Bioinformatics, 37(7), 2021, 913–920
doi: 10.1093/bioinformatics/btaa772
Advance Access Publication Date: 8 September 2020
Original Paper

Sequence analysis

SMI-BLAST: a novel supervised search framework based on PSI-BLAST for protein remote homology detection

Xiaopeng Jin[1], Qing Liao[1], Hang Wei[1], Jun Zhang[1] and Bin Liu[1,2,3,*]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen, Guangdong 518055, China[2]School of Computer Science and Technology and [3]Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing 100081, China

*To whom correspondence should be addressed.

Associate Editor: Xu Jinbo

Received on May 21, 2020; revised on August 14, 2020; editorial decision on August 27, 2020; accepted on August 28, 2020

Abstract

Motivation: As one of the most important and widely used mainstream iterative search tool for protein sequence search, an accurate Position-Specific Scoring Matrix (PSSM) is the key of PSI-BLAST. However, PSSMs containing non-homologous information obviously reduce the performance of PSI-BLAST for protein remote homology.

Results: To further study this problem, we summarize three types of Incorrectly Selected Homology (ISH) errors in PSSMs. A new search tool Supervised-Manner-based Iterative BLAST (SMI-BLAST) is proposed based on PSI-BLAST for solving these errors. SMI-BLAST obviously outperforms PSI-BLAST on the Structural Classification of Proteins-extended (SCOPe) dataset. Compared with PSI-BLAST on the ISH error subsets of SCOPe dataset, SMI-BLAST detects 1.6–2.87 folds more remote homologous sequences, and outperforms PSI-BLAST by 35.66% in terms of ROC1 scores. Furthermore, this framework is applied to JackHMMER, DELTA-BLAST and PSI-BLASTexB, and their performance is further improved.

Availability and implementation: User-friendly webservers for SMI-BLAST, JackHMMER, DELTA-BLAST and PSI-BLASTexB are established at http://bliulab.net/SMI-BLAST/, by which the users can easily get the results without the need to go through the mathematical details.

Contact: bliu@bliulab.net

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Protein remote homology detection is an increasingly important task in analysing protein structures and functions (Chen *et al.*, 2018). With the rapid growth of protein sequences, more and more computational methods have been proposed to address this important and challenging problem.

Among those methods, PSI-BLAST (Altschul, 1997) is one of the most widely used and famous tools in this field. It represents the protein sequence as the Position-Specific Scoring Matrix (PSSM) in an iteration manner, leading to better performance than sequence alignment methods, such as BLAST (Altschul *et al.*, 1990) and FASTA (Pearson, 1990; Pearson and Lipman, 1988). Because of its effectiveness and accuracy, several improvements for PSI-BLAST have been proposed. For e-value score, some methods re-calculate or replace it for more accurate ranking list. Because early iterations are highly specific, and later iterations are sensitive for weaker remote homologs, SIB-BLAST (Lee *et al.*, 2008) re-calculates e-value by combining the second ranking list and the final ranking list to improve the accuracy of PSI-BLAST. PSI-BLASTFDR (Carroll *et al.*, 2015) replaces the e-value threshold with false discovery rate (FDR) to improve retrieval performance. For PSSM profile, PSI-BLASTexB (Oda *et al.*, 2017) points out that the narrow block in multiple sequence alignment (MSA) leads to an inaccurate PSSM, and it solves this problem by setting the minimum block width (MBW) in PSI-BLAST. CS-BLAST (Biegert and Soding, 2009) constructs a new context-specific amino acid similarities to replace the PSSM for higher sensitivity. DELTA-BLAST (Boratyn *et al.*, 2012) searches a database with a new constructed PSSM, which contains the information of conserved domain after using RPS-BLAST (Marchler-Bauer, 2002) to search on Conserved Domain Database (Marchler-Bauer, 2011). On the library of complete protein sequences, PSI-SEARCH2 (Pearson *et al.*, 2017) solves the homologous over-extension (HOE) errors (Gonzalez and Pearson, 2010) of PSSM by inserting the original query sequence residues into gapped positions in the aligned subject sequences. Therefore, the frequency of false-positive alignments is reduced by 5–20 folds compared with PSI-BLAST and JackHMMER. HHblits (Remmert *et al.*, 2012) and HMMER (Eddy, 2011) are based on HMM profile alignments rather than PSSM, which is more sensitive because HMM profile considers not only emission probability but also state transition probability. JackHMMER (Johnson *et al.*, 2010) and HHblits

(Remmert *et al.*, 2012) also use different iterative strategies to enhance the ability of detecting remote homology protein. Those methods are combined to further improve the performance, such as SAM-HMMER (Wistrand and Sonnhammer, 2005), CHASE (Alam *et al.*, 2004) and ProtDec-LTR (Liu *et al.*, 2015a).

For databases with whole protein sequences, such as UniProt dataset (The UniProt, 2017), the most important problems of PSI-BLAST are non-homologous false-positive (FP) alignments (NH-FP) and HOE FPs defined by Gonzalez and Pearson (2010). HOE errors severely affect the performance of PSI-BLAST on the databases with whole proteins. Fortunately, PSI-SEARCH series (Li *et al.*, 2012; Pearson *et al.*, 2017; Yang *et al.*, 2019) have solved the HOE errors with excellent performance. However, unrelated protein domain errors of NH-FP become the main errors in the search process of PSI-BLAST on protein domain databases, which are widely used to evaluate the performance of protein remote detection and fold recognition (Remmert *et al.*, 2012; Senior *et al.*, 2020; Yan *et al.*, 2019). To fully study the non-homologous protein errors in protein domain databases, in this study, we make an attempt to solve the non-homologous protein errors based on the PSI-BLAST search list using a Supervised-Manner-based Iterative framework (SMI-BLAST).

Based on the analysis of PSI-BLAST search results and non-homologous protein errors, we summarize three situations as incorrectly Selected Homology (ISH) errors. ISH errors indicate that true positives (TPs) exist in the ranking list but the selected list is null or contains false positives. Figure 1 shows three types of ISH errors and other situations of PSSM: (i) True-PSSM (Fig. 1A). PSSM is constructed by TPs, which is an ideal situation for PSSM and can describe the correct evolutionary information of query sequences; (ii) ISH-MIX error (Fig. 1B). The selected list contains false positives and TPs, leading to more false positives detected at later iterations; (iii) ISH-NULL error (Fig. 1C). No sequence in the selected list can be used to construct PSSM but there are TPs in the candidate list; (iv) ISH-ALL error (Fig. 1D). The sequences in the selected list are all false positives but TPs exist in the candidate list. For PSI-BLAST with ISH-ALL error, it is hard to detect the TPs at later iterations and (v) False-PSSM (Fig. 1E). The ranking list contains no TP, and therefore there is no more adjustment space for PSSM. To construct and keep an ideal situation during the iteration process, the above ISH errors of PSSM should be solved.

