Briefings in Bioinformatics Advance Access published November 22, 2016

Briefings in Bioinformatics, 2016, 1–14

doi: 10.1093/bib/bbw108
Paper

A comprehensive review and comparison of different computational methods for protein remote homology detection

Junjie Chen, Mingyue Guo, Xiaolong Wang and Bin Liu

Corresponding author: Bin Liu, Harbin Institute of Technology Shenzhen Graduate School, HIT Campus Shenzhen University Town, Xili, Shenzhen 518055, China. Tel.: (+86) 0755-86011630; E-mail: binliu@hitsz.edu.cn

Abstract

Protein remote homology detection is one of the most fundamental and central problems for the studies of protein structures and functions, aiming to detect the distantly evolutionary relationships among proteins via computational methods. During the past decades, many computational approaches have been proposed to solve this important task. These methods have made a substantial contribution to protein remote homology detection. Therefore, it is necessary to give a comprehensive review and comparison on these computational methods. In this article, we divide these computational approaches into three categories, including alignment methods, discriminative methods and ranking methods. Their advantages and disadvantages are discussed in a comprehensive perspective, and their performance is compared on widely used benchmark data sets. Finally, some open questions in this field are further explored and discussed.

Key words: protein remote homology detection; protein structure and function; alignment methods; discriminative methods; ranking methods

Introduction

Protein remote homology detection refers to the identification of the homologous proteins, which are similar in structure and function but sharing low sequence identity. In the long-term natural evolutionary process, because protein structures and functions are more conserved than their sequences [1], proteins sharing similar structures and functions may have low sequence identities [2]. For protein homology detection, detecting the homologs with high sequence identity is much easier than detecting those with low sequence identity.

Sequence alignment methods [3, 4] can unambiguously distinguish between protein pairs of similar and non-similar structure when the pairwise sequence identity is high (>40% for long alignments). However, when the sequence identity is within the 'twilight zone' of 20–35%, so-called remote homology detection, homology detection becomes a difficult task [5, 6].

Detecting remote homolog proteins has an important impact on the proteomics [7], biomedical sciences [8], etc., and it is one of the fundamental techniques for protein structure and function prediction. Christian B. Anfinsen, the 1972 Noble Prize winner in Chemistry, confirms the connection between the amino acid sequence and the biologically active conformation [9], which provides reliable evidence that it is possible to predict the protein structures based only on their amino acid sequences. However, this problem is far from being solved.

As the development of sequencing techniques, the number of protein sequences is growing rapidly. Up to June 2016, there

Junjie Chen is a PhD candidate at the School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, China. His research areas include bioinformatics, nature language processing and machine learning.
Mingyue Guo is a master student at the School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, China. Her research areas include bioinformatics and machine learning.
Xiaolong Wang, PhD, is a professor at the School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, China. His research areas include nature language processing, bioinformatics and artificial intelligent.
Bin Liu, PhD, is a professor at the School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, China. His expertise is in bioinformatics, nature language processing and machine learning.

Submitted: 3 August 2016; **Received (in revised form):** 14 September 2016

are >64 million protein sequences in UniProtKB/TrEMBL database [10] (http://www.uniprot.org/), and millions of sequences are added into this database per month. In contrast, the number of proteins with known structures grows much slower. Up to June 2016, there are only about 119 000 structures deposited in protein data bank (PDB) [11] (http://www.rcsb.org/pdb/home/home.do), and only thousands of structures are added per year. Therefore, the huge gap between protein sequences and structures is obvious and quickly increasing. It is an emergency task to explore effective and low-cost approaches to reduce this gap. Because the traditional biological techniques for protein remote homology detection are expensive and ineffective, computational approach is an alternative scheme with low cost.

During the past decades, many new computational approaches have been proposed, which significantly promote the development of protein remote homology detection. However, the latest review papers [12, 13] in this field were published a decade ago. To give the researchers a catching-up view on the recent development in this area, an updated review is highly needed. In this article, we will give a comprehensive review and comparison on the computational approaches in the field of protein remote homology detection, especially emphasizing on the recently proposed methods. First, the golden standard database of structural classification of proteins (SCOPs) is introduced; second, computational methods are reviewed and divided into three categories based on the used machine learning techniques; and finally, the performance of various methods are compared and discussed.

Structural classification of proteins database

Some databases classify proteins into different groups according to their structures and evolutionary relationships, such as SCOP [14, 15], SCOP extended (SCOPe) [16], etc. Once a novel protein is classified into a known group, its structural and functional properties can be inferred according to the homologous proteins in this group.

The SCOP [14, 15] is one of the commonly used databases for protein remote homology detection, which is constructed manually by visual inspection and comparison of structures. As reported in [17], 571 articles (published during 2012–13) cited SCOP data sets. It has become the gold standard database for evolutionary classification. Proteins in SCOP are classified in a hierarchy way to reflect their structures and evolutionary relationships, as shown in Figure 1.

By 2009, about 38 000 PDB entries are manually classified into a strictly hierarchical structure in SCOP database. In general, the proteins in the same superfamily are homologous, and

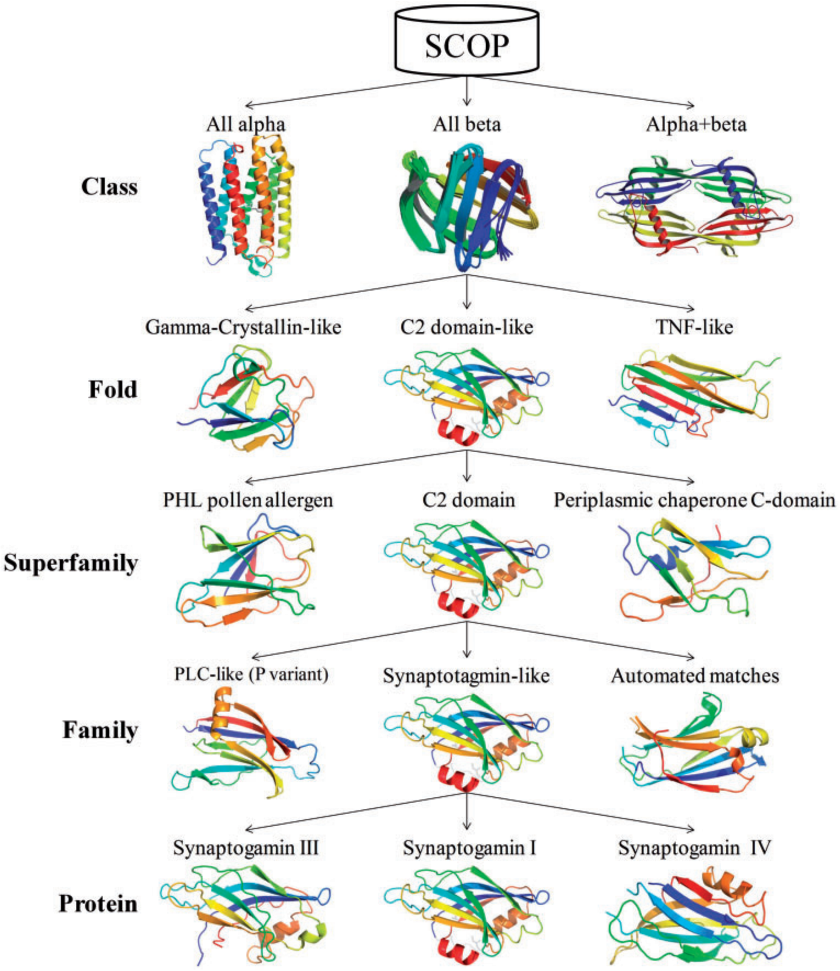

Figure 1. The tree hierarchy of SCOP database. All proteins in SCOP are organized in four levels: class, fold, superfamily and family. In general, proteins in the same family have clearly evolutionary relationship. Proteins in the same superfamily but not in the same family are remote homologous proteins. Proteins in the same fold but not in the same superfamily have unclearly evolutionary relationship. The proteins in different classes have no evolutionary relationship.

those in different folds are not homologous. Proteins in the same superfamily but not in the same family are remote homologous proteins. Most of the computational methods for remote homology detection methods are trained and evaluated based on the SCOP database [18].

SCOPe [16] is an extension of SCOP database by using the automatic annotation techniques, which uses the same hierarchical system as SCOP database, and is fully compatible with SCOP. Besides, there are some other databases that can be used to build predictive models for protein remote homology detection, such as CATH [19–21] and Pfam [22, 23]. The CATH database [19–21] is a hierarchical domain classification of proteins according to their structures in the PDB. These protein structures are classified by using a combination of both automated and manual procedures. There are four major levels in CATH: class, architecture, topology and homology. Pfam database [22, 23] is a large collection of protein families and domains, and each protein is represented by multiple sequence alignment and hidden Markov model (HMM). A summary of the widely used protein classification databases is shown in Table 1.

