SIVTKSIVNADAEARYLSPGELDRIKSFVSSGEKRLRIAQILTDNRERIVKQAGDQLFQKRPDVVSPGGN
AYGQEMTATCLRDLDYYLRLITYGIVAGDVTPIEEIGIVGVREMYKSLGTPIDAVAAGVSAMKNVASSIL
SAEDAAEAGAYFDYVAGALA
```

3D Structure

Biological Assembly Asymmetric Unit

The structural summary on RCSB PDB

The structural biology knowledge on SBKB.org

Multiple sequence alignment Results (Download)

```
Query protein: d1eyxa_

d1eyxa_   MKSVITTVISAADSAGRFPSSSDLESVQGNIQRASARLEAAEKLASNHEAVVKEAGDACF
d1liab_   MLDAFSRVVVNSDSKAAYVSGSDLQALKTFINDGNKRLDAVNYIVSNSSCIVSDAISGMI
d1f99b_   MLDAFAKVVAQADARGEFLSNTQIDALLAIVSEGNKRLDVVNKITNNASAIVTNAARALF
d1eyxb_   MLDAFSRVISNADAKAAYVGGSDLQALRTFISDGNKRLDAVNYIVSNSSCIVSDAISGMI
*  ..::  *:  :*:  .  :  ..::::::  :.  ..  **:..:  :..*  ..:*.:*  .  :

d1eyxa_   GKYGYLKNPGEAGENQEKINKCYRDIDHYMRLVNYSLVIGGTGPLDEWGIAGAREVYRTL
d1liab_   CENPGLITPGGNCYTNRRMAACLRDGEIILRYVSYALLAGDASVLEDRCLNGLKETYIAL
d1f99b_   AEQPQLISPGGNAYTSRRMAACLRDMEIVLRYVSYAMIAGDASVLDDRCLNGLRETYQAL
d1eyxb_   CENPGLITPGGNCYTNRRMAACLRDGEIILRYISYALLAGDSSVLEDRCLNGLKETYIAL
:  *  .**  ...::  *  **  :  :*  :.*:::  *.:.  *::  :  *  :*.*  :*

d1eyxa_      NLPTSAYIAAFAFTRDR-LCGPRDMSAQAGVEY------------STALDYIINSLS
d1liab_      GVPTNSTVRAVSIMKAAAVCFISNTASQRKVEVIEGDCSALASEVASYCDRVVAAVS
d1f99b_      GTPGASVAVAIQKMKDAALALVNDTT-----GTPAGDCASLVAEIATYFDRAAAAVA
d1eyxb_      GVPTNSTVRAVSIMKAAVGAFISNTASQRKGEVIEGDCSALAAEIASYCDRISAAVS
. * : *. : . : : :: * :::
```

**Fig. 2.** The workflow for detecting remote homologous proteins by using *ProtDec-LTR2.0* web server. *ProtDec-LTR2.0* cannot only detect the remote homologous proteins but also provide result visualization and interpretation functions, such as homologous protein 3D structure visualization and multiple sequence alignment interpretation

page will be reloaded for each 5s, and the results will be shown on your screen when the processing is finished or sent you by email if provided. On the result page, the detected remote homologous proteins are shown in a table, and they are sorted in descending order by LTR scores. The sequence and structure information of the detected proteins are provided as well. The user can access to their protein primary sequence and 3D structure by clicking the 3D Structure button in Visualization column of the table. Besides, the user can also easily find detailed structure summary and biology knowledge by clicking the RCSB PDB (Rose *et al.*, 2017) button and PSI SBKB (Gabanyi *et al.*, 2011) button, and find the conserved domains by launching a multiple sequences alignment tool Clustal Omega (Larkin *et al.*, 2007).

## Acknowledgements

We wish to thank the three anonymous reviewers, whose constructive comments were very helpful for strengthening the presentation of this article.

## Funding

This work was supported by the National Natural Science Foundation of China (No. 61672184), the Natural Science Foundation of Guangdong Province (2014A030313695), Guangdong Natural Science Funds for Distinguished Young Scholars (2016A030306008), and Scientific Research Foundation in Shenzhen (grant no. JCYJ20150626110425228).

*Conflict of Interest*: none declared.

## References

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Chandonia,J.-M. *et al.* (2017) SCOPe: Manual curation and artifact removal in the structural classification of proteins–extended database. *J. Mol. Biol.*, **429**, 348–355.

Chen,J. *et al.* (2016a) A comprehensive review and comparison of different computational methods for protein remote homology detection, *Brief. Bioinformatics*, 10.1093/bib/bbw1108.

Chen,J. *et al.* (2016b) Protein remote homology detection based on an ensemble learning approach. *BioMed. Res. Int.*, **2016**, 5813645.

Chen,J. *et al.* (2016c) dRHP-PseRA: detecting remote homology proteins using profile-based pseudo protein sequence and rank aggregation. *Sci. Rep.*, **6**, 32333.

Chen,W. *et al.* (2017) Detecting N6-methyladenosine sites from RNA transcriptomes using ensemble Support Vector Machines. *Sci. Rep.*, **7**, 40242.

Chou,K.-C. (2011) Some remarks on protein attribute prediction and pseudo amino acid composition. *J. Theor. Biol.*, **273**, 236–247.

Fox,N.K. *et al.* (2014) SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.*, **42**, D304–D309.

Gabanyi,M.J. *et al.* (2011) The Structural Biology Knowledgebase: a portal to protein structures, sequences, functions, and methods. *J. Struct. Funct. Genomics*, **12**, 45–54.

Gribskov,M. and Robinson,N.L. (1996) Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput. Chem.*, **20**, 25–33.

Larkin,M.A. *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.

Liu,B. *et al.* (2016) Protein Remote Homology Detection by Combining Pseudo Dimer Composition with an Ensemble Learning Method. *Current Proteomics*, **13**, 86–91.

Liu,B. *et al.* (2015a) Application of Learning to Rank to protein remote homology detection. *Bioinformatics*, **31**, 3492–3498.

Liu,B. *et al.* (2015b) Protein remote homology detection by combining Chou's distance-pair pseudo amino acid composition and principal component analysis. *Molecular Genetics and Genomics*, **290**, 1919–1931.

Liu,B. *et al.* (2008) A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis. *BMC Bioinformatics*, **9**, 510.

Liu,B. *et al.* (2014a) Using distances between Top-n-gram and residue pairs for protein remote homology detection. *BMC Bioinformatics*, **15**, S3.

Liu,B. *et al.* (2014b) Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics*, **30**, 472–479.

Melvin,I. *et al.* (2011) Detecting remote evolutionary relationships among proteins by large-scale semantic embedding. *PLoS Comput. Biol.*, **7**, e1001047.

Mistry,J. *et al.* (2013) Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.*, **41**, e121-e121.

Murzin,A.G. *et al.* (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Rose,P.W. *et al.* (2017) The RCSB protein data bank: integrative view of protein, gene and 3D structural information. *Nucleic Acids Res.*, **45**, D271–D281.

Wei,L. *et al.* (2015) Enhanced protein fold prediction method through a novel feature extraction technique. *IEEE Trans. Nanobiosci.*, **14**, 649–659.

Wei,L. and Zou,Q. (2016) Recent progresses in machine learning-based methods for protein fold recognition. *Int. J. Mol. Sci.*, **17**, 2118.