ELSEVIER

COMPUTATIONAL ANDSTRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

Protein profiles: Biases and protocols

Gregor Urban [a,1], Mirko Torrisi [b,1], Christophe N. Magnan [a,1], Gianluca Pollastri [b], Pierre Baldi [a,*]

[a] Department of Computer Science & Institute for Genomics and Bioinformatics, University of California, Irvine, CA 92697, USA
[b] UCD Institute for Discovery, University College Dublin, Dublin, 4, Ireland

ARTICLE INFO

Article history:
Received 14 May 2020
Received in revised form 14 August 2020
Accepted 15 August 2020
Available online 27 August 2020

Keywords:
Evolutionary information
Profile-based prediction
Evaluation protocols

ABSTRACT

The use of evolutionary profiles to predict protein secondary structure, as well as other protein structural features, has been standard practice since the 1990s. Using profiles in the input of such predictors, in place or in addition to the sequence itself, leads to significantly more accurate predictions. While profiles can enhance structural signals, their role remains somewhat surprising as proteins do not use profiles when folding in vivo. Furthermore, the same sequence-based redundancy reduction protocols initially derived to train and evaluate sequence-based predictors, have been applied to train and evaluate profile-based predictors. This can lead to unfair comparisons since profiles may facilitate the bleeding of information between training and test sets. Here we use the extensively studied problem of secondary structure prediction to better evaluate the role of profiles and show that: (1) high levels of profile similarity between training and test proteins are observed when using standard sequence-based redundancy protocols; (2) the gain in accuracy for profile-based predictors, over sequence-based predictors, strongly relies on these high levels of profile similarity between training and test proteins; and (3) the overall accuracy of a profile-based predictor on a given protein dataset provides a *biased* measure when trying to estimate the actual accuracy of the predictor, or when comparing it to other predictors. We show, however, that this bias can be mitigated by implementing a new protocol (EVALpro) which evaluates the accuracy of profile-based predictors as a function of the profile similarity between training and test proteins. Such a protocol not only allows for a fair comparison of the predictors on equally hard or easy examples, but also reduces the impact of choosing a given similarity cutoff when selecting test proteins. The EVALpro program is available in the SCRATCH suite (www.scratch.proteomics.ics.uci.edu) and can be downloaded at: www.download.igb.uci.edu/#evalpro.

© 2020 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Protein structure prediction is usually decomposed into simpler but still difficult tasks like the prediction of secondary structure, relative solvent accessibility, domains, or contact/distance maps. Despite the variety of methods proposed to tackle each of these tasks, the use of evolutionary information, notably sequence profiles, in the input of the prediction methods has been a constant since the 90s when it was shown to significantly improve prediction accuracies [28]. It is not uncommon to report an improvement of 10% or more when using profiles instead of sequences alone.

A key reason for why profiles improve accuracy is that they can amplify structural signals against the noisy background of evolution. For instance, an alternating pattern of buried and exposed residues in a profile, typically signals the presence of an alpha helix on the surface of a protein [3]. The same pattern can be less visible at the level of an individual sequence. However, this observation alone does not provide a full explanation for their usefulness for two main reasons. First, proteins do not use profile information when folding in vivo. Thus, in principle, one may expect sequence-based predictors to be able to achieve the same level of accuracy as profile-based predictors; however, this has not been observed in the past 30 years. Second, over the same time period, the same sequence-based redundancy reduction protocols–initially derived to train and evaluate sequence-based predictors–have been applied to train and evaluate profile-based predictors. However, if we visualize a profile as creating a sort of “ball” around a sequence in protein space, then profiles increase the volume occupied by both training and test sequences, increasing the chance of an overlap, i.e. of information bleeding between training and test sets (Fig. 1), thereby reducing the quality and fairness of

[*] Corresponding author.
E-mail addresses: gurban@uci.edu (G. Urban), cmagnan@ics.uci.edu (C.N. Magnan), gianluca.pollastri@ucd.ie (G. Pollastri), pfbaldi@uci.edu (P. Baldi).
[1] Authors contributed equally to this work.

https://doi.org/10.1016/j.csbj.2020.08.015
2001-0370/© 2020 The Author(s). Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