Computational methods

Computational methods for protein remote homology detection have been studies for decades, and many powerful approaches have been proposed. To help the readers to understand their development, we roughly group these computational methods into three categories according to their methodologies and machine learning techniques, including alignment methods, discriminative methods and ranking methods. Please note that it does not mean that the methods in one category are completely different from the others. In fact, methods in different categories also share some similar techniques, for example the alignment methods are one of the most important approaches for constructing the feature vectors for both the discriminative and ranking methods [2, 24]. However, the detection strategies of the methods from three categories are different, for example the alignment methods search the best local alignments (LAs) by using dynamic programming, discriminative methods construct classification models by using machine learning classifiers and ranking methods treat the remote homology detection problem as a ranking task by searching the target proteins against protein databases.

Alignment methods

Alignment methods are one of the earliest and most widely used techniques for protein remote homology detection, which find the best-matching local or global alignments of two proteins with gap penalties. These alignment methods can be further categorized into three groups based on the different alignment strategies, including sequence alignment, profile alignment and HMM alignment.

(i) Sequence alignment methods

Sequence alignment methods are the foundational techniques for inferring the homology between a pair of proteins. In these methods, the sequence alignments between two sequences are calculated by using dynamic programming algorithms, such as global alignment (Needleman–Wunsch) [25] and LA (Smith–Waterman) [26].

Global alignments perform well when the lengths of sequences in the data set are roughly the same because they attempt to align every residue in each sequence. Local alignments are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context. To improve the computational efficiency, some heuristic algorithms have been developed to find a near optimal alignment in a more efficient fashion, such as BLAST [27] and FASTA [28]. These approaches reduce accuracy for improving efficiency. For more information of these methods, please refer to [29].

Analysis of overlap between structurally alignments shows that sequence identity and sequence similarity measures are poor indicators of structural relatedness in the 'twilight zone' [5, 6], while the alignment score allows much better discrimination between alignments of structurally related and unrelated sequences for a wide variety of alignment settings, as reported in [30]. The sequence alignment methods make great contributions to infer the homology of protein pairs. Later methods such as profile alignment and HMM alignment methods are all based on the framework and idea of sequence alignment methods.

(ii) Profile alignment methods

To further improve the sensitivity of the aforementioned sequence alignment methods, some profile alignment methods have been proposed. A profile is calculated based on the Multiple Sequence Alignments (MSAs) generated by searching against a nonredundant database [31] in an unsupervised manner. Each protein sequence in an MSA has statistically

Table 1. The summary of protein classification databases

Type	Latest version	Description	Websites
SCOP	v1.75 Feb 23, 2009	7 classes 1195 folds 1962 superfamilies 3902 families	http://scop.mrc-lmb.cam.ac.uk/scop/
SCOPe	v2.06 Jan 13, 2016	7 classes 1221 folds 2008 superfamilies 4851 families	http://scop.berkeley.edu/
CATH	v4.0 Mar 26, 2013	2 35 858 CATH domains 2738 CATH superfamilies 69 058 annotated PDBs	http://www.cathdb.info/
Pfam	V30.0 Jul 1, 2016	16 306 entries 559 clans	http://pfam.xfam.org/

significant sequence identity with the query protein. A profile can be represented as a Position-Specific Weight Matrix or Position-Specific Scoring Matrix (PSSM) [4]. The profile incorporates the evolutionary information extracted from MSAs [2, 18, 32], and therefore it is a more powerful representation than the amino acid sequence. The profile alignment methods can be split into three categories: (i) searching a profile query against a database of sequences; (ii) searching a sequence query against a database of profiles; and (iii) searching a profile query against a database of profiles. The flowchart of profile alignment methods is shown in Figure 2.

The approaches in the first category search a profile query against a database of sequences. One of the most widely used approaches is Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) [3, 4]. If an input query is a single protein sequence, PSI-BLAST starts to run a sequence alignment program, such as BLASTp [3], to generate an MSA. The query's profile is calculated from the MSA. In the second round, some newly detected sequences are added into the MSA generated in the first round. The profile then is refined for next iteration. This process is iteratively searched until no new sequence is detected.

Different from those in the first category, some methods in the second category search a query sequence against a database of profiles, such as IMPALA [33]. Methods in this group first build a profile database, in which each protein sequence is represented by a profile. One of the key searching techniques is sequence–profile alignment. The performance of IMPALA is comparable with PSI-BLAST for protein remote homology detection [33].

Methods in the third category are based on profile–profile alignment, which search a profile query against a database of profiles, such as COMPASS [34, 35], FFAS [36] and COMA [37]. These methods construct profiles for the query and each protein in the database, and then detect the homology relationships among proteins via local profile–profile alignment. The profile–profile alignment is more sensitive than the sequence–profile alignment, and thus the methods in this group outperform those in the first and second groups.

Some variants of the profile alignment methods are proposed so as to further improve the detection performance. They attempt to incorporate the secondary structure information into the framework of these methods, such as Phyre [38], FORTE [39], PFRES [40], SPARK-X [41], BioShell [42], etc. This is because protein structures are more conserved than their sequences during the evolutionary process, and therefore the structure information is a more accurate indicator for the remote homology relationships.

The profile alignment methods achieve promising performance, and do play an important role in stimulating the development of this important area. Meanwhile, they also have some disadvantages, such as low quality of MSA prevents the performance of these methods. Recent studies [43] show that the profile–profile scoring functions are critical for MSA. Therefore, designing more accurate profile–profile scoring functions would be a key to improve these methods. Furthermore, a profile only accounts for single position-specific information. However, adjacent position-specific information is also useful for extracting richer evolutionary information [44, 45].

(iii) MM alignment methods

HMMs [46] are applied to protein remote homology detection, which provide a probabilistic measurement of remote homologous sequences based on the pairwise comparison of HMMs [46]. HMM transforms a multiple sequence alignment into a position-specific scoring system [47], which cannot only

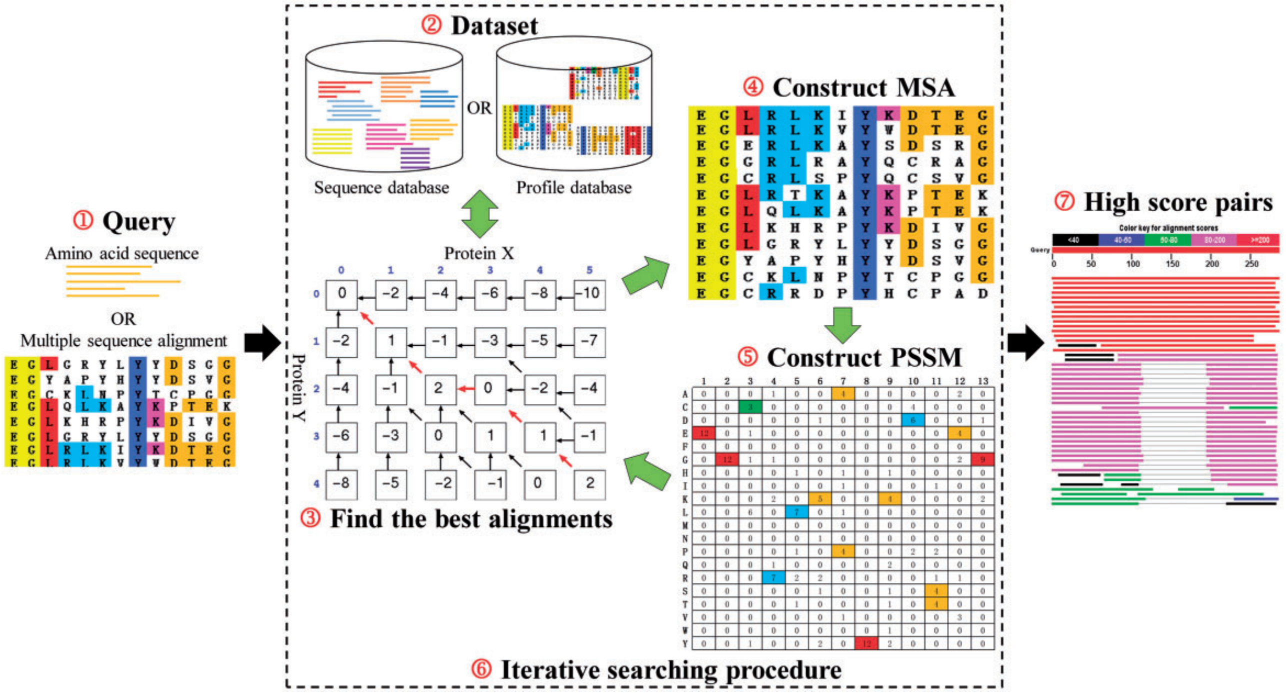

Figure 2. The flowchart of profile alignment methods. First, a query is aligned with all sequences in the database to find out some statistically significantly similar sequences with the query. Second, a profile is calculated based on the generated multiple sequence alignment. The next iteration takes the new profile as a query to search the whole database. The profile alignment methods can be split into three categories: (i) searching a profile query against a database of sequences, such as PSI-BLAST; (ii) searching a sequence query against a database of profiles, such as IMPALA; and (iii) searching a profile query against a database of profiles, such as COMA.

produce a single highest-scoring sequence, but also output a family of possible alignments. Thus, the HMM alignment models are more sensitive than profile alignment methods, and can be used for evaluating the biological significance [48].