In this study, we propose a framework (SMI-BLAST) to correct ISH errors by embedding Supervised-Manner-based Iterative framework (SMI-based framework) in PSI-BLAST. SMI-BLAST can not only correct the ISH errors of PSI-BLAST, but also can improve its performance for protein remote homology detection. Furthermore, SMI-based framework is successfully applied to JackHMMER (Johnson *et al.*, 2010), DELTA-BLAST (Boratyn *et al.*, 2012) and PSI-BLASTexB (Oda *et al.*, 2017).

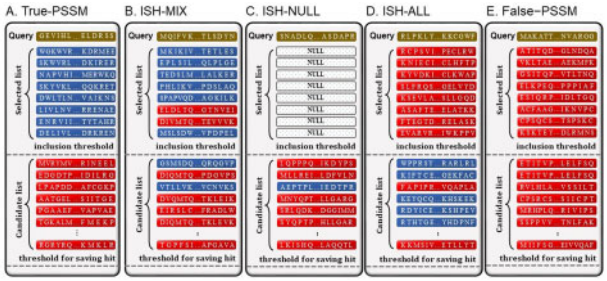

Fig. 1. The five situations of PSI-BLAST selecting sequences to construct PSSM profile for protein remote homology detection. Brown bars, blue bars and red bars within black rectangle represent query sequences, homologous sequences (in the same superfamily) and non-homologous sequences (not in the same superfamily), respectively. Blue bars and red bars constitute the ranking list of search results. The grey dotted lines represent the inclusion threshold for next alignments (default e-value = 0.002) and the threshold for saving hit (default e-value = 10) in PSI-BLAST (Altschul, 1997). Selected lists represent those sequences are selected to construct PSSM for the next iteration when their e-value scores are lower than inclusion threshold of PSI-BLAST. Candidate lists represent that those sequences are abandoned in the next iteration when their scores are between inclusion threshold and threshold of PSI-BLAST.(Color version of this figure is available at *Bioinformatics* online.)

2 Materials and methods

2.1 Benchmark dataset

For evaluating the performance of SMI-BLAST, the Structural Classification of Proteins-extended (SCOPe2.06) dataset (Chandonia, 2019) is used, which is a golden benchmark for protein remote homology detection with less than 95% identity to each other. According to structural and evolutionary relationships, 28 010 protein sequences in SCOPe are classified into the following five hierarchy levels: protein, family, superfamily, fold and class.

For analysing the ISH errors of PSSM, those sequences suffering from ISH errors after the first iteration of PSI-BLAST are separated from SCOPe benchmark dataset. ISH error subset is represented as:

$$\mathbb{S}^{\mathrm{ISH}} = \mathbb{S}^{\mathrm{ISH}}_{\mathrm{MIX}} \cup \mathbb{S}^{\mathrm{ISH}}_{\mathrm{NULL}} \cup \mathbb{S}^{\mathrm{ISH}}_{\mathrm{ALL}}, \tag{1}$$

where those subsets represent the three ISH error situations shown in Figure 1. (i) ISH error subset $\mathbb{S}^{\mathrm{ISH}}_{\mathrm{MIX}}$ with 229 sequences: the selected lists of those sequences contain TPs and false positives. (ii) ISH error subset $\mathbb{S}^{\mathrm{ISH}}_{\mathrm{NULL}}$ with 907 sequences: the selected lists of those sequences contain no sequence, but TPs exit in the result list. (iii) ISH error subset $\mathbb{S}^{\mathrm{ISH}}_{\mathrm{ALL}}$ with 6 sequences: the selected lists of those sequences only contain false positives, but TPs exit in the result list.

For the total SCOPe benchmark dataset in this study, it consisted of ISH error subset $\mathbb{S}^{\mathrm{ISH}}$ and other sequences that belong to the other two situations of PSSM. The SCOPe benchmark dataset is represented as:

$$\mathbb{S}_{\mathrm{scope}} = \mathbb{S}^{\mathrm{ISH}} \cup \mathbb{S}^{\mathrm{T}} \cup \mathbb{S}^{\mathrm{F}}, \tag{2}$$

where those two subsets represent the corresponding situations shown in Figure 1. (i) Subset $\mathbb{S}^{\mathrm{T}}$ with 25 014 sequences: the selected lists of those sequences after the first iteration are all TPs. (ii) Subset $\mathbb{S}^{\mathrm{F}}$ with 1854 sequences: their ranking lists after the first iteration contain no TP, meaning that there are no TPs for constructing correct PSSM from those ranking lists.

2.2 Flowchart of SMI-BLAST

The flowchart of SMI-based framework is shown in Figure 2. The flowchart contains four models: (i) first unsupervised PSI-BLAST model is used to produce initial search result with 1 iteration; (ii) the first supervised learning to rank model (Li, 2011) LTR-First is used to solve the ISH errors by selecting a more accurate selected list. The false positives used to construct PSSMs are obviously reduced by LTR-First model; (iii) the second unsupervised PSI-BLAST model is used to detect more homology protein sequences with the corrected

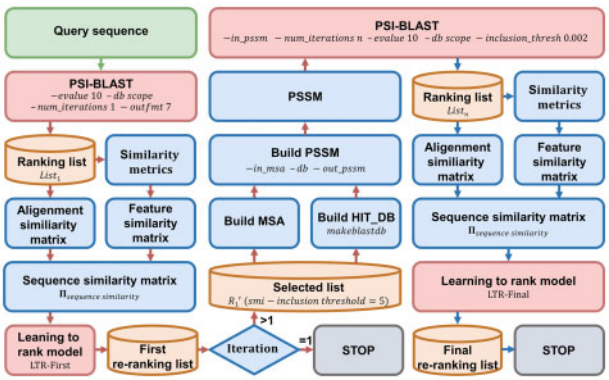

Fig. 2. The flowchart of SMI-BLAST. The red arrows show how to adjust PSSM to solve the ISH errors. The blue arrows show re-sorting the final ranking list by learning to rank model based on sequence similarity matrix when the number of iteration is greater than 1. (Color version of this figure is available at *Bioinformatics* online.)

PSSM; (iv) the second supervised learning-to-rank model LTR-Final is used to re-rank the final ranking list from previous step for more accurate results. Although more accurate PSSM can be obtained by LTR-First, the final ranking lists should be re-ranked by LTR-Final. The reason is that the True-PSSM cannot guarantee the accuracy of ranking list. Furthermore, sequence similarity matrix is the most critical module for the two learning-to-rank models, because it describes the similarity of ranking list and query sequence from various aspects.

2.3 Sequence similarity matrix construction

How to measure the similarity of remote homology protein sequence pairs is a challenging task. PSI-BLAST relies on profile alignment scores. Therefore, its ranking lists are confined to sequence alignment information. The ranking list of PSI-BLAST can be represented as:

$$\mathrm{List}_n(q) = \{p_1,\ p_2, \ldots,\ p_l\}, \tag{3}$$

where n represents the nth iteration, q represents the query protein, p_i represents the ith feedback protein in the ranking list and $e_{value}(q, p_i) \leq e_{value}(q, p_{i+1})$, l represents the number of feedback proteins (default $0 \leq l \leq 500$).