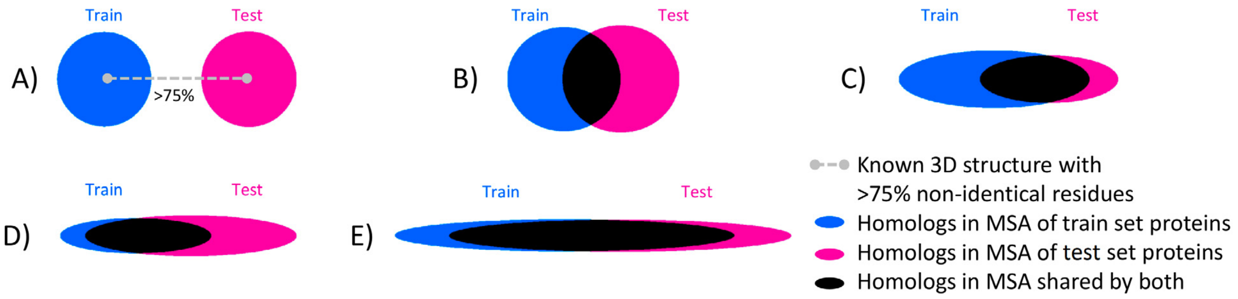
Train
Test
A)
>75%
Train
Test
B)
Train
Test
C)
Train
Test
D)
Train
Test
E)
Known 3D structure with
>75% non-identical residues
Homologs in MSA of train set proteins
Homologs in MSA of test set proteins
Homologs in MSA shared by both

Fig. 1. Illustration of various cases of protein space coverage obtained when using evolutionaryprofiles, derived from multiple sequence alignments (MSA). Figures generated using UNIREF90 data from pairs of proteins sharing less than 25% sequence identity as described in Section 1.2. The area of the ellipsoids corresponds to the number of homologs in the protein's MSA, the distance is inversely related to the sequence identity of the train-test protein pairs, while the overlapping area corresponds to the fraction of shared homologs in both train and test proteins. (A) corresponds to the case initially anticipated in the early 90s (Fig. 2 of Rost and Sander [28]). Cases observed in actual data are categorized into (B), (C), (D), and (E), depending on the fraction of homologs in the train and test groups that are part of the intersection of all homologs, with a threshold of 50%: (B) contains protein pairs sharing less than 50% of their homologs; (C) contains pairs where over 50% of the test set's homologs are part of the intersection, but less than 50% of the train set's homologs are part of the intersection; (D) equivalent to (C) with inverted thresholds; (E) contains protein pairs sharing over 50% of their homologs.

the evaluation. Thus here we set out to study these subtle effects and consider the possibility that the observed gain in accuracy could be at least in part due to an evaluation bias - the bias that results from having sequence-based redundancy reduction protocols for evaluating profile-based predictors.

1.1. Three decades of profile-based predictors

The transition from sequence-based to profile-based predictors occurred after a series of landmark studies in the 80s and 90s, on one side revealing the relationship between sequence and structure [10,8,29], and on the other side providing fast alignment methods to detect putative homologous proteins in large unannotated protein databases [30,20,1,2]. It became clear at this point that a single protein sequence was sufficient to retrieve information about the entire protein family and its evolution. Evolutionary profiles, calculated from multiple sequence alignments (MSA) of the putative homologous proteins and expressed in the form of amino acid frequencies at each alignment position or position-specific scoring matrices (PSSM), were rapidly selected as a solution to represent and incorporate this newly available information into prediction systems in place of the previously used sequence-based features. The resulting gain in accuracy observed in these studies was striking. For instance, sequence-based secondary structure predictors available in the early 90s with an estimated accuracy between 60% and 65% [25] were rapidly replaced by a new generation of profile-based predictors with an estimated accuracy between 70% and 76% [28,19]. Since then, predictors have kept improving thanks notably to more sophisticated prediction methods and larger databases [34,18], but have remained in the same generation of profile-based predictors. The ~10% gain in accuracy initially observed is still visible nowadays as recently showed in Heffernan et al. [16] and Torrisi et al. [32]. Similar trends can be observed for predictors beyond secondary structure to the point that today most state-of-the-art predictors include evolutionary profiles in their input representations.