The procedure of constructing a profile HMM based on an MSA is shown in Figure 3A, from which we can see that the frequencies in each column of an MSA are counted, and each of them are modeled by a match state M_k. An ungapped HMM mode is generated by connecting all match states. However, this ungapped model only can represent protein sequences matching the consensus sequences of the MSA without any gap. The insert states I_k and delete states D_k are added into ungapped model so that it can account for the insertions and deletions in new observation sequence.

As shown in Figure 3B, each path in an HMM model corresponds to an emitted sequence. The available transitions between alignment pairs include MM, MI, IM, DG and GD. With dynamical programming, the best alignments of two HMMs can be calculated by maximizing the log-sum-of-odds score.

Several variants of HMM-based methods are also proposed, such as HMMER [49], SAM [50], HHsearch/HHblits [51], etc. These methods significantly reduce the computational cost, especially for the sequences with overlapping regions [52, 53]. A comparison between SAM and HMMER has been conducted in [54], and the results show that SAM model estimation is more sensitive, while HMMER model scoring is more accurate. Besides, several profile HMM resources have been built for HMM alignment, such as the PROSITE database [55, 56] and the Pfam database [22]. They are collections of protein families, which are represented by multiple sequence alignments and HMMs.

HMM alignment methods are useful tools for remote homology detection. They have more sensitive performance than both sequence alignment methods and profile alignment methods. However, the alignment methods use only positive samples (sequences in the same superfamily) to train the model, and therefore the trained model tends to predict more false positives.

Alignment methods are the milestones for protein remote homology detection, which play a key role in promoting the development of the computational approaches in this field. Because of their importance, some discriminative and ranking methods are constructed based on these alignment methods, which will be introduced in the following parts.

Discriminative methods

Different from the alignment methods, discriminative methods treat the protein remote homology detection as a superfamily-level classification task. These methods train classification models in a supervised manner using both the positive and negative samples, which are then used to predict the unseen samples. Therefore, the number of false-positive samples can be efficiently reduced compared with alignment methods. To share the advantages of the alignment methods, the feature vectors of some discriminative methods are constructed based on the alignment methods, such as SVM-Pairwise [58], SVM-LA [59], etc.

The remote homology detection is a multiclass classification problem. Its aim is to detect the superfamily of a query protein. Most of these discriminative methods convert the multiclass classification problem into a series of binary classification tasks [58]. In the training process, for each family, a classification model is constructed. The positive test samples are proteins within a target family, and the positive training samples come from the proteins outside this family but within the same superfamily. Negative samples are selected from outside of the same fold [18, 58]. The flowchart of discriminative methods is shown in Figure 4. Some widely used machine learning classifiers are used to construct the discriminative methods, such as support vector machine (SVM) [2], Artificial Neural Network [60, 61], random forest [62, 63], etc. Among these approaches, the SVM-based methods [64] are among the top-performing methods because of the advantage of the kernel tricks.

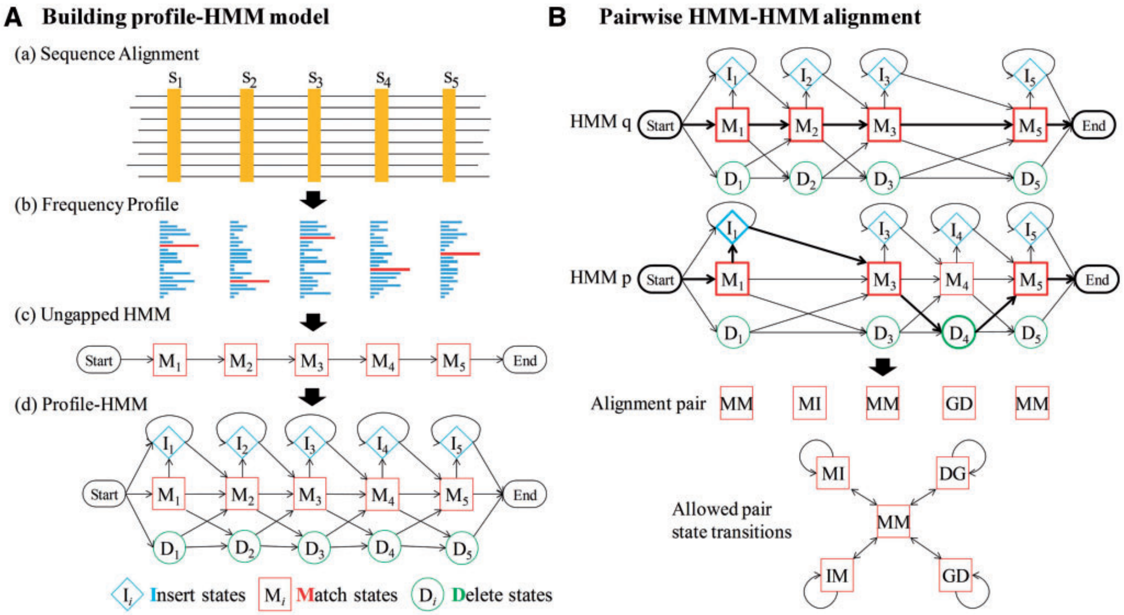

Figure 3. (A) Building a profile HMM model from a multiple sequence alignment. The MSA is transformed into a ungapped HMM model, and then the inert states and delete states are added into the ungapped HMM model. (B) Pairwise alignment of two HMMs. The aim of the alignment is to find the maximization of the logsum-of-odds score by using dynamical programming algorithm.

Nearly, all the discriminative methods require fixed-length feature vectors as inputs. Their performance mainly depends on how to accurately represent protein sequences as feature vectors. However, it is a difficult task because the proteins have different length [65–69]. In this regard, many powerful protein representations have been proposed to capture the characteristics of proteins in different aspects, especially recently a web server called Pse-in-One is established to generate various features of protein, RNA and DNA sequences [65]. The discriminative methods are further divided into four categories according to their different vectorization strategies:

(i) Direct kernel-based methods

Direct kernel functions calculate an explicit similarity of pairwise proteins. There are two classes of direct kernel functions: string kernel and profile kernel [18].

String kernels compute the similarity between a pair of protein sequences. Spectrum kernel [70] is a string kernel, whose features are the set of all possible subsequences of amino acids with fixed length *k*. However, when *k* is large, the feature set grows exponentially. To overcome this problem, mismatch kernel [71] is proposed, which measures the sequence similarity based on the occurrences of fixed-length subsequence patterns in proteins, allowing mutations in these patterns. Later, the reduced amino acid alphabets [72, 73] are used so as to further reduce the dimension of the feature vectors. To incorporate the structure and function information of proteins into the predictors, the motif kernels [74] based on the occurrences of discrete sequence motifs are constructed. The aforementioned string kernels only consider the local composition of proteins, ignoring the sequence-order information and biological evolutionary event information. To overcome these shortcomings, the LA kernel [59] is proposed, which measures the similarity between two sequences by summing up scores obtained from LAs with gaps penalty. This approach is able to detect the subtle sequence similarity.

Profile kernels compute the similarity between a pair of profiles. Window-based kernels are a class of profile kernels, which determine the similarity between a pair of profiles by combining the ungapped profile alignment scores of the subsequences. The window of alignments can be all possible fixed-length sub-profile (AF-PSSM kernel) [75], best fixed-width sub-profile (BF-PSSM kernel) [75] or best variable-width sub-profile (BV-PSSM kernel) [75]. Local alignment-based kernel is another profile kernel, which computes the sequence similarity by finding an optimal LA between two profiles. SW-PSSM [75] finds the optimal LA by using Smith–Waterman alignments with gap penalties. Because the profile kernels consider the evolutionary information extracted from MSA, these approaches outperform the string kernels.