In SMI-BLAST, the $l \times 89$ sequence similarity matrix is constructed to describe the similarity of query sequences and feedback sequences. This matrix is composed of one alignment similarity matrix and four feature similarity matrices, defined as:

$$\boldsymbol{\Pi}_{\text{sequence similarity}} = [\boldsymbol{\Pi}_{\mathrm{ALI}}\boldsymbol{\Pi}_{\mathrm{AAC}}\boldsymbol{\Pi}_{\mathrm{ACR}}\boldsymbol{\Pi}_{\mathrm{PSEAAC}}\boldsymbol{\Pi}_{\mathrm{PROFILE}}], \tag{4}$$

where $l \times 5$ $\boldsymbol{\Pi}_{\mathrm{ALI}}$ is alignment similarity matrix, $l \times 18$ $\boldsymbol{\Pi}_{\mathrm{AAC}}$ is amino acid composition feature similarity matrix, $l \times 24$ $\boldsymbol{\Pi}_{\mathrm{ACR}}$ is autocorrelation feature similarity matrix, $l \times 12$ $\boldsymbol{\Pi}_{\mathrm{PSEAAC}}$ is pseudo amino acid composition feature similarity matrix, $l \times 30$ $\boldsymbol{\Pi}_{\mathrm{PROFILE}}$ is profile-based feature similarity matrix.

For $l \times 5$ alignment similarity matrix, the alignment scores are extracted from the ranking list generated by PSI-BLAST, defined as:

$$\boldsymbol{\Pi}_{\mathrm{ALI}} = \begin{bmatrix} \varphi_{\mathrm{i}}(\mathrm{q},\ \mathrm{p}_1) & \varphi_{\mathrm{e}}(\mathrm{q},\ \mathrm{p}_1) & \varphi_{\mathrm{b}}(\mathrm{q},\ \mathrm{p}_1) & \varphi_{\mathrm{al}}(\mathrm{q},\ \mathrm{p}_1) & \varphi_{\mathrm{o}}(\mathrm{q},\ \mathrm{p}_1) \\ \varphi_{\mathrm{i}}(\mathrm{q},\ \mathrm{p}_2) & \varphi_{\mathrm{e}}(\mathrm{q},\ \mathrm{p}_2) & \varphi_{\mathrm{b}}(\mathrm{q},\ \mathrm{p}_2) & \varphi_{\mathrm{al}}(\mathrm{q},\ \mathrm{p}_2) & \varphi_{\mathrm{o}}(\mathrm{q},\ \mathrm{p}_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \varphi_{\mathrm{i}}(\mathrm{q},\ \mathrm{p}_l) & \varphi_{\mathrm{e}}(\mathrm{q},\ \mathrm{p}_l) & \varphi_{\mathrm{b}}(\mathrm{q},\ \mathrm{p}_1) & \varphi_{\mathrm{al}}(\mathrm{q},\ \mathrm{p}_l) & \varphi_{\mathrm{o}}(\mathrm{q},\ \mathrm{p}_l) \end{bmatrix}, \tag{5}$$

where $\varphi_{\mathrm{i}}(q,\ p_l)$, $\varphi_{\mathrm{e}}(q,\ p_l)$, $\varphi_{\mathrm{b}}(q,\ p_1)$, $\varphi_{\mathrm{al}}(q,\ p_l)$ represent identity, e-value, bit score, align length calculated by PSI-BLAST, respectively. $\varphi_{\mathrm{o}}(q,\ p_l)$ represents the reciprocal of ranking position in ranking list of PSI-BLAST.

For $l \times 84$ feature similarity matrix, six similarity metrics are used to calculate the feature similarity scores between query sequence and feedback sequences in the ranking list. Those features of query sequences and feedback sequences are extracted by Pse-in-One (Liu *et al.*, 2015b), which are useful for protein sequence analysis problems (Zou *et al.*, 2016). The feature similarity matrix is divided into four sub-matrices according to different feature types defined as:

$$\boldsymbol{\Pi}_{\mathrm{AAC}} = \begin{bmatrix} \phi_{\mathrm{Kmer}}(q,\ p_1) & \phi_{\mathrm{DR}}(q,\ p_1) & \phi_{\mathrm{DP}}(q,\ p_1) \\ \phi_{\mathrm{Kmer}}(q,\ p_2) & \phi_{\mathrm{DR}}(q,\ p_2) & \phi_{\mathrm{DP}}(q,\ p_2) \\ \vdots & \vdots & \vdots \\ \phi_{\mathrm{Kmer}}(q,\ p_l) & \phi_{\mathrm{DR}}(q,\ p_{\ l}) & \phi_{\mathrm{DP}}(q,\ p_l) \end{bmatrix}, \tag{6}$$

$$\boldsymbol{\Pi}_{\mathrm{ACR}} = \begin{bmatrix} \phi_{\mathrm{AC}}(q,\ p_1) & \phi_{\mathrm{CC}}(q,\ p_1) & \phi_{\mathrm{ACC}}(q,\ p_1) & \phi_{\mathrm{PDT}}(q,\ p_1) \\ \phi_{\mathrm{AC}}(q,\ p_2) & \phi_{\mathrm{CC}}(q,\ p_2) & \phi_{\mathrm{ACC}}(q,\ p_2) & \phi_{\mathrm{PDT}}(q,\ p_1) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_{\mathrm{AC}}(q,\ p_l) & \phi_{\mathrm{CC}}(q,\ p_{\ l}) & \phi_{\mathrm{ACC}}(q,\ p_l) & \phi_{\mathrm{PDT}}(q,\ p_l) \end{bmatrix}, \tag{7}$$

$$\boldsymbol{\Pi}_{\mathrm{PSEAAC}} = \begin{bmatrix} \phi_{\mathrm{PC}_{\mathrm{PSEAAC}}}(q,\ p_1) & \phi_{\mathrm{SC}_{\mathrm{PSEAAC}}}(q,\ p_1) \\ \phi_{\mathrm{PC}_{\mathrm{PSEAAC}}}(q,\ p_2) & \phi_{\mathrm{SC}_{\mathrm{PSEAAC}}}(q,\ p_2) \\ \vdots & \vdots \\ \phi_{\mathrm{PC}_{\mathrm{PSEAAC}}}(q,\ p_l) & \phi_{\mathrm{SC}_{\mathrm{PSEAAC}}}(q,\ p_{\ l}) \end{bmatrix}, \tag{8}$$

$$\boldsymbol{\Pi}_{\mathrm{PROFILE}} = \begin{bmatrix} \phi_{\mathrm{AP}}(q,\ p_1) & \phi_{\mathrm{T1}}(q,\ p_1) & \phi_{\mathrm{T2}}(q,\ p_1) & \phi_{\mathrm{PP}}(q,\ p_1) & \phi_{\mathrm{DT}}(q,\ p_1) \\ \phi_{\mathrm{AP}}(q,\ p_2) & \phi_{\mathrm{T1}}(q,\ p_2) & \phi_{\mathrm{T2}}(q,\ p_2) & \phi_{\mathrm{PP}}(q,\ p_1) & \phi_{\mathrm{DT}}(q,\ p_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \phi_{\mathrm{AP}}(q,\ p_l) & \phi_{\mathrm{T1}}(q,\ p_{\ l}) & \phi_{\mathrm{T2}}(q,\ p_l) & \phi_{\mathrm{PP}}(q,\ p_1) & \phi_{\mathrm{DT}}(q,\ p_l) \end{bmatrix}, \tag{9}$$

where AAC represents the amino acid composition based on Kmer (Liu *et al.*, 2008), DR (Liu *et al.*, 2014b) and DP (Liu *et al.*, 2014a). ACR represents the autocorrelation based on AC, CC, ACC (Dong *et al.*, 2009; Guo *et al.*, 2008) and PDT (Liu *et al.*, 2012). PSEAAC represents the pseudo amino acid composition based on PC-PseAAC (Chou, 2001), SC-PseAAC (Chou, 2005). PROFILE represents the profile-based features. AP, T1, T2, PP, DT represent features based on Top-1-gram (Liu *et al.*, 2008), Top-2-gram (Liu *et al.*, 2008), ACC-PSSM (Dong *et al.*, 2009), PDT-Profile (Liu *et al.*, 2012), DT (Liu *et al.*, 2014a). $\phi_m(q,p_l)$ represents the feature similarity score, m represents the feature type. $\phi_m(q,p_l)$ can be calculated by:

$$\phi_m(q,p_l) = \begin{bmatrix} \sqrt{\sum_{i=1}^{N}(f_{q,i}^m - f_{p_l,i}^m)^2} \\ \sum_{i=1}^{N} f_{q,i}^m - f_{p_l,i}^m \\ \dfrac{\sum_{i=1}^{N}\left(f_{q,i}^m - -f_q^m\right)\cdot\left(f_{p_l,i}^m - -f_{p_l}^m\right)}{\sqrt{\sum_{i=1}^{N}\left(f_{q,i}^m - -f_q^m\right)^2}\cdot\sqrt{\sum_{i=1}^{N}\left(f_{p_l,i}^m - -f_{p_l}^m\right)^2}} \\ max_l f_{q,i}^m - f_{p_l,i}^m \\ \dfrac{\sum_{i=1}^{N} f_{q,i}^m \cdot f_{p_l,i}^m}{\sqrt{\sum_{i=1}^{N}\left(f_{q,i}^m\right)^2}\cdot\sqrt{\sum_{i=1}^{N}\left(f_{p_l,i}^m\right)^2}} \\ \dfrac{\sum_{i=1}^{N} f_{q,i}^m - f_{p_l,n}^m}{\sum_{i=1}^{N} f_{q,i}^m + \sum_{i=1}^{N} f_{p_l,i}^m} \end{bmatrix}^{\mathrm{T}}, \tag{10}$$

where $f_{q,i}^m$ and $f_{p,i}^m$ represent the features of query sequence and feedback sequence, i $(0 < i \leq N)$ represents the ith location of N-dimensional feature. According to the order in matrix $\phi_m(q,p_l)$, Euclidean distance (Danielsson, 1980), Manhattan distance (Borgefors, 1984), Pearson's Correlation score (Lee Rodgers and Nicewander, 1988), Chebyshev distance (Klove *et al.*, 2010), Cosine similarity (Singhal, 2001) and Bray Curtis dissimilarity (Somerfield, 2008) are used to measure the similarity scores between query sequence's feature and feedback sequence's feature. These similarity metrics have been widely used in biological sequence analysis problems (Bass *et al.*, 2013; Hou *et al.*, 2018).

2.4 Training the two learning-to-rank models

In sequence similarity matrix, 89 types of sequence similarity scores are calculated to represent the sequence pairs' similarity. How to re-rank the result list based on those similarity scores is important for generating a more accurate ranking list. Learning to rank (LTR) algorithms solve the ranking task in a supervised manner, which are

widely used in Information Retrieval (IR) and Natural Language Processing (NLP) (Li, 2011). LambdaMART (Burges, 2010) combining LambdaRank (Burges *et al.*, 2006) and MART (Multiple Additive Regression Trees) is used to train the two LTR models (LTR-First and LTR-Final) in SMI-BLAST with default parameters of RankLib-2.10 (Burges *et al.*, 2005) except that the loss function is set as Normalized Discounted Cumulative Gain (Donmez *et al.*, 2009). To improve the generalizability of SMI-BLAST, the two LTR models are trained and evaluated by fivefold cross-validation (Bengio and Grandvalet, 2004) (see Supplementary Fig. S1 in Supplementary Data). In other words, the training samples and test samples of the LTR models are fully independent. It should be noticed that the two LTR models are independently trained, and the PSSMs generated by LTR-First are fed into the LTR-Final for the final prediction.

2.5 PSSM construction and iterative search of SMI-BLAST

To complete the PSSM adjustment, PSSM should be constructed according to the selected list $R_1{}'$ which is offered by LTR-First. First, unabridged multiple sequence alignment from $List_1$ is constructed. It contains all the detailed alignment information between the query sequence and feedback sequences. Second, only the sequences in multiple sequence alignments with the same id with $R_1{}'$ are retained. Third, the original sequences with the same id as $R_1{}'$ are used to construct the HIT_DB by *makeblastdb* command. Finally, similar command with PSI-SEARCH (Pearson *et al.*, 2017) from PSI-BLAST 2.7.1+ is used to construct PSSM with MSA and HIT_DB.

For later iteration search, the new constructed PSSM in asntxt format is used as input to run *psiblast* 2.7.1+ program. When the final ranking lists are produced, LTR-Final model is used to re-sort it.

2.6 Evaluation

To evaluate the improvement of different methods, the ranking quality and the ability of detecting remote homology sequences are measured. In this study, if the feedback protein sequences and the query protein sequence are in the same superfamily, the feedback protein sequences are considered as TPs, otherwise false positives (Chen *et al.*, 2017). For the ranking quality, ROC1 and ROC50 scores are used. If ROC1 or ROC50 score is 1, it indicates the ranking list gets a perfect ranking. For the ability of detecting remote homology sequences, the TP number and coverage in the same Errors Per Query (EPQ) score (Reid *et al.*, 2007) are used as evaluation metrics. TP number represents the number of detected homology proteins. Coverage represents the proportion of the detected homologous proteins in the total query protein's superfamilies. EPQ represents the proportion of detected non-homology proteins in the total detected proteins in the search results.

Table 1. The performance of PSI-BLAST with different PSSM situations after two iterations

Dataset	Method	ROC1[a]	ROC50[a]	TP number[a]
$\mathbb{S}^{T}$	PSI-BLAST[b]	0.9225	0.9636	70.7306
$\mathbb{S}^{ISH}_{MIX}$	PSI-BLAST[b]	0.6432	0.8747	31.9432
$\mathbb{S}^{ISH}_{NULL}$	PSI-BLAST[b]	0.4268	0.6723	3.7067
$\mathbb{S}^{ISH}_{ALL}$	PSI-BLAST[b]	0.0000	0.2989	6.6667

[a]Performance at homology level that contains close homology and remote homology (belonging to the same SCOP superfamily).

[b]The results of PSI-BLAST are obtained by PSI-BLAST 2.7.1+ (https://blast.ncbi.nlm.nih.gov/Blast.cgi). It is worth noting that PSI-BLAST 2.7.1+ can produce some homology protein sequences on set $\mathbb{S}^{ISH}_{NULL}$ and set $\mathbb{S}^{ISH}_{ALL}$.

3 Results

3.1 Incorrectly Selected Homology errors obviously decrease the performance of PSI-BLAST

As introduced in Section 1, PSI-BLAST is suffering from the ISH errors. In this section, we study the influence of these errors on the performance of PSI-BAST. We analyse the performance of PSI-BLAST on the true PSSM subset $\mathbb{S}^{T}$ and ISH error subsets $\mathbb{S}^{ISH}_{MIX}$, $\mathbb{S}^{ISH}_{NULL}$ and $\mathbb{S}^{ISH}_{ALL}$ (cf. Eq. 1), and the predictive results are shown in Table 1, from which we can see: (i) PSI-BLAST achieves good performance in terms of ranking quality and detected true-positive number on True PSSM subset; (ii) compared with the results of PSI-BLAST on True PSSM subset, the ranking quality and detected number of TP obviously decrease on ISH error subsets. Based on the above results, we conclude that the performance of PSI-BLAST obviously decreases because of the ISH errors.

To directly show the effect of ISH errors on PSI-BLAST, an example is given in Figure 3. SMI-BLAST achieves good performance at the second iteration (Fig. 3C) after extracting true homology protein sequence to construct true PSSM (Fig. 3A and B). However, the performance of PSI-BLAST 2.7.1+ (https://blast.ncbi.nlm.nih.gov/Blast.cgi) at the second iteration is extremely low (Fig. 3F) because the PSSM of PSI-BLAST suffers from ISH-NULL error (Fig. 3D and E).

3.2 SMI-BLAST outperforms PSI-BLAST by solving the Incorrectly Selected Homology errors

As can been seen from Figure 4A and B, compared with PSI-BLAST, SMI-BLAST improves the performance by 35.66% in terms of average ROC1, and detects 1.6–2.87-folds TPs on the ISH error subset $\mathbb{S}^{ISH}$ (cf. Eq. 1).

The most obvious performance improvements by SMI-BLAST are on set $\mathbb{S}^{ISH}_{NULL}$ and set $\mathbb{S}^{ISH}_{ALL}$ (Table 2). On set $\mathbb{S}^{ISH}_{NULL}$, more than 3-fold TPs are detected and nearly 2-fold improvement in terms of ROC1. The homology proteins of more than half of the query sequences are correctly detected (Fig. 4C and D), and more ranking lists do not contain any false positive (Fig. 4F). On set $\mathbb{S}^{ISH}_{ALL}$, the ranking lists of PSI-BLAST are obviously improved by SMI-BLAST. The top hit in more than half of all the ranking lists is

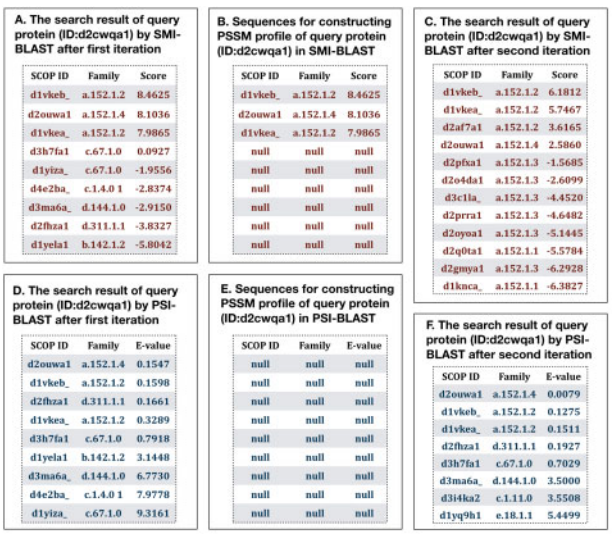

Fig. 3. The results of query protein (SCOP ID: d2cwqa1 and Family: a.152.1.4) in the iteration process of SMI-BLAST and PSI-BLAST 2.7.1+ (https://blast.ncbi.nlm.nih.gov/Blast.cgi). (A) and (D) The result of SMI-BLAST and PSI-BLAST after the first iteration, respectively. (B) There are three protein sequences providing alignment information for PSSM profile in SMI-BLAST. (C) The search result of SMI-BLAST after the second iteration. (E) There is no sequence providing alignment information for PSSM profile in PSI-BLAST. (F) The search result of PSI-BLAST after the second iteration

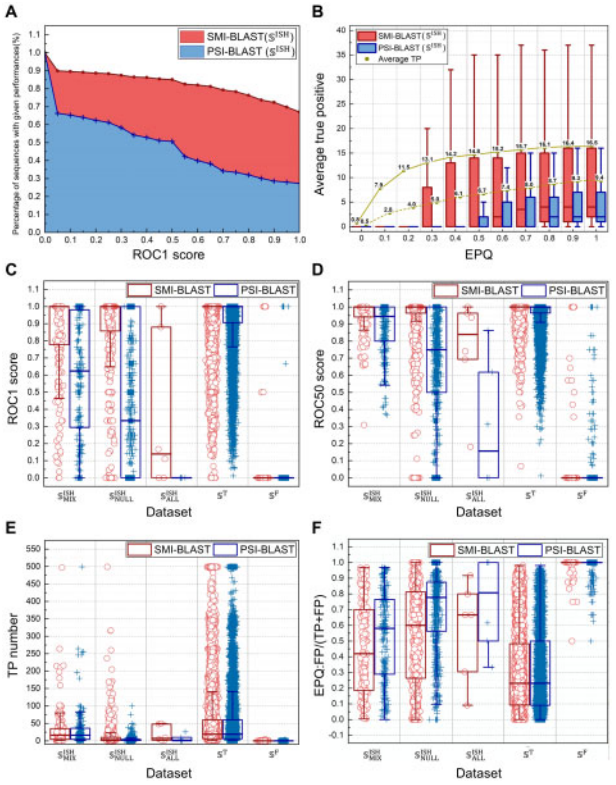

Fig. 4. SMI-BLAST (○) outperforms PSI-BLAST (+) after two iterations on the benchmark dataset. Comparison between SMI-BLAST and PSI-BLAST in terms of ROC1 score (**A**) on the ISH error subset $\mathbb{S}^{ISH}$. Comparison between SMI-BLAST and PSI-BLAST in terms of true-positives number at the same EPQ (**B**) on the ISH error subset $\mathbb{S}^{ISH}$. Comparison between SMI-BLAST and PSI-BLAST in terms of ROC1 score, ROC50 score and true-positive number and EPQ on different subsets of SCOPe benchmark dataset (**C**, **D**, **E** and **F**). In boxplots, the first and third quartiles are shown in the box, the lines inside the box are median and the whiskers outside the box are 1.5 times the interquartile. Yellow points show the average score (B). The results of PSI-BLAST are obtained by PSI-BLAST 2.7.1+ (https://blast.ncbi.nlm.nih.gov/Blast.cgi). (Color version of this figure is available at *Bioinformatics* online.)

Table 2. The performance comparison between SMI-BLAST and PSI-BLAST after two iterations

Dataset	Method	ROC1[a]	ROC50[a]	TP number[a]	EPQ[a]
$\mathbb{S}^{ISH}_{MIX}$	PSI-BLAST[b]	0.6432	0.8747	31.9432	0.5304
	SMI-BLAST	0.8487	0.9575	30.9782	0.4614
$\mathbb{S}^{ISH}_{NULL}$	PSI-BLAST[b]	0.4268	0.6723	3.7067	0.7059
	SMI-BLAST	0.8214	0.9292	12.8247	0.5421
$\mathbb{S}^{ISH}_{ALL}$	PSI-BLAST[b]	0.0000	0.2989	6.6667	0.7415
	SMI-BLAST	0.3596	0.7529	19.1667	0.5742
$\mathbb{S}^{T}$	PSI-BLAST[b]	0.9225	0.9636	70.7306	0.3027
	SMI-BLAST	0.9680	0.9905	70.6471	0.3016
$\mathbb{S}^{F}$	PSI-BLAST[b]	0.0047	0.0131	0.0367	0.9951
	SMI-BLAST	0.0032	0.0047	0.0119	0.9989

[a]Performance at homology level that contains close homology and remote homology (belonging to the same SCOP superfamily).

[b]The results of PSI-BLAST are obtained by PSI-BLAST 2.7.1+ (https://blast.ncbi.nlm.nih.gov/Blast.cgi). It is worth noting that PSI-BLAST 2.7.1+ can produce some homology protein sequences on set $\mathbb{S}^{ISH}_{NULL}$ and set $\mathbb{S}^{ISH}_{ALL}$.

TP (Fig. 4C), SMI-BLAST is able to detect more TPs (Fig. 4E). On set $\mathbb{S}^{T}$ and set $\mathbb{S}^{F}$, SMI-BLAST shows stable performance compared with PSI-BLAST, further confirming the better results of SMI-BLAST.

Table 3. Performance comparison between PSI-BLAST and SMI-BLAST at different iterations for protein remote homology detection on the SCOPe 2.06 benchmark dataset

Iteration	PSI-BLAST[a]			SMI-BLAST		
	ROC1[b]	ROC50[b]	Coverage[b]	ROC1[b]	ROC50[b]	Coverage[b]
1	0.8318	0.8896	0.3978	0.9082	0.9269	0.3978
2	0.8432	0.8904	0.4636	0.8983	0.9230	0.4690
5	0.8513	0.8941	0.5134	0.8894	0.9192	0.5200
10	0.8523	0.8945	0.5235	0.8878	0.9190	0.5309

[a]The results of PSI-BLAST are obtained by PSI-BLAST 2.7.1+ (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

[b]Performance at homology level containing close homology and remote homology (belonging to the same SCOP superfamily).

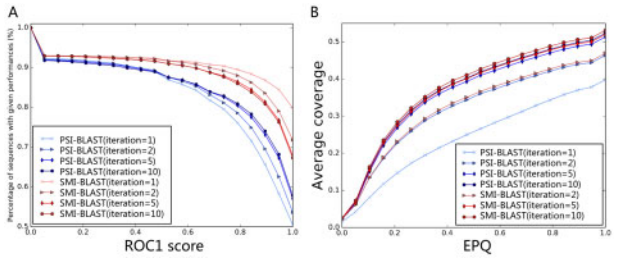

Fig. 5. SMI-BLAST outperforms PSI-BLAST in terms of ROC1 score at any iteration on the SCOPe 2.06 benchmark dataset

3.3 The influence of iteration number on the performance of SMI-BLAST

To investigate the influence of iteration number on the performance of SMI-BLAST, we investigate the ranking quality and detected true-positive number of SMI-BLAST at different iterations.

As can been seen from Table 3 and Figure 5, the performance of SMI-BLAST obviously outperforms PSI-BLAST at any iteration. It should be noted that the ranking quality of SMI-BLAST declines slightly with the growth of iteration number, but it is still obviously higher than that of PSI-BLAST (Table 3 and Fig. 5). The reason is that more low similarity protein sequences are detected at the following iterations, making it difficult to resort them by the final resort step of SMI-BLAST. Therefore, it is necessary for users to make a decision between ranking quality and the number of TPs.

3.4 SMI-BLAST solves the ISH errors of PSSM

For SMI-BLAST, resolving the ISH errors of PSSM mainly contributes to its better performance for protein remote homology detection. When ISH errors are solved by SMI-BLAST, higher ROC score and more TPs can be obtained in the iteration process. To further explore how many ISH errors are solved by SMI-BLAST, two situations are analysed: (i) ISH errors are completely corrected, which means that the PSSMs with ISH errors can be all converted into True-PSSMs; (ii) ISH errors are mitigated, which denotes that PSSMs with ISH-NULL error and ISH-ALL error can be converted into PSSMs with ISH-MIX error. The reason for mitigating ISH errors is that the performance of PSI-BLAST on set $\mathbb{S}^{ISH}_{MIX}$ is obviously higher than that on set $\mathbb{S}^{ISH}_{NULL}$ and set $\mathbb{S}^{ISH}_{ALL}$ (Table 1).

About 90.8% of ISH-MIX errors are corrected into True-PSSM (Supplementary Fig. S2B and Supplementary Table S1), indicating that the incorrect evolutionary information of ISH-MIX can be removed by the first learning to rank model, contributing to lower EPQ scores on set $\mathbb{S}^{ISH}_{MIX}$ (Fig. 4F and Table 2). About 79% ISH-NULL errors and 50% ISH-ALL errors are corrected or mitigated, leading to about 3-fold incensement in terms of the number of TP (Supplementary Fig. S2C, D and Supplementary Table S1). Based on these results, we conclude that: (i) Removing the false positives from

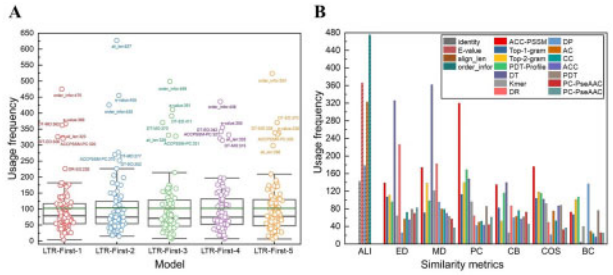

Fig. 6. The usage frequencies of sequence similarity scores in LTR-First models. The labels in Figure 7A represent how to calculate the usage frequencies in the model, such as DT-MD: 362 represents the usage frequencies score is 362 and it is calculated by DT feature and Manhattan distance

Table 4. Performance comparison of SMI-BLAST with different feature combinations for protein remote homology detection on the SCOPe 2.06 benchmark dataset

Methods	ROC1[a]	ROC50[a]	Coverage[a]
PSI-BLAST	0.8432	0.8904	0.4636
SMI-BLAST[b]	0.8957	0.9216	0.4625
SMI-BLAST[c]	0.8977	0.9227	0.4676
SMI-BLAST[d]	0.8980	0.9229	0.4689
SMI-BLAST[e]	0.8983	0.9230	0.4690

[a]Performance at homology level containing close homology and remote homology (belonging to the same SCOP superfamily).

[b]The iteration number of SMI-BLAST is 2, and 7 features with usage frequencies greater than 200 in Supplementary Tables S2 and S3 are used to train the learning-to-rank models of SMI-BLAST.

[c]The iteration number of SMI-BLAST is 2, and 31 features with usage frequencies greater than 100 in Supplementary Tables S2 and S3 are used to train the learning-to-rank models of SMI-BLAST.

[d]The iteration number of SMI-BLAST is 2, and 70 features with usage frequencies greater than 50 in Supplementary Tables S2 and S3 are used to train the learning-to-rank models of SMI-BLAST.

[e]The iteration number of SMI-BLAST is 2, and all the 89 features are used to train the learning-to-rank models of SMI-BLAST.

the selected list of ISH-MIX is essential, leading to more sensitive PSSMs; (ii) The feedback sequences in candidate list can provide useful alignment information, although the alignment similarity between query sequences and feedback sequences in candidate list is relatively low.

3.5 Sequence similarity matrix analysis

The importance of similarity scores in sequence similarity matrix should be explored, because the sequence similarity matrix is the most crucial module for the two learning to rank models. In this section, the FeatureManager tool of Ranklib-2.10 (Burges *et al.*, 2005) is used to generate the usage frequencies of sequence similarity scores of LTR-First model, where higher usage frequencies indicate more important contribution to this model. The distributions of usage frequencies of five LTR-First models trained by 5-fold cross-validation are similar (Fig. 6A), indicating that the contribution of these similarity features for LTR-First model is stable.

To investigate the effect of sequence similarity features on the performance of the SMI-BLAST, different feature combinations are used to train SMI-BLAST according to their usage frequencies obtained by the FeatureManger tool. From Figure 6B, Table 4, Supplementary Tables S2 and S3, we can see the followings: (i) the SMI-BLAST trained with all the 89 features can achieve the best performance; (ii) the SMI-BLAST trained with the top 7 most important features can obviously improve the performance of the PSI-BLAST;

Table 5. Performance of various methods for protein remote homology detection on the SCOPe 2.06 benchmark dataset

Methods	Performance on SCOPe benchmark dataset			
	ROC1[a]	ROC10[a]	ROC20[a]	ROC50[a]
PSI-BLAST	0.8513	0.8885	0.8921	0.8941
SMI-BLAST	0.8894	0.9146	0.9175	0.9192
JackHMMER	0.8919	0.9027	0.9043	0.9059
SMI-HMMER	0.8975	0.9103	0.9123	0.9138
DELTA-BLAST	0.8910	0.9157	0.9199	0.9233
SMI-DELTABLAST[b]	0.9051	0.9299	0.9333	0.9357
PSI-BLASTexB	0.8754	0.9041	0.9081	0.9112
SMI-PSIBLASTexB[b]	0.8929	0.9235	0.9276	0.9306

[a]Performance at homology level that contains close homology and remote homology (belonging to the same SCOP superfamily).

[b]The parameters are given in Supplementary Table S4.

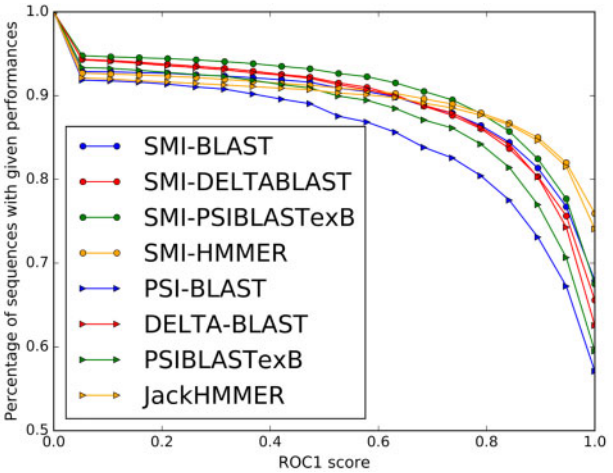

Fig. 7. SMI-based framework improves the performance of PSI-BLAST, DELTA-BLAST, PSI-BLASTexB and JackHMMER on the SCOPe 2.06 benchmark dataset

(iii) when more features are added according to their frequencies, the performance improvement of SMI-BLAST decreases gradually; (iv) the e-value score is one of the most important features, indicating that the search results of PSI-BLAST have an important influence on the performance of SMI-BLAST; (v) the profile-based features are more important than the other sequence features in this framework, which is consistent with previous studies (Liu *et al.*, 2019).

3.6 Application of SMI-based framework to related search methods

To evaluate its generalization, the proposed SMI-based framework is applied to JackHMMER and two PSI-BLAST related methods: DELTA-BLAST (Boratyn *et al.*, 2012) and PSI-BLASTexB (Oda *et al.*, 2017). JackHMMER (Johnson *et al.*, 2010) is a state-of-the-art iterative similarity search method based on HMM profiles with a similar iterative process as that of PSI-BLAST. DELTA-BLAST and PSI-BLASTexB are two improved versions of PSI-BLAST. Unfortunately, they also suffer from the ISH errors on protein domain dataset, because HMM profile and PSSM profile are both constructed based on the multiple sequence alignments.

From Table 5 and Figure 7, we can see: (i) SMI-based framework can not only obviously improve the performance of PSI-BLAST, but also improve the performance of the three related search methods. The reason for the performance improvement is that those methods also suffer from the ISH errors when iteratively searching on protein

domain database; (ii) SMI-based framework and two PSI-BLAST-based search methods (DELTA-BLAST and PSI-BLASTexB) are complementary, because they improve the performance of PSI-BLAST using different theories and techniques; (iii) the SMI-based framework can improve the performance of all the four methods, and SMI-DELTABLAST achieves the best performance. Based on the analysis of the sequence similarity matrix in the previous section, it is reasonable to conclude that the original search method can provide important sequence similarity characteristics; (iv) the improvement of JackHMMER is less than that of PSI-BLAST when applying the SMI-based framework. The most important reason is that most of the JackHMMER search results achieve the best ranking quality, leading to a limited optimization space for SMI-based framework.

4 Conclusion

In this study, we summarize three types of ISH errors of PSSM for protein remote homology detection. To overcome those errors, we propose SMI-BLAST by applying the SMI-based framework to PSI-BLAST. Experimental results show that SMI-BLAST outperforms PSI-BLAST by solving the ISH errors of PSSM on protein domain database. Sequence similarity matrix plays an important role in the proposed framework, whose sequence similarity features are the key to extract correct homology information from the ranking list and improve the ranking quality search results. When applied to DELTA-BLAST, PSI-BLASTexB and JackHMMER, the proposed SMI-based framework can also improve the predictive performance of those methods. The web servers of the proposed methods are constructed (http://bliulab.net/SMI-BLAST/). For more information of the web servers, please refer to the 'Web server and user guide' section in Supplementary Data. Sequence search is an important task in protein sequence analysis, and the proposed SMI-based framework is a general framework, which would be also applied to solve other related problems, such as protein fold recognition, protein–protein interaction prediction, etc.

Acknowledgements

The authors are very much indebted to the three anonymous reviewers, whose constructive comments are very helpful for strengthening the presentation of this article.

Funding

This work was supported by the National Natural Science Foundation of China [61822306, 61672184 and 61702134], the Beijing Natural Science Foundation [JQ19019], National Key R&D Program of China [2018AAA0100100] and Guangdong Special Support Program of Technology Young talents [2016TQ03X618].

Conflict of Interest: none declared.

References

Alam,I. *et al.* (2004) Comparative homology agreement search: an effective combination of homology-search methods. *Proc. Natl. Acad. Sci. USA*, **101**, 13814–13819.

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

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bass,J.I.F. *et al.* (2013) Using networks to measure similarity between genes: association index selection. *Nat. Methods*, **10**, 1169–1176.

Bengio,Y. and Grandvalet,Y. (2004) No unbiased estimator of the variance of K-fold cross-validation. *J. Mach. Learn. Res.*, **5**, 1089–1105.

Biegert,A. and Soding,J. (2009) Sequence context-specific profiles for homology searching. *Proc. Natl. Acad. Sci. USA*, **106**, 3770–3775.

Boratyn,G.M. *et al.* (2012) Domain enhanced lookup time accelerated BLAST. *Biol. Direct*, **7**, 12–12.

Borgefors,G. (1984) Distance transformations in arbitrary dimensions. *Comput. Graph. Image Process.*, **27**, 321–345.

Burges,C.J.C. (2010) From ranknet to lambdarank to lambdamart: an overview. *Learning*, **11**, 81.

Burges,C.J.C. *et al.* (2005) Learning to rank using gradient descent. In: Proceedings of the 22nd International Conference on Machine Learning. ACM. New York, NY, USA. pp. 89–96.

Burges,C. *et al.* (2006) Learning to rank with nonsmooth cost functions. In Schölkopf,B. et al. (eds) Proceedings of the 19th International Conference on Neural Information Processing Systems (NIPS' 06). MIT Press, Cambridge, MA, USA, pp. 193–200.

Carroll,H.D. *et al.* (2015) Improving retrieval efficacy of homology searches using the false discovery rate. *IEEE ACM Trans. Comput. Biol.*, **12**, 531–537.

Chandonia,J.-M. *et al.* (2019) SCOPe: classification of large macromolecular structures in the structural classification of proteins—extended database. *Nucleic Acids Research*,**47**, D475–D481.

Chen,J.J. *et al.* (2017) ProtDec-LTR2.0: an improved method for protein remote homology detection by combining pseudo protein and supervised learning to rank. *Bioinformatics*, **33**, 3473–3476.

Chen,J. *et al.* (2018) A comprehensive review and comparison of different computational methods for protein remote homology detection. *Brief. Bioinf.*, **19**, 231–244.

Chou,K.C. (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins Struct. Funct. Genet.*, **43**, 246–255.

Chou,K.C. (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, **21**, 10–19.

Danielsson,P.-E. (1980) Euclidean distance mapping. *Comput. Graph. Image Process.*, **14**, 227–248.

Dong,Q.W. *et al.* (2009) A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. *Bioinformatics*, **25**, 2655–2662.

Donmez,P. *et al.* (2009) On the local optimality of LambdaRank. *In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, *New York*, NY, USA, pp. 460–467.

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

Gonzalez,M.W. and Pearson,W.R. (2010) Homologous over-extension: a challenge for iterative similarity searches. *Nucleic Acids Res.*, **38**, 2177–2189.

Guo,Y.Z. *et al.* (2008) Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic Acids Res.*, **36**, 3025–3030.

Hou,J. *et al.* (2018) DeepSF: deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics*, **34**, 1295–1303.

Johnson,L.S. *et al.* (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, **11**, 431.

Klove,T. *et al.* (2010) Permutation Arrays Under the Chebyshev Distance. *IEEE Transactions on Information Theory*, **56**, 2611–2617.

Lee,M.M. *et al.* (2008) Simple is beautiful: a straightforward approach to improve the delineation of true and false positives in PSI-BLAST searches. *Bioinformatics*, **24**, 1339–1343.

Lee Rodgers,J. and Nicewander,W.A. (1988) Thirteen ways to look at the correlation coefficient. *Am. Stat.*, **42**, 59–66.

Li,H. (2011) A short introduction to learning to rank. *IEICE Trans. Inf. Syst.*, **E94-D**, 1854–1862.

Li,W.Z. *et al.* (2012) PSI-Search: iterative HOE-reduced profile SSEARCH searching. *Bioinformatics*, **28**, 1650–1651.

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**, 510–510.

Liu,B. *et al.* (2012) Using amino acid physicochemical distance transformation for fast protein remote homology detection. *PLoS One*, **7**, e46633.

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) iDNA-Prot|dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition. *PLoS One*, **9**, e106691.

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

Liu,B. *et al.* (2015b) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.*, **43**, W65–W71.

Liu,B. *et al.* (2018b) iPromoter-2L: a two-layer predictor for identifying promoters and their types by multi-window-based PseKNC. *Bioinformatics*, **34**, 33–40.

Liu,B. *et al.* (2019) BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.*, **47**, e127–e127.

Marchler-Bauer,A. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.*, **30**, 281–283.

Marchler-Bauer,A. *et al.* (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.*, **39**, D225–D229.

Oda,T. *et al.* (2017) Simple adjustment of the sequence weight algorithm remarkably enhances PSI-BLAST performance. *BMC Bioinformatics*, **18**, 288.

Pearson,W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.*, **183**, 63–98.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.

Pearson,W.R. *et al.* (2017) Query-seeded iterative sequence similarity searching improves selectivity 5-20-fold. *Nucleic Acids Res.*, **45**, e46–e46.

Reid,A.J. *et al.* (2007) Methods of remote homology detection can be combined to increase coverage by 10% in the midnight zone. *Bioinformatics*, **23**, 2353–2360.

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

Senior,A.W. *et al.* (2020) Improved protein structure prediction using potentials from deep learning. *Nature*, **577**, 706–710.

Singhal,A. (2001) Modern information retrieval: a brief overview. *IEEE Data Eng. Bull.*, **24**, 35–43.

Somerfield,P.J. (2008) Identification of the Bray-Curtis similarity index: Comment on Yoshioka (2008) *Marine Ecology Progress Series*, **372**, 303–306.

The UniProt,C. (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res.*, **45**, D158–D169.

Wistrand,M. and Sonnhammer,E.L. (2005) Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. *BMC Bioinformatics*, **6**, 99.

Yan,K. *et al.* (2019) Protein fold recognition based on multi-view modeling. *Bioinformatics*, **35**, 2982–2990.

Yang,M. *et al.* (2019) Combined alignments of sequences and domains characterize unknown proteins with remotely related protein search PSISearch2D. *Database (Oxford)*, **2019**, baz092.

Zou,Q. *et al.* (2016) Pretata: predicting TATA binding proteins with novel features and dimensionality reduction strategy. *BMC Syst. Biol.*, **10**, 401–412.