1.2. Assessing predictors: redundancy versus evolutionary profiles

The general evaluation protocol used in the field to assess the accuracy of profile-based predictors was proposed in Rost and Sander [28]. With this protocol, training and test proteins are selected from the Protein Data Bank (PDB, Gilliland et al. [12]), or from databases directly derived from the PDB like SCOPe [6] or CATH [22], via a redundancy reduction step performed at the sequence identity level using tools like PSI-BLAST [2], CD-HIT [13], or Pisces [33]. This protocol was suggested based on the assumption that a protein sequence sharing less than 25% identity with the protein sequences used to train the predictor is a suitable independent test protein to evaluate a profile-based predictor, an assumption clearly visible from Fig. 2 of Rost and Sander [28] (depicted by case (A) in Fig. 1 of this study). However, we have known for a long time that it isn't particularly unlikely that such a protein may belong to the same family as some of the training proteins [29,4,27,31]. Such cases would inevitably result in high levels of similarity between the corresponding profiles despite the low level of sequence identity (cases (B) to (E) in Fig. 1), and may be the reason for much of the gain in accuracy of profile-based predictors. Likewise, overall differences in accuracy between predictors could result from there being different proportions of cases (A) to (E) in Fig. 1 between their training sets and the sets they are tested on. To create Fig. 1, pairs of proteins are sampled from PDB-derived train and test data sets (reduced to less than 25% sequence identity), while MSA-derived profiles are generated using UNIREF90. To compute the intersection of homologs that are shared between train and test proteins, we ensured that only matches of the same regions of the same homologs are considered, instead of naively counting homologs with potentially disjoint matching regions.

1.3. A profile-induced evaluation bias?

To better understand the mechanism by which evolutionary profiles contribute to a predictor's accuracy, we first designed and implemented a simple evaluation protocol to assess the accuracy of a predictor as a function of maximal profile similarity between training and test proteins. The similarity is calculated using a sliding window of fixed length, comparing any given segment of the test set to all possible sliding window positions in the training set, and choosing the most similar one. We then applied this protocol to six state-of-the-art profile-based secondary structure predictors using both their respective training and test datasets, whenever available, and a separate test dataset specifically prepared for these predictors. The results of this first set of experiments confirm our initial suspicions, notably: (1) high levels of profile similarity between training and test examples can be observed despite the low level of sequence identity between the corresponding proteins; (2) the accuracy of the predictors is strongly correlated to the level of profile similarity; and (3) high levels of profile similarity are necessary for the predictors to perform significantly better than sequence-based predictors. We then confirm that the redundancy between training and test datasets introduced by the use of evolutionary profiles is the consequence of a larger than initially anticipated coverage of the protein space by these profiles (illustrated in Fig. 1). We finally show these conclusions are not specific to secondary structure predictors by showing that comparable results are obtained when using other kinds of predictors. Taken together, these results suggest that the use of evolutionary profiles introduces evaluation biases within the current protocols, and that the level and distribution of profile similarity between training and test sets should be explicitly considered when evaluating and comparing different predictors.

2. Methods

2.1. Profile similarity measure

Measuring the similarity between evolutionary profiles is a problem usually addressed in the context of identifying homologous regions of proteins based on their respective evolutionary profiles. For instance, HHblits [26] detects homologous sequences by performing numerous pairwise alignments of profile HMMs. While remarkably efficient at identifying homologous protein regions, these approaches are more complex and costly to implement than needed in this study. Indeed, the similarity we wanted here is a numerical similarity between the training and test examples of a predictor, independent from the biology, fast to compute on any pair of profiles, and with no need to consider possible insertions in the profiles. We also wanted this measure to be calculated between short protein regions of fixed length in order to address the likely variations of profile similarity along the protein sequences and to limit this measure to a very rough approximation of the information used by the predictors to make a prediction for a given sequence position. We experimented with different window lengths, and a detailed analysis of the effect of varying the window length is given in the Supplementary materials. In the end, we selected a window length of 30 amino acids which provides a good tradeoff between the actual window sizes used by the predictors and the significance of the measure for the selected length. This simplified the problem to measuring the similarity between two 30x20 numerical matrices that we further simplified by flattening the matrices into vectors of dimension 600 without any loss of generality since the matrix structure is not exploited by any of the predictors considered in this study. We selected the cosine similarity measure between the vectors of dimension 600 in our experiments for its desirable properties including notably its low complexity, amplitude independence, and $[0,1]$ boundaries for positive spaces ($[-1,1]$ otherwise). For any profile window A of length 30 in a test protein, we therefore define the similarity of A with the training dataset as the maximum cosine similarity value calculated between A and all profile windows of length 30 in the corresponding training dataset.