(ii) Indirect kernel-based methods

Indirect kernel-based methods construct the kernel space according to the predictive scores calculated by alignment

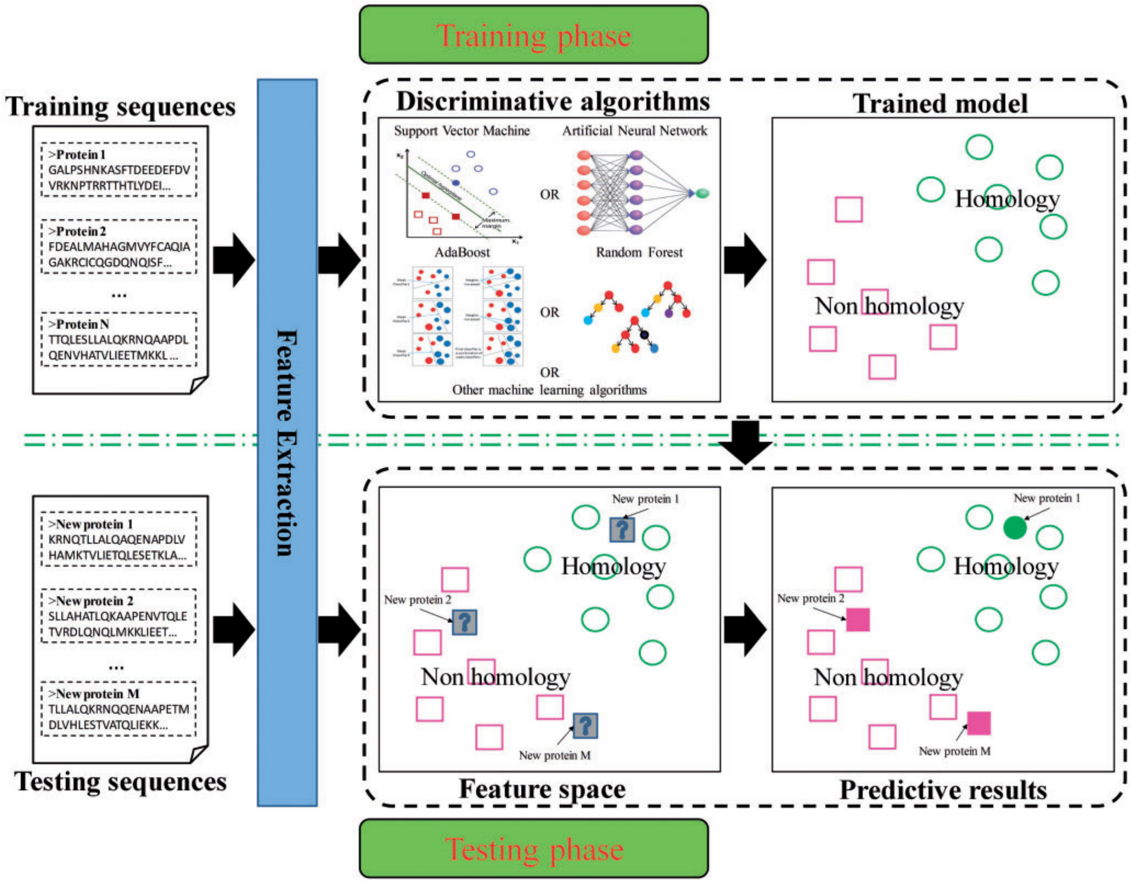

Figure 4. The general framework of discriminative methods. There are two phases in the framework: training phase and testing phase. In training phase, the training sequences are first mapped into a feature space by using feature extraction techniques, such as Top-n-grams [18, 44], PDT [45], etc., and fed into discriminative algorithms to train the model, which can be used to identify the homology proteins.

methods. SVM-Fisher [76] is the first indirect kernel-based method for detecting remote protein homology, which maps all protein sequences into points in a Euclidean feature space based on the sequence similarity scores output by HMM alignment. The feature vectors of SVM-pairwise [58] are constructed based on the pairwise similarity scores calculated by Smith–Waterman algorithm. SVM-BALSA [77] takes into account the Bayesian scores, reflecting all possible alignments to represent the protein sequences. SVM-HUSTLE [78] builds an iterative semi-supervised discriminative model, which classifies a query sequence by training on a collection of representative high-confidence training sets, recruits additional sequences and assigns a statistical measure of homology between a pair of sequences.

(iii) Physicochemical property-based methods

Amino acid physicochemical properties are highly related with protein structure and function [45, 79]. Thus, several methods are developed based on these properties, for examples SVM-RQA [80] proposes a scheme for remote homology detection by using both the amino acid properties and recurrence quantification analysis (RQA). SVM-PCD [81] uses the normalized physicochemical distributions of the 4-mers in protein sequences [82]. This method is further improved by incorporating the sequence-order information, and a predictor based on the physicochemical distance transformation is proposed, called SVM-PDT [45]. To capture more sequence-order information, the disPseAAC [83] and PseAACIndex [84] are proposed, which combine the physicochemical properties and pseudo amino acid composition [85]. The performance of these methods is highly comparable with other discriminative methods, but their computational cost is much lower.

(iv) Methods based on multiple feature combination

As introduced above, various features and predictors have been proposed for protein remote homology detection, which are based on different theories and techniques, and therefore their predictive results are complementary. Recently, several methods combine different features or predictors to further improve the predictive performance, for examples a hybrid machine learning approach [86] is developed, which combines the nearest neighbor methods and multiclass SVMs. Damoulas and Girolami [87] merge multiple feature groups (physicochemical properties, Smith–Waterman scores) together and combine these available feature groups by using a multiclass kernel function. PFP-Pred [88] is formed by a set of basic classifiers constructed based on Optimized Evidence-Theoretic K-Nearest Neighbors rule. Liu *et al.* [2] propose an approach that combines three top-performing string kernels (SVM-Ngram, SVM-pairwise and SVM-LA) by using multiple kernel learning. SVM-Ensemble [89] is an ensemble classifier with a weighted voting strategy, in which it combines three basic classifiers based on different feature spaces, including K-mer, auto-cross covariance (ACC) and SC-PseAAC. PDC-Ensemble [90] is constructed by combining PDC via an ensemble learning approach, in which a new feature of proteins called Pseudo Dimer Composition is presented.

Some efficient feature extraction techniques derived from other fields are applied to protein remote homology detection, such as natural language processing (NLP), speech signal processing and image processing. Some researchers treat the building blocks of protein sequences (N-gram, mismatch, *k*-mer, pattern, etc.) as the words in the natural languages, and treat the protein sequences as the documents. Based on these similarities, some techniques from NLP are applied to protein remote homology detection, such as latent semantic analysis [91, 92] and word correlation matrices [93]. ACC [94] from speech signal processing is applied to transform a PSSM into a fixed-length vector. Top-n-gram [18, 44] is a novel profile-based building blocks of proteins, which contains the evolutionary information extracted from MSAs.

Although discriminative methods perform well for protein remote homology detection, they cannot be widely used in practical applications. There are three main reasons: (1) Because discriminative methods are based on supervised manners, labeled training samples are required. However, the number of proteins with known evolutionary relationship is limited. Therefore, it is difficult to train a useful predictor for the superfamiles with only a few labeled samples. (2) Most of the discriminative methods treat the protein remote homology detection as a series of binary classification tasks. A predictor is constructed for one specific superfamily. There are about 2000 superfamilies in SCOP database, and about 2000 individual predictors should be constructed, which is computational inefficient. (3) If a test protein belongs to an unknown superfamily, the discriminative methods cannot make correct prediction without the labeled training samples of this superfamily.

Ranking methods

Recently, ranking methods have attracted more and more attentions, which treat protein remote homology detection as a ranking task or database searching problem. Similar as the alignment methods, for ranking methods, the query is searched against a protein database that consists of proteins with known structures and functions, and the proteins in the database are sorted according to their evolutionary similarities to the query. Furthermore, ranking methods are also able to incorporate other important features into the feature space, such as physicochemical properties, and sequence features used in the discriminative methods. Therefore, the ranking methods take the advantages of both alignment methods and discriminative methods, and achieve better predictive performance. The flowchart of a typical ranking method is shown in Figure 5.

The performance of these ranking methods depends on how to accurately measure the similarity between two proteins. Several methods have been proposed, for example RankProp [95] is one of the most widely used ranking methods, which uses an unsupervised ranking algorithm that exploits the entire network of similarity relationships among proteins in a database by performing a diffusion operation on a weighted network. The weights on the network are the similarities output by PSI-BLAST or HHbilts. The query is added into the protein similarity network, and then remote homologous proteins can be detected based on the propagation of network. Later, a semi-supervised version of RankProp is developed by using labeled samples in SCOP [96]. A large-scale implementation of RankProp [97] web server is publicly accessible, where 1.1 million proteins are added. Motived by the language models from NLP, ProtEmbed [98] converts a large-scale embedding of protein feature vectors into a low-dimensional ‘semantic space’. Its main idea is to learn a mapping of protein sequences into a feature space that captures their ‘semantic similarity’. Taking the advantage of protein embedding vectors, ProtEmbed outperforms RankProp.

Some ranking methods are developed and achieve the state-of-the-art performance, which raises the possibility to combine these methods to improve the predictive performance. In this regard, ProDec-LTR [83] adopts the learning to rank (LTR)

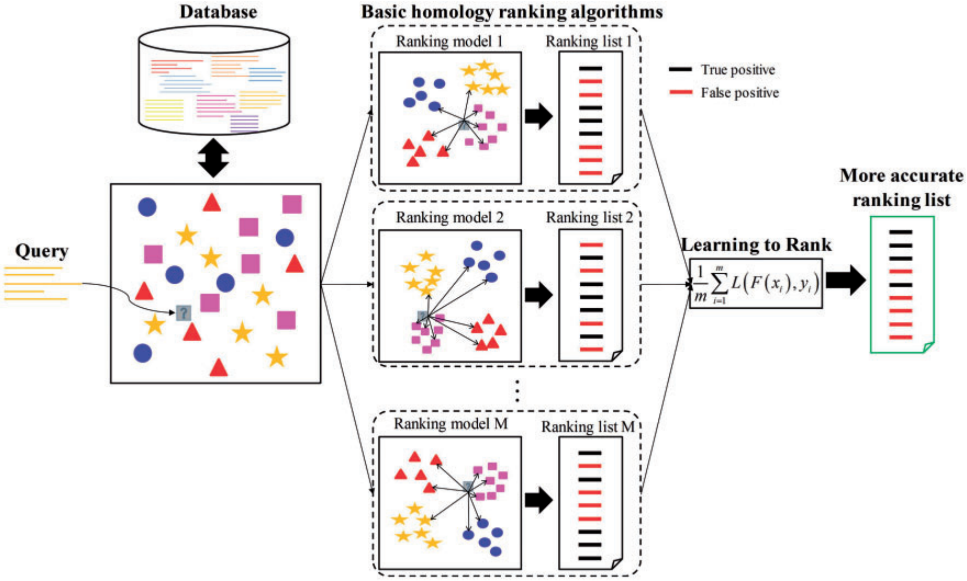

Figure 5. The general framework of ranking methods. All proteins in the database are mapped into a feature space, where the distribution of proteins is scattered arrangement. Then the feature space is transformed by ranking algorithms to cluster the homologous proteins. The distance in the transformed feature space between the protein pairs represents their evolutionary relationship. For a query, it is first mapped the transformed feature space, and then the distance between the query and all the other proteins is measured. The ranking lists are sorted in descending order. In this example, the Learning to Ranking (LTR) algorithm is used to re-rank these ranking lists in a supervised manner in order to improve the predictive performance [24].

technique to combine three different models, including PSI-BLAST, HHblits and ProtEmbed. LTR is a supervised ranking algorithm from NLP. For each query sequence, these three basic predictors are performed on it to generate a candidate list, and then this list is re-ranked by using the LTR model so as to reduce the number of false positives. The dRHP-PseRA [89] combines the profile-based protein representation and four state-of-the-art predictors (PSI-BLAST, HHblits, Hmmer and Coma) via the rank aggregation approach.

Discussion

As introduced and discussed above, because of its importance, many computational predictors for protein remote homology detection have been proposed. In this section, we will make a comprehensive comparison among these methods on two widely used benchmark data sets.

Receiver operating characteristic (ROC), ROC1 and ROC50 are three widely used performance measures for evaluating the performance of a computational predictor for protein remote homology detection. ROC represents the area under the ROC curve [99]. ROC1 and ROC50 scores are the areas when the first false positive and 5th false positives appear, respectively. The discriminative methods treat protein remote homology detection as a series of binary classification tasks, the trained model assigns a probability value for each test sample and their performance is calculated based on all the samples in the test set. In contrast, for the alignment and ranking methods, the samples in the whole data set are sorted according to their similarities with the query, and their performance for each query is calculated based on the whole data set. Therefore, the performance of the discriminative methods and the methods in the other two groups cannot be compared directly. Therefore, they are evaluated on two widely used benchmark data sets, SCOP-D [2, 18] for discriminative methods (Table 2), and SCOP-O [24, 98] for the other computational methods (Table 3).

From Tables 2 and 3, we can see that the protein representation and detection strategy are two important factors for a computational predictor. We will discuss these two points in more details.

The performance comparison from protein representation perspective

Efficient and accurate protein representation is one of the keys for the computational predictors. Sequence-based representations are only based on the amino acid composition of protein sequences, while profile-based and HMM-based representations are constructed from the MSAs. Because MSAs consider the evolutionary events, such as mutations, deletions and insertions of residues in the protein sequences, the profile-based methods and HMM-based methods are more sensitive than the sequence-based methods, which are fully consistent with the results listed in Table 2. As also shown in this table, SW-PSSM achieves an ROC score of 0.982, outperforming other discriminative method. The reason is that both these two methods use the evolutionary information extracted from multiple sequence alignments. As can be seen from Table 3, ProtDec-LTR is the best-ranking method, and the HMM-based methods show better results than the profile-based methods. This is because HMM representation cannot only produce a highest scoring consensus sequence, but also can output a family of all possible homologous proteins.

The performance comparison from detection strategy perspective

The existing detection strategies can be divided into three categories, including alignment strategy, discriminative strategy and ranking strategy.

The alignment strategy is one of the most widely used techniques, such as PSI-BLAST and HHblits. Alignment strategy sums the scores of each position-specific alignment to indicate

Table 2. Performance comparison among different discriminative methods on SCOP-D [2, 18]

Methods	Protein representation[a]	Detection strategy[b]	ROC	ROC50[c]	Source
Mismatch	S	D	0.872	0.400	[94]
GPkernel	S	D	0.899	NA	[74]
SVM-Ngram	S	D	0.791	NA	[91]
SVM-Pattern	S	D	0.835	NA	[91]
SVM-Motif	S	D	0.814	0.616	[91]
SVM-Ngram-LSA	S	D	0.859	NA	[91]
SVM-Pattern-LSA	S	D	0.879	NA	[91]
SVM-Motif-LSA	S	D	0.859	NA	[91]
SVM-LA (β=0.5)	S	D	0.925	0.649	[59]
AF-PSSM (w=3)	P	D	0.976	0.833	[75]
BF-PSSM (w=2)	P	D	0.980	0.854	[75]
BV-PSSM (w=2)	P	D	0.973	0.855	[75]
SW-PSSM (3.0, 0.75 1.5)	P	D	0.982	0.904	[75]
LSTM	P	D	0.932	0.652	[100]
ACCRe_ACC	P	D	0.954	0.894	[94]
SVM-Fisher	H	D	0.773	0.250	[59]
SVM-Pairwise	P	D	0.896	0.464	[59]
SVM-BALSA	S	D	0.935	NA	[77]
SVM-HUSTLE	S	D	0.812	NA	[78]
Profile (5 7.5)	P	D	0.980	0.794	[75]
SVM-RQA	S	D	0.912	0.441	[80]
SVM-DR	S	D	0.919	0.715	[44]
SVM-DT	S	D	0.948	0.800	[44]
SVM-PCD	S	D	0.906	NA	[81]
SVM-PDT (β=8)	S	D	0.916	0.626	[45]
disPseAAC	S	D	0.922	0.721	[83]
SVM-PDT-Profile(β=8, n=2)	P	D	0.950	0.740	[45]
PseAAcIndex-Profile(λ=5)	P	D	0.922	0.712	[84]
SVM-WCM (k=6)	S	D	0.904	0.447	[93]
SVM-Top-n-gram (n =2)	P	D	0.923	0.713	[18]

[a] The second column denotes the feature type of protein representation: 'S' represents the sequence-based feature; 'P' represents the profile-based feature; and 'H' represents the HMM-based feature.
[b] The third column denotes the detection strategy: 'A' represents the alignment methods; 'D' represents the discriminative methods; and 'R' represents the ranking methods.
[c] The 'NA' means not available.

Table 3. Performance comparison among different alignment and ranking methods on SCOP-O [24, 98]

Methods	Protein representation[a]	Detection strategy[b]	ROC1[c]	ROC50
PSI-BLAST	P	A	0.750	0.800
HHblits	H	A	0.840	0.882
Coma	H	A	0.699	0.779
Hmmer	H	A	0.789	0.792
RankProp	P	R	NA	0.707
ProtEmbed	H	R	0.814	0.890
ProtDec-LTR (HHblits+ProtEmbed)	H+P	R	0.844	0.902

[a] The second column denotes the feature type of protein representation: 'S' represents the sequence-based feature; 'P' represents the profile-based feature; and 'H' represents the HMM-based feature.
[b] The third column denotes the detection strategy: 'A' represents the alignment methods; 'D' represents the discriminative methods; and 'R' represents the ranking methods.
[c] The 'NA' means not available.

the evolutionary similarity between a pair of proteins, such as sequence alignment, profile alignment and HMM alignment. Alignment strategy is straightforward, and it is suitable for large-scale analysis of homologous relationship. However, some important features, such as Top-n-gram [18, 44] and physicochemical properties [44, 45, 83], are hard to be incorporated into these alignment methods, which prevent the performance improvement. As shown in Table 3, PSI-BLAST only achieves an ROC50 score of 0.800, which is lower than the performance of the top-performing method ProtDec-LTR by 10.2%.

Another kind of popular methods is discriminative methods, which partly overcome the disadvantages of the alignment methods by using the kernel tricks. The performance of these methods depends on the kernel functions. In this study, the performance of 30 widely used discriminative methods is compared, and their results are shown in Table 2, from which, we can see that the kernel functions constructed based on profiles outperform the ones constructed based on sequence information and HMM alignments. However, as discuss above, these discriminative methods are difficult to be used in real-world applications.

Recently, the ranking strategy is attracting more and more attentions, because it takes the advantages of both alignment strategy and discriminative strategy by introducing the discriminative models into the alignment framework. As shown in Table 3, the ranking methods outperform the alignment-based methods, and the ProtDec-LTR achieves the best performance.

The existing open problems of computational methods for protein remote homology detection

(i) The influence of benchmark data set

Although several benchmark data sets have been established to evaluate the performance of various methods as introduced and discussed above, they suffer from three main shortcomings: redundancy data, sample number limitation and class-imbalanced data.

Redundancy data are one important factor for constructing the benchmark data set. If a benchmark data set contains protein sequences with high sequence identity, the performance of methods tested on this data set would be overestimated when using cross-validation strategy. However, for new proteins from an independent data set, their performance would decrease significantly. To avoid this problem, for discriminative strategy, the predictors are trained on proteins that are in the same superfamily but not in the same family with test samples, such as SVM-Fisher [59], SVM-Top-n-gram [18], etc. For alignment and ranking strategies, any pairs of proteins in the benchmark data set have <20% sequence similarity, such as FFAS [36], COMER [101], etc.

For discriminative methods and ranking methods, the training set is used to train the prediction models. Therefore, the large-scale benchmark data sets are required to improve the generalization ability, otherwise, the generalization ability of these methods would be reduced significantly. Unfortunately, some superfamilies in database only have a few available samples, for example there are only one family g.1.1.1 (insulin-like) in superfamily g.1.1 according to SCOPe V2.06. Therefore, the discriminative methods and ranking methods cannot perform well for proteins from this superfamiles. Class-imbalanced data are a critical problem in protein remote homology detection. SCOPe is one of the golden data sets for protein classification, which is organized as tree hierarchy with four levels, including class, fold, superfamily and family. The number of proteins in these families varies significantly, for example in SCOPe 2.06, there are only 28 PDB entries in family a.1.1.1, while protein family a.1.1.2 has 1274 PDB entries, leading to an extremely class-imbalanced problem. Trained and tested on such benchmark data set, more false-positive samples would be predicted by a predictor.

(ii) Protein representations have important impact on the predictive performance

As introduced and emphasized in the above section, the feature extraction process has important impact on the performance of the computational methods for protein remote homology detection. In this regard, many protein features have been proposed, such as sequence-based representation, profile-based representation, HMM-based representation, etc. Although these methods do make contributions to the development of the computational methods in this field, more accurate protein representations are still needed [65]. The new representations should consider the sequence-order information, the structure and function information of proteins, etc. Almost all the currently available protein features are extracted based on the experiences and the knowledge of protein structures and functions. Can we extract protein features automatically by using statistical theories only based on the protein sequences? The deep learning and neural networks would be suitable techniques to answer this question, which has shown advantages to automatically learn a smart representation from big data, such as image processing, natural languages, etc. [101–104]. This point will be discussed in the “Conclusion and Perspectives” section.

(iii) The problems in practical applications

Because the number of protein sequences is increasing rapidly, the computational predictors with low computational cost and high precision are more and more urgently needed for the real-world applications.

Although the discriminative approaches achieve the state-of-the-art performance, it is hard to build easy-to-use online web servers. As reports in [100], SVM-LA [59] and SW-PSSM [75] take >550 h for predicting 20 000 proteins in one superfamily. Compared with the discriminative methods, the alignment and ranking methods are promising approaches for real-world application purpose because they are efficient, and the ranking results are easily to interpret. Some commonly used online web servers are summarized in Table 4.

Conclusion and perspectives

Finally, in this part, we will discuss the state and future of computational methods for protein remote homology detection, and provide some practical guidelines for readers who are non-experts in this field. More powerful computational predictors can be developed by considering the following seven aspects:

1. Updated benchmark data sets should be constructed, which should overcome the three main shortcomings discussed in the Discussion section. For redundancy data problem, serval tools, such as CD-HIT [108] and PISCES [109], can be used to remove the redundancy sequences. These software tools are able to efficiently cluster millions of sequences with reasonable time-consuming. For sample number limitation and class-imbalanced data problems, although the labeled data in SCOP/SCOPe [14–16] are limited, the unlabeled data in nonredundant database UniRef [110] are abundant. The semi-supervised learning model [111] would be a suitable

Table 4. Summary of web-servers for protein remote homology detection

Name	Software/server	Source
PSI-BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi	[3]
HMMER3	http://hmmer.janelia.org/	[105]
HHblits	http://toolkit.tuebingen.mpg.de/hhblits	[53]
COMPASS	http://prodata.swmed.edu/compass/	[34]
PROCAIN	http://prodata.swmed.edu/procain/	[106]
COMA	http://www.ibt.lt/bioinformatics/coma/	[37]
PRC	http://www.ibi.vu.nl/programs/prcwww/	[107]
FFAS	http://ffas.burnham.org/	[36]
SPARK-X	http://sparks-lab.org/yueyang/server/SPARKS-X/	[41]

approach for using both labeled and unlabeled data. By using this strategy, several predictors have been proposed, and performance improvement is observed, for example SVM-HUSTLE [78] is an iterative semi-supervised machine learning approach, which can recruit additional sequences that are assigned a significant statistical measure of homology between a pair of sequences. RankProp [97], ProtEmbed [98] and LSTM [100] expand their training database by searching against a large-scale nonredundant database.

2. How to accurate represent the protein sequences is a key to improve the performance of the predictors. However, it is never an easy task because the protein lengths vary significantly. Recently, studies [65, 112] show that the sequence-order information [44, 113], the physicochemical properties [45, 80], pseudo protein representation [18, 89] and profile/HMM representation [37, 49, 51] are useful features for constructing computational predictors, and a web server called Pse-in-One [65] has been constructed to generate these features. These features or their combinations would be useful for constructing more powerful predictors. Furthermore, techniques from other fields, such as image processing, NLP, signal processing, etc., will play more and more important roles to solve this problem. For example, protein sequences and natural languages share some similarities. The protein sequences can be treated as the documents, and the building blocks of proteins, such as K-mers, Top-n-grams, motifs [74], can be treated as the words of the languages [18]. Some 'grammars' of proteins are also discovered, and several computational predictors have been proposed based on them, such as SVM-Top-n-gram [18], iDNA-Prot|dis [114], SVM-WCM [93], etc. These methods provide new idea and techniques to improve the predictive performance for protein remote homology detection. Furthermore, the deep learning has shown powerful representational ability in many other fields, which can be used to automatically extract the discriminative patterns from protein sequences, protein profiles or even protein structures, for example the LSTM [100] uses the model-based recurrent neural network for protein remote homology detection, which outperforms some alignment and discriminative methods, indicating that the deep learning technique works for protein remote homology detection. Performance improvement can be anticipated by applying other advanced deep learning methods, such as Convolutional Neural Networks.
3. The performance of the alignment methods depends on the alignment algorithms, and the scoring function is the key for an alignment algorithm [43]. Various scoring functions for profile–profile alignments have been proposed and compared in [43]. For protein remote homology detection, more sensitive scoring functions are highly needed, which will be a key to improve the predictive performance of the current alignment methods.
4. New machine learning frameworks of combining various detecting strategy should be further explored. The computational methods for protein remote homology detection can be divided into three categories. Because methods from different groups are based on distinct techniques and theories, most of them are complementary. Therefore, new machine learning frameworks for combining these different strategies will contribute to the development of computational predictors in this field, and recent studies show that this approach is a promising way to improve the predictive performance [2, 89, 90].
5. New evolutionary similarity indicators should be explored. The alignment methods assign alignment scores between protein pairs as the indicator to decide whether they are homologous or not. The discriminative methods compute the probabilities of a query belonging to a specific superfamily. The ranking methods measure the spatial distance of protein pairs in the feature space. These three indicators are different but complementary. The alignment score is computed based on the global or LAs, which is important for inferring protein structure. However, the discriminative probability and the spatial distance are more accurate than alignment score for protein remote homology detection. A more accurate evolutionary similarity indicator can be achieved by combining the existing three indicators.
6. Performance measures are important for evaluating the performance of various predictors. ROC1, ROC50 and ROC are the three most widely used performance measures in this field. For example, the performance of alignment methods and ranking methods is often evaluated by ROC1 and ROC50, while the performance of discriminative methods is evaluated by ROC50 and ROC. However, ROC score is more suitable for evaluating the performance of binary classification tasks (homology or not), but it is not able to evaluate the performance for detecting the hierarchy homology relationships in SCOP and SCOPe databases. The normalized discounted cumulative gain criterion [116] is a more suitable performance measure for this purpose because it evaluates the performance of a predictor based on the graded relevance of the candidate entities.
7. For real-world application purpose, accurate protein remote homology detection software tools with low computational cost is critical. However, because most the state-of-the-art methods are based on time-consuming alignment algorithms, high-performance computing is urgently needed to reduce the computing cost. Although several parallel detection systems have been constructed, such as PSI-BLAST [3, 4], HHblist [51], etc., their computational cost is still too high to be applied for large-scale protein database searching tasks. Fortunately, some graphical processor units-based (GPU-based) software tools have been proposed, for example GPU-BLAST [117] is three to four times faster than NCBI-BLAST. GHOSTM [118] is 4–15 times faster than NCBI-BLAST by using four GPUs. GPU-HMMER [119] achieves up to 6.5 times speedup over previous fast single-threaded implementation. Therefore, these GPU computing techniques and tools are important for constructing efficient remote homology detection systems.

Key Points

- As the number of protein sequences is growing rapidly, it is an emergency task to find out an effective computational approach for protein remote homology detection.
- SCOP is the gold standard data set, where proteins are classified into different groups according to their structures and evolutionary relationships.
- The computational approaches for protein remote homology detection are introduced and discussed. Furthermore, comprehensive performance comparisons of these methods are given based on two widely used benchmark data sets.

- Some open questions for protein remote homology detection are discussed, and some useful web server or stand-alone tools are summarized.

Funding

The National High Technology Research and Development Program of China (863 Program) (grant number 2015AA015405), the National Natural Science Foundation of China (grant numbers 61672184, 61300112 and 61573118), the Natural Science Foundation of Guangdong Province (grant number 2014A030313695), the Guangdong Natural Science Funds for Distinguished Young Scholars (grant number 2016A030306008), and the Scientific Research Foundation in Shenzhen (grant number JCYJ20150626110425228).

References

1. Pál C, Papp B, Lercher MJ. An integrated view of protein evolution. *Nat Rev Genet* 2006;**7**:337–48.
2. Liu B, Zhang D, Xu R, *et al*. Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection. *Bioinformatics* 2014;**30**:472–9.
3. Altschul SF, Madden TL, Schäffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
4. Altschul SF, Koonin EV. Iterated profile searches with PSI-BLAST—a tool for discovery in protein databases. *Trends Biochem Sci* 1998;**23**:444–7.
5. Rost B. Twilight zone of protein sequence alignments. *Protein Eng* 1999;**12**:85–94.
6. Blake JD, Cohen FE. Pairwise sequence alignment below the twilight zone. *J Mol Biol* 2001;**307**:721–35.
7. Kim M-S, Pinto SM, Getnet D, *et al*. A draft map of the human proteome. *Nature* 2014;**509**:575–81.
8. Standley DM, Kinjo AR, Kinoshita K, *et al*. Protein structure databases with new web services for structural biology and biomedical research. *Brief Bioinform* 2008;**9**:276–85.
9. Anfinsen CB. *Studies on the Principles that Govern the Folding of Protein Chains*. Singapore: World Scientific, 1972.
10. UniProt Consortium. Ongoing and future developments at the universal protein resource. *Nucleic acids research* 2011;**39**:D214–9.
11. Rose PW, Prlić A, Bi C, *et al*. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res* 2015;**43**:D345–56.
12. Fariselli P, Rossi I, Capriotti E, *et al*. The WWWH of remote homolog detection: the state of the art. *Brief Bioinform* 2007;**8**:78–87.
13. Wan X-F, Xu D. Computational methods for remote homolog identification. *Curr Protein Pept Sci* 2005;**6**:527–46.
14. Andreeva A, Howorth D, Brenner SE, *et al*. SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res* 2004;**32**:D226–9.
15. Murzin AG, Brenner SE, Hubbard T, *et al*. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;**247**:536–40.
16. Fox NK, Brenner SE, Chandonia J-M. SCOPe: structural classification of proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res* 2014;**42**:D304–9.
17. Fox NK, Brenner SE, Chandonia JM. The value of protein structure classification information—surveying the scientific literature. *Proteins* 2015;**83**:2025–38.
18. Liu B, Wang X, Lin L, *et al*. A discriminative method for protein remote homology detection and fold recognition combining top-n-grams and latent semantic analysis. *BMC Bioinformatics* 2008;**9**:510.
19. Orengo CA, Michie A, Jones S, *et al*. CATH–a hierarchic classification of protein domain structures. *Structure* 1997;**5**: 1093–109.
20. Pearl FMG, Bennett C, Bray JE, *et al*. The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res* 2003;**31**:452–5.
21. Greene LH, Lewis TE, Addou S, *et al*. The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. *Nucleic Acids Res* 2007;**35**:D291–7.
22. Bateman A, Coin L, Durbin R, *et al*. The Pfam protein families database. *Nucleic Acids Res* 2004;**32**:D138–41.
23. Finn RD, Coggill P, Eberhardt RY, *et al*. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* 2016;**44**(D1):D279–85.
24. Liu B, Chen J, Wang X. Application of learning to rank to protein remote homology detection. *Bioinformatics* 2015;**31**(21): 3492–8.
25. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 1970;**48**:443–53.
26. Smith TF, Waterman MS. Identification of common molecular subsequences. *J Mol Biol* 1981;**147**:195–7.
27. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990;**215**:403–10.
28. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* 1988;**85**:2444–8.
29. Lesk A. *Introduction to Bioinformatics*. Oxford University Press, Oxford, UK, 2013.
30. Abagyan RA, Batalov S. Do aligned sequences share the same fold? *J Mol Biol* 1997;**273**:355–68.
31. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci USA* 1987;**84**:4355–8.
32. Liu B, Wang S, Dong Q, *et al*. Identification of DNA-binding proteins by combining auto-cross covariance transformation and ensemble learning. *IEEE Trans Nanobioscience* 2016;**15**:328–334.
33. Schäffer AA, Wolf YI, Ponting CP, *et al*. IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics* 1999;**15**:1000–11.
34. Sadreyev RI, Tang M, Kim BH, *et al*. COMPASS server for homology detection: improved statistical accuracy, speed and functionality. *Nucleic Acids Res* 2009;**37**:W90–4.
35. Sadreyev R, Grishin N. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J Mol Biol* 2003;**326**:317–36.
36. Jaroszewski L, Li Z, Cai XH, *et al*. FFAS server: novel features and applications. *Nucleic Acids Res* 2011;**39**:W38–44.
37. Margelevicius M, Laganeckas M, Venclovas C. COMA server for protein distant homology search. *Bioinformatics* 2010;**26**:1905–6.
38. Kelley LA, Sternberg MJ. Protein structure prediction on the Web: a case study using the Phyre server. *Nat Protoc* 2009;**4**: 363–71.

39. Tomii K, Akiyama Y. FORTE: a profile-profile comparison tool for protein fold recognition. *Bioinformatics* 2004;**20**:594–5.
40. Chen K, Kurgan L. PFRES: protein fold classification by using evolutionary information and predicted secondary structure. *Bioinformatics* 2007;**23**:2843–50.
41. Yang Y, Faraggi E, Zhao H, *et al*. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics* 2011;**27**: 2076–82.
42. Gront D, Blaszczyk M, Wojciechowski P, *et al*. BioShell Threader: protein homology detection based on sequence profiles and secondary structure profiles. *Nucleic Acids Res* 2012;**40**:W257–62.
43. Ye X, Wang G, Altschul SF. An assessment of substitution scores for protein profile–profile comparison. *Bioinformatics* 2011;**27**:3356–63.
44. Liu B, Xu J, Zou Q, *et al*. Using distances between top-n-gram and residue pairs for protein remote homology detection. *BMC Bioinformatics* 2014;**15**:S3.
45. Liu B, Wang X, Chen Q, *et al*. Using amino acid physicochemical distance transformation for fast protein remote homology detection. *PLoS One* 2012;**7**:e46633.
46. Krogh A, Brown M, Mian IS, *et al*. Hidden Markov models in computational biology: applications to protein modeling. *J Mol Biol* 1994;**235**:1501–31.
47. Yoon B-J. Hidden Markov models and their applications in biological sequence analysis. *Curr Genom* 2009;**10**:402–15.
48. Eddy SR. Profile hidden Markov models. *Bioinformatics* 1998;**14**:755–63.
49. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 2011;**39**:W29–37.
50. Hughey R, Krogh A. SAM: Sequence alignment and modeling software system. Computer Research Laboratory, Santa Cruz, USA, 1995.
51. Söding J. Protein homology detection by HMM–HMM comparison. *Bioinformatics* 2005;**21**:951–60.
52. Mount DW, Mount DW, *Bioinformatics: Sequence and Genome Analysis*. New York, NY: Cold spring harbor laboratory press, 2001.
53. Remmert M, Biegert A, Hauser A, *et al*. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat Methods* 2012;**9**:173–5.
54. Wistrand M, Sonnhammer EL. Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER. *BMC Bioinformatics* 2005;**6**:99.
55. Sigrist CJ, Cerutti L, De Castro E, *et al*. PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res* 2010;**38**:D161–6.
56. Hulo N, Bairoch A, Bulliard V, *et al*. The 20 years of PROSITE. *Nucleic Acids Res* 2008;**36**:D245–9.
57. Finn RD, Coggill P, Eberhardt RY, *et al*. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* 2016;**44**:D279–D285.
58. Liao L, Noble WS. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J Comput Biol* 2003;**10**:857–68.
59. Saigo H, Vert JP, Ueda N, *et al*. Protein homology detection using string alignment kernels. *Bioinformatics* 2004;**20**: 1682–9.
60. Shen Y, Bax A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J Biomol NMR* 2013;**56**:227–41.
61. Faraggi E, Kloczkowski A. GENN: a general neural network for learning tabulated data with examples from protein structure prediction. *Methods Mol Biol* 2015;**1260**:165–78.
62. Da Silva F, Desaphy J, Bret G, *et al*. IChemPIC: a random forest classifier of biological and crystallographic protein-protein interfaces. *J Chem Inf Model* 2015;**55**:2005–14.
63. Zhao X, Zou Q, Liu B, *et al*. Exploratory predicting protein folding model with random forest and hybrid features. *Curr Proteomics* 2014;**11**:289–99.
64. Vapnik VN, Vapnik V. *Statistical Learning Theory*. New York, NY: Wiley, 1998.
65. Liu B, Liu F, Wang X, *et al*. Pse-in-one: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res* 2015;**43**:W65–71.
66. Lin H, Chen W. Prediction of thermophilic proteins using feature selection technique. *J Microbiol Methods* 2011;**84**: 67–70.
67. Lin H, Ding C, Song Q, *et al*. The prediction of protein structural class using averaged chemical shifts. *J Biomol Struct Dyn* 2012;**29**:643–9.
68. Wang B, Chen P, Huang D-S, *et al*. Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *FEBS Lett* 2006;**580**:380–4.
69. Song L, Li D, Zeng X, *et al*. nDNA-prot: identification of DNA-binding proteins based on unbalanced classification. *BMC Bioinformatics* 2014;**15**:298.
70. Leslie CS, Eskin E, Noble WS. The spectrum kernel: a string kernel for SVM protein classification. In: *Pacific Symposium on Biocomputing*. 2002, pp. 566–75.
71. Leslie CS, Eskin E, Cohen A., *et al*. Mismatch string kernels for discriminative protein classification. *Bioinformatics* 2004;**20**:467–76.
72. Oğul H, Mumcuoğlu E&Uuml;. A discriminative method for remote homology detection based on n-peptide compositions with reduced amino acid alphabets. *BioSystems* 2007;**87**:75–81.
73. Chen W, Feng P, Lin H. Prediction of ketoacyl synthase family using reduced amino acid alphabets. *J Ind Microbiol Biotechnol* 2012;**39**:579–84.
74. Håndstad T, Hestnes AJ, Sætrom P. Motif kernel generated by genetic programming improves remote homology and fold detection. *BMC Bioinformatics* 2007;**8**:1.
75. Rangwala H, Karypis G. Profile-based direct kernels for remote homology detection and fold recognition. *Bioinformatics* 2005;**21**:4239–47.
76. Jaakkola T, Diekhans M, Haussler D. A discriminative framework for detecting remote protein homologies. *J Comput Biol* 2000;**7**:95–114.
77. Webb-Robertson B-J, Oehmen C, Matzke M. SVM-BALSA: remote homology detection based on Bayesian sequence alignment. *Comput Biol Chem* 2005;**29**:440–3.
78. Shah AR, Oehmen CS, Webb-Robertson B-J. SVM-HUSTLE—an iterative semi-supervised machine learning approach for pairwise protein remote homology detection. *Bioinformatics* 2008;**24**:783–90.
79. Cheng X-Y, Huang W-J, Hu S-C, *et al*. A global characterization and identification of multifunctional enzymes. *PLoS One* 2012;**7**:e38979.
80. Yang Y, Tantoso E, Li K-B. Remote protein homology detection using recurrence quantification analysis and amino

acid physicochemical properties. *J Theor Biol* 2008;**252**:145–54.

81. Webb-Robertson B-JM, Ratuiste KG, Oehmen CS. Physicochemical property distributions for accurate and rapid pairwise protein homology detection. *BMC Bioinformatics* 2010;**11**:1.
82. Kawashima S, Pokarowski P, Pokarowska M, *et al.* AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res* 2008;**36**:D202–5.
83. Liu B, Chen J, Wang X. Protein remote homology detection by combining Chou's distance-pair pseudo amino acid composition and principal component analysis. *Mol Genet Genomics* 2015;**290**:1919–31.
84. Liu B, Wang X, Zou Q, *et al.* Protein remote homology detection by combining Chou's pseudo amino acid composition and profile-based protein representation. *Mol Inform* 2013;**32**:775–82.
85. Chou KC. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins* 2001;**43**:246–55.
86. Melvin I, Weston J, Leslie CS, *et al.* Combining classifiers for improved classification of proteins from sequence or structure. *BMC Bioinformatics* 2008;**9**:389.
87. Damoulas T, Girolami MA. Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. *Bioinformatics* 2008;**24**:1264–70.
88. Shen H-B, Chou K-C. Ensemble classifier for protein fold pattern recognition. *Bioinformatics* 2006;**22**:1717–22.
89. Chen J, Liu B, Huang D. Protein remote homology detection based on an ensemble learning approach. *Biomed Res Int* 2016;**2016**:5813645.
90. Liu B, Chen J, Wang S. Protein remote homology detection by combining pseudo dimer composition with an ensemble learning method. *Curr Proteomics* 2016;**13**:86–91.
91. Dong Q-W, Wang X-L, Lin L. Application of latent semantic analysis to protein remote homology detection. *Bioinformatics* 2006;**22**:285–90.
92. Dumais ST. Latent semantic analysis. *Annu Rev Inf Sci Technol* 2004;**38**:188–230.
93. Lingner T, Meinicke P. Word correlation matrices for protein sequence analysis and remote homology detection. *BMC Bioinformatics* 2008;**9**:259.
94. Liu X, Zhao L, Dong Q. Protein remote homology detection based on auto-cross covariance transformation. *Comput Biol Med* 2011;**41**:640–7.
95. Weston J, Elisseeff A, Zhou D, *et al.* Protein ranking: from local to global structure in the protein similarity network. *Proc Natl Acad Sci USA* 2004;**101**:6559–63.
96. Weston J, Kuang R, Leslie C, *et al.* Protein ranking by semi-supervised network propagation. *BMC Bioinformatics* 2006;**7**:S10.
97. Melvin I, Weston J, Leslie C, *et al.* RANKPROP: a web server for protein remote homology detection. *Bioinformatics* 2009;**25**:121–2.
98. Melvin I, Weston J, Noble WS, *et al.* Detecting remote evolutionary relationships among proteins by large-scale semantic embedding. *PLoS Comput Biol* 2011;**7**:e1001047.
99. Gribskov M, Robinson NL. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Comput Chem* 1996;**20**:25–33.
100. Hochreiter S, Heusel M, Obermayer K. Fast model-based protein homology detection without alignment. *Bioinformatics* 2007;**23**:1728–36.
101. Margelevičius M. Bayesian nonparametrics in protein remote homology search. *Bioinformatics* 2016;**32**(18):2744–52.
102. Zhao Z-Q, Huang D-S, Sun B-Y. Human face recognition based on multi-features using neural networks committee. *Pattern Recognit Lett* 2004;**25**:1351–8.
103. Huang D-S, Du J-X. A constructive hybrid structure optimization methodology for radial basis probabilistic neural networks. *IEEE Trans Neural Netw* 2008;**19**:2099–115.
104. Huang D-S. A constructive approach for finding arbitrary roots of polynomials by neural networks. *IEEE Trans Neural Netw* 2004;**15**:477–91.
105. Mistry J, Finn RD, Eddy SR, *et al.* Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res* 2013;**41**:e121.
106. Wang Y, Sadreyev RI, Grishin NV. PROCAIN server for remote protein sequence similarity search. *Bioinformatics* 2009;**25**:2076–7.
107. Brandt BW, Heringa J. webPRC: the profile comparer for alignment-based searching of public domain databases. *Nucleic Acids Res* 2009;**37**:W48–52.
108. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9.
109. Wang G, Dunbrack RL. PISCES: a protein sequence culling server. *Bioinformatics* 2003;**19**:1589–91.
110. Suzek BE, Huang H, McGarvey P, *et al.* UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* 2007;**23**:1282–8.
111. Chapelle O, Scholkopf B, Zien A. Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006)[Book reviews]. *IEEE Trans Neural Netw* 2009;**20**:542–542.
112. Liu B, Liu F, Fang L, *et al.* repDNA: a python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics* 2015;**31**:1307–9.
113. Lingner T, Meinicke P. Remote homology detection based on oligomer distances. *Bioinformatics* 2006;**22**:2224–31.
114. Liu B, Xu J, Lan X, *et al.* 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* 2014;**9**:e106691.
115. Valizadegan H, Jin R, Zhang R, *et al.* Learning to rank by optimizing ndcg measure. In: *Advances in neural information processing systems*. 2009, pp. 1883–1891.
116. Wang Y, Wang L, Li Y, *et al.* A theoretical analysis of NDCG ranking measures. In: *Proceedings of the 26th Annual Conference on Learning Theory* (COLT 2013). 2013.
117. Vouzis PD, Sahinidis NV. GPU-BLAST: using graphics processors to accelerate protein sequence alignment. *Bioinformatics* 2011;**27**:182–8.
118. Suzuki S, Ishida T, Kurokawa K, *et al.* GHOSTM: a GPU-accelerated homology search tool for metagenomics. *PLoS One* 2012;**7**:e36060.
119. Li X, Han W, Liu G, *et al.* A speculative HMMER search implementation on GPU. In: *Parallel and Distributed Processing Symposium Workshops & PhD Forum (IPDPSW), 2012 IEEE 26th International*. 2012, pp. 735–741. IEEE, Washington, DC, USA.