2.2. Profile-based secondary structure predictors and corresponding training/test datasets

We selected six widely used protein secondary structure predictors such that (1) evolutionary profiles constitute the largest part of their input features and (2) the training and test datasets used in the corresponding studies were prepared using a 25% sequence identity threshold. Namely, we used:

- SSpro [24,7,21]
- JPred [9,11]
- PSIPRED [19,5]
- SPIDER [15,17]
- Porter [23,32]
- SPOT-1D [14]

For each predictor, we also collected the corresponding training and test datasets whenever made available by the authors and retrieved the protein databases used by these predictors to generate the profiles. A summary of the software releases, datasets, and profile generation methods used in this study is provided in Table 1. Note however that:

- SSpro and Porter were evaluated without using their template-based prediction modules to remove the evaluation bias they would introduce.
- The datasets used for the latest release of PSIPRED were not retained by the authors so we extracted a fairly large representative training set from the PDB snapshot matching with the release date of the predictor using Pisces [33] with default settings.
- SSpro was originally evaluated using a cross-validation procedure on a large protein dataset. We therefore considered this dataset as being only the training set of the predictor in this study.
- Porter's test set is suitable for an independent evaluation of SSpro as all the proteins in this dataset were released after June 2017 and share less than 25% sequence identity with any protein in the training set of SSpro. We also used it to test PSIPRED in our experiments despite the high redundancy levels with the PSIPRED training set mentioned above, and visible in Fig. 2.

2.3. Sequence-based secondary structure predictor

In order to get a baseline accuracy for sequence-based predictions of the secondary structure in our experiments, and to compare the performances between sequence-based and profile-based predictions, we used the most recent predictor we found in that category: SPIDER3_single [16]. Other sequence-based predictors were considered during our study but were all systematically outperformed by SPIDER3_single so only comparisons with this predictor's accuracy are reported here.

2.4. Independent test protein dataset

We extracted from the PDB an independent test set of proteins for the seven predictors listed in Sections 2.2 and 2.3 so that all predictors could be evaluated on the same set of proteins with less than 25% sequence identity with any of the proteins used to train all the predictors, i.e. following the evaluation protocol currently used in the field. PDB entries deposited after February 2017, with less than 25% sequence identity with any of the proteins in the seven training datasets (sequence identity was estimated using PSI-BLAST), and not violating any of the predictor-specific restrictions on the protein sequences (minimal and maximal sequence lengths, no non-standard or unknown amino acids) were selected. The process resulted in 409 such proteins. We name the corresponding protein dataset PDB409.

2.5. Profile similarity based evaluation protocol

We implemented the protocol described below to assess the accuracy of a profile-based predictor as a function of the similarity level calculated between the profiles of its training and test proteins.

1. Evolutionary profiles are extracted for each training and test protein following the methods reported in Table 1, i.e. using the same tools and protein databases as the ones originally used to train each predictor.

Table 1
Description of the profile-based predictors used in this study. The reported release date for each predictor corresponds to the date the models were trained whenever available, to the release date of the software otherwise. The reported date for each dataset is such that all proteins in the corresponding training set had their structure available prior to that date. The protein database release reported in the last column is the release currently used by the online version of the predictors whenever the information was available, a close match with the predictor's release date otherwise.

Predictors		Datasets			Profiles			
Name & Release	Date	Date	Training	Test	Method	Type	Database	Release
SSpro, ACCpro 5.1	10/2013	08/2013	5,772	–	PSI-Blast	Frequency	UniRef50	06/2015
JPred 4	12/2014	07/2014	1,348	149	PSI-Blast	PSSM	UniRef90	07/2014
PSIPRED 4.02	03/2016	03/2016	10,739	–	PSI-Blast	PSSM	UniRef90	05/2016
SPIDER 3	10/2016	06/2014	4,590	1,199	PSI-Blast	PSSM	UniRef90	05/2016
					HHblits	Frequency	UniProt20	02/2016
Porter, PaleAle 5	03/2018	12/2014	15,753	3,154	PSI-Blast	Frequency	UniRef90	05/2016
					HHblits	Frequency	UniProt20	02/2016
SPOT-1D	08/2018	02/2017	10,029	1,213	PSI-Blast	PSSM	UniRef90	04/2018
					HHblits	Frequency	UniClust30	10/2017

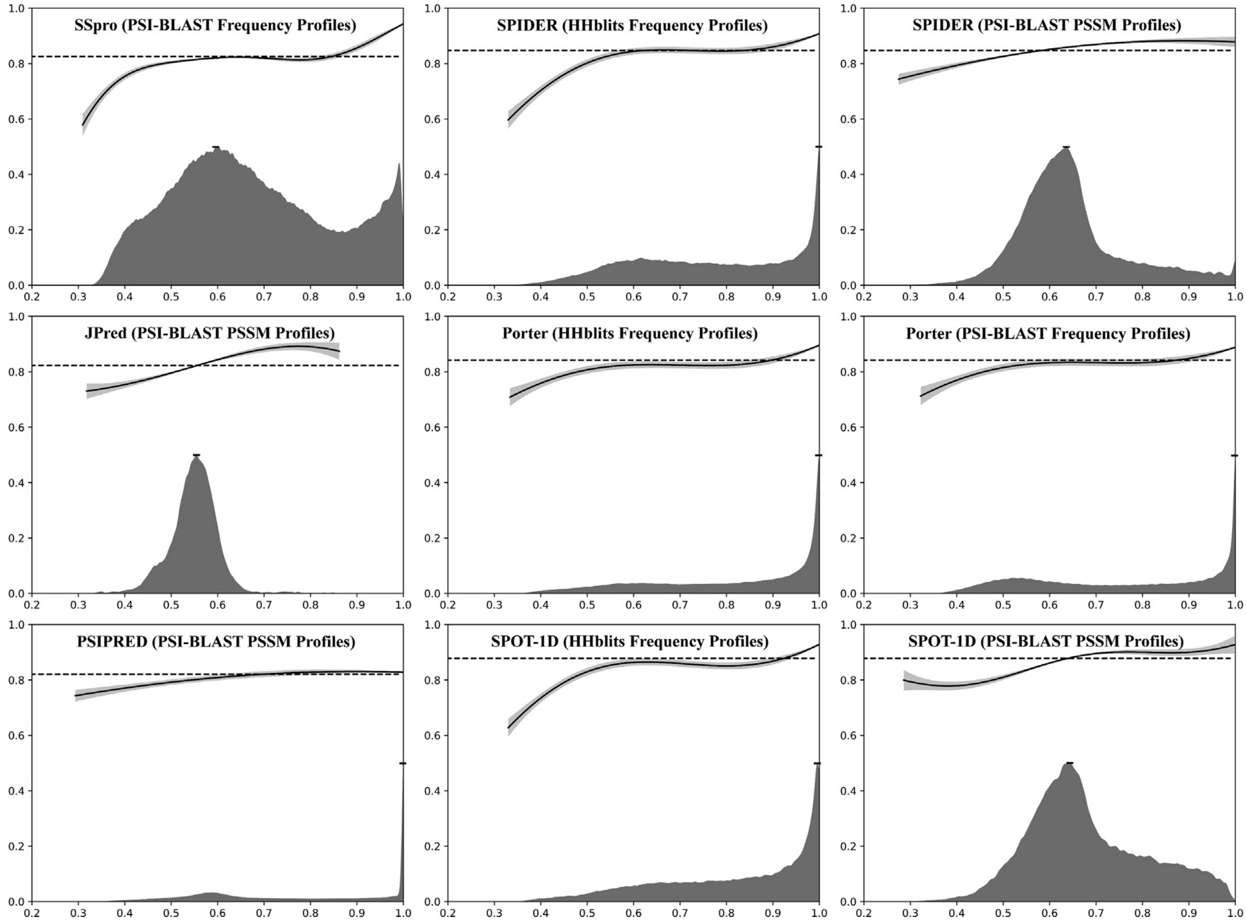

Fig. 2. Evaluation results for the six profile-based secondary structure predictors considered in this study on their own test datasets (Porter's test set for SSpro and PSIPRED) following the protocol described in Section 2.5. The legend for each plot is as follows: (a) the x-axis represents the profile similarity level calculated as indicated in Section 2.1, (b the y-axis represents the estimated predictor accuracy, (c) the GPR-learned functions interpolating the numerous observations at each profile similarity level are drawn using continuous black curves with 95% confidence intervals drawn around the curve in grey color, (d) the predictor's average accuracy on the entire test dataset is depicted using discontinuous horizontal black lines, and (e) the relative frequency of profile fragments observed at each profile similarity level is depicted by plain grey areas in the lower parts of the plots; note that it is re-scaled to reach 50% at its peak (indicated with a dash) for improved visibility, and in some cases almost all fragments concentrate at a cosine similarity of 1.0, as in the case of PSIPRED.

2. Predicted secondary structures are obtained for each test protein by running the corresponding predictor with the same protein database as the one used to generate the profiles in the previous step.
3. Prediction accuracy and profile similarity level with the proteins in the training dataset are calculated for each possible fragment of length 30 in the test proteins using a sliding window approach. Profile similarity levels are calculated as described in Section 2.1. For the three predictors using two different profiles in input, we consider each profile separately and report the results for both types of profiles.
4. We use the implementation of the Gaussian Process Regression (GPR) method publicly available in the scikit-learn python library to interpolate the numerous pairs (profile similarity

level, accuracy) obtained during the previous step. The resulting functions are used to estimate and plot the predictor's accuracy at each profile similarity level.

3. Results

The results obtained following this protocol are reported individually for each the six profile-based secondary structure predictors in Fig. 2 when evaluating the predictors using their own test sets and in Fig. 3 when evaluating them using the PDB409 dataset. A combined view for each set of results and a comparison with the accuracy of the top-performing sequence-based predictor (SPIDER3_single) on the same test datasets are provided respectively in Figs. 4 and 5. Note that the high occurrence of high-similarity protein fragments for PSIPRED in Fig. 2 is due to the unavailability of its training set, forcing us to use a surrogate dataset. Given the resulting atypical plot for PSIPRED in Fig. 2, we conclude that the availability of the actual training set is important for obtaining meaningful results. For the other predictors we observe a clear positive correlation of train-test dataset similarity with higher predictor accuracy. In addition, we also observe that more recent predictors perform better than older ones, which can be due to larger training sets and better profile features due to a larger number of homologs.

3.1. Origin of the redundancy observed when using profiles

The results obtained following the protocol described in Section 2.5 and reported in Figs. 2 and 3 clearly show that high frequencies of test profile windows very similar or identical to the training profile windows can be observed in several cases despite the very low level of sequence identity between the proteins. A quick observation of the corresponding MSAs revealed large intersections between the members of each MSA, providing a natural explanation for the high profile similarity values obtained on these examples. The various cases we observed (illustrated in Fig. 1) suggest that the protein space covered by the members of an MSA may be much larger than initially anticipated in the 90s (depicted by case (A) in Fig. 1) and could be a major factor influencing the overall accuracy of the profile-based predictors developed during the last decades.

We test this assumption by evaluating both a profile-based predictor on sequences and a sequence-based predictor on profiles. The usual process is to only test a predictor on the same type of features that it was trained on. In Table 2 the sequence- and profiles-based predictors are both tested on the one-hot-coded sequence test set, as well as the profile-based test set. One reason for why meaningful results can be expected is that sequence and profile representations are similar: one-hot-coded amino acid sequences are equivalent to profiles in the extreme case of zero identified homologs, while regular profile features can be seen as 'augmented' one-hot-codes, with added values in place of zeros.

On the one hand, this experiment aims to check if the high accuracy of a profile-based predictor is strongly dependent on the large intersections mentioned above by observing the change in accuracy resulting from removing these intersections. On the other side, the experiment aims to check if the low accuracy of sequence-based predictors is improved by adding such intersections. We trained the two predictors using the same encoding for both sequences and profiles so that the two types of input data could be used to evaluate the predictors. Sequences were represented as frequency profiles containing only 0 and 1 values as commonly done by sequence-based predictors. No extra features were used in these experiments. The dataset and methods used to train the two predictors are not detailed here but are identical for each predictor and closely follow the protocols used in Torrisi et al. [32]. The PDB409 dataset described in Section 2.4 is used to evaluate the trained models both on sequences and profiles generated using HHblits. The accuracy of each model on both types of input data is reported in Table 2.

3.2. Profile-based prediction of other structural features

As mentioned in the introduction, the use of evolutionary profiles to predict structural features of a protein is not limited to the secondary structure prediction problem. We performed the same analysis on different kinds of predictors to make sure that the main results of this study are not specific to the profile-based prediction of the secondary structure. All these experiments led to highly similar results and conclusions so we decided to report only some of these results for the three prediction problems listed below.

- Secondary Structure (8-class)
- Relative Solvent Accessibility (2-class)
- Torsion Angles (14-class)

We selected three predictors for each prediction problem among the ones already used during the previous experiments as most of these predictors are also trained to predict other structural features than the secondary structure 3-class, occasionally distributed under a different name. Datasets and profile generation methods reported in Table 1 are also valid for the corresponding predictors evaluated in this experiment. Evaluation results are reported in Fig. 6 and were obtained by evaluating each predictor on its own test dataset similarly to the results reported in Fig. 2.

4. Discussion

Studying the accuracy of profile-based predictors as a function of the profile similarity between training and test datasets provides interesting results at three different levels. First, it reveals the mechanism by which evolutionary profiles increase the prediction accuracy. Second, it reveals that their use introduces a fairly significant evaluation bias with the protocols used in the field. And finally, as profile-similarity is predictive of the accuracy of individual secondary structure predictors, it can be used to create a composite predictor that combines predictions from multiple predictors at the protein fragment level in an informed way.

The first two aspects are discussed in more detail below.

4.1. How profiles improve prediction accuracy

The common belief that evolutionary profiles are more informative for structural feature prediction, compared to single sequences, must be qualified in light of the results obtained in this study. All the predictors evaluated in our experiments show a clear correlation between their prediction accuracy and the level of profile similarity. While the predictors that we compared were released between 2013 and 2018, and were thus trained on very different training sets, we still observe a high degree of similarly between their performance graphs in Figs. 4 and 5. As one would expect, the most recent predictors outperform the oldest ones due to larger training sets, larger protein databases, and higher overall profile similarity. All predictors perform overall poorly on low profile similarity fragments with, in some cases, a level of accuracy which is even below the level achieved by pure sequence-based predictors. And all predictors improve steadily as the profile similarity level increases. A proper evaluation of a predictor should use test proteins that are unrelated to the training proteins. We have seen that this is not systematically the case with

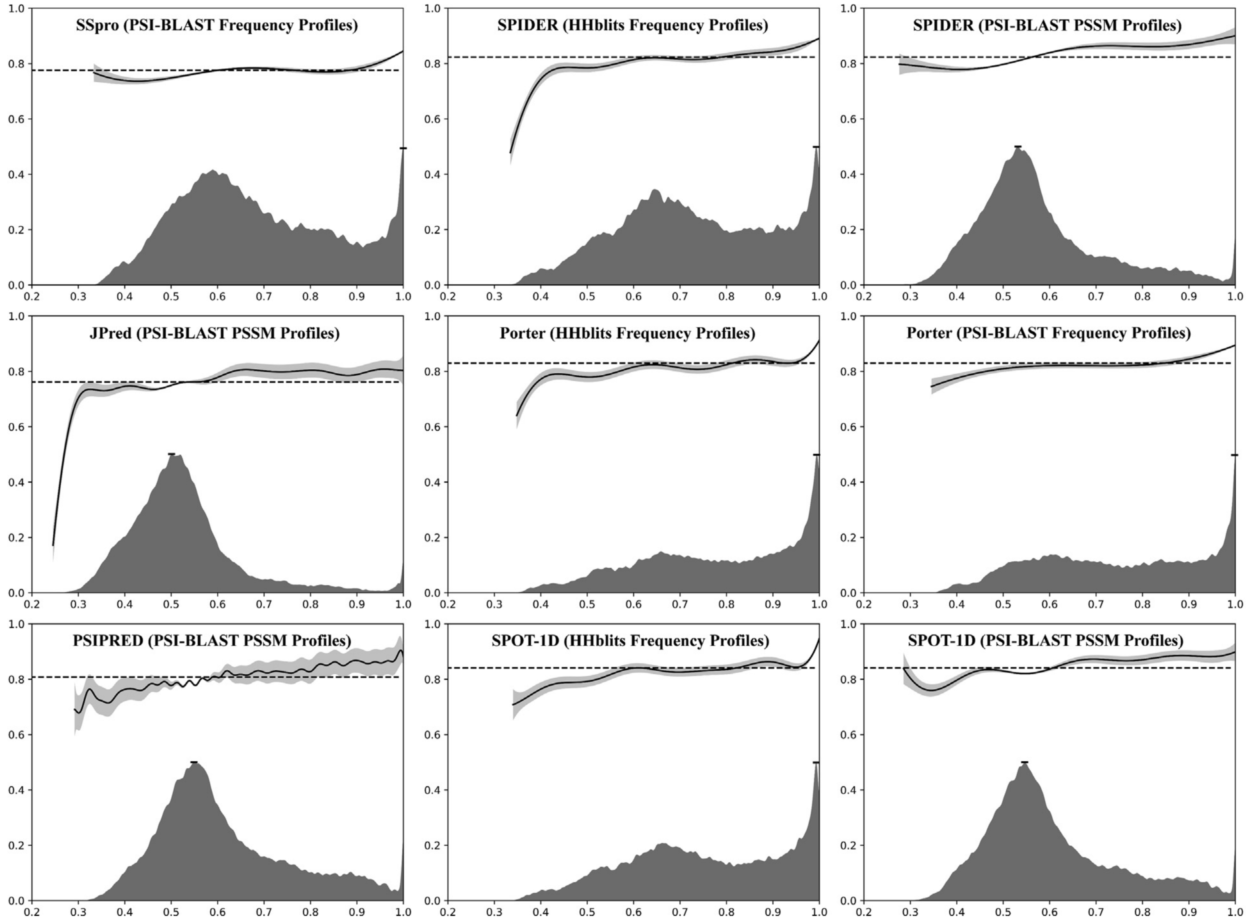

Fig. 3. Evaluation results for the six profile-based secondary structure predictors considered in this study on the PDB409 test dataset using the same representation as Fig. 2. The legend for each plot is as follows: (a) the x-axis represents the profile similarity level calculated as indicated in Section 2.1, (b) the y-axis represents the estimated predictor accuracy, (c) the GPR-learned functions interpolating the numerous observations at each profile similarity level are drawn using continuous black curves with 95% confidence intervals drawn around the curve in grey color, (d) the predictor's average accuracy on the entire test dataset is depicted using discontinuous horizontal black lines, and (e) the relative frequency of profile fragments observed at each profile similarity level is depicted by plain grey areas in the lower parts of the plots; note that it is re-scaled to reach 50% at its peak (indicated with a dash) for improved visibility.

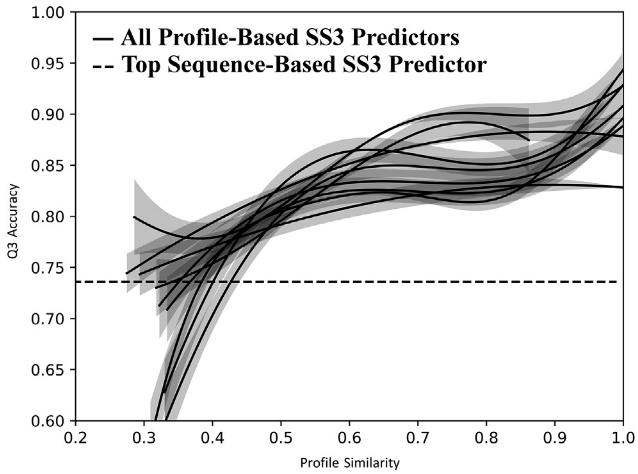

Fig. 4. Superimposed GPR graphs from Fig. 3 for all profile-based secondary structure predictors that are part of this study and tested on their own test datasets. Details such as predictor names are omitted for visibility purposes, as this figure's objective is to compare overall trends. The average accuracy of SPIDER3_single is shown by a dashed line for comparison.

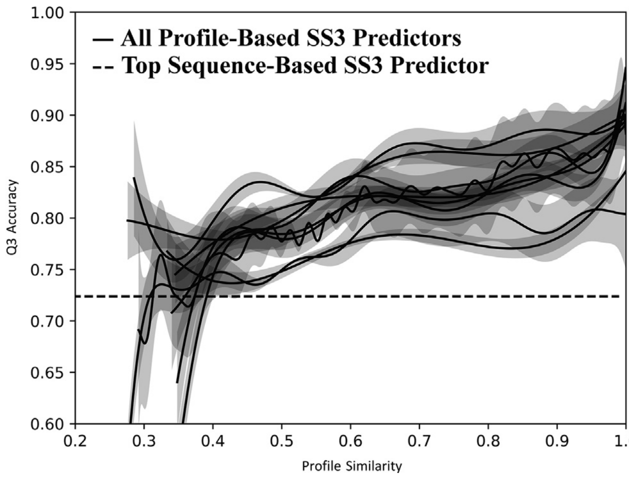

Fig. 5. Superimposed GPR graphs from Fig. 3 for all profile-based secondary structure predictors that are part of this study and tested on the PDB409 dataset. Details such as predictor names are omitted for visibility purposes, as this figure's objective is to compare overall trends. The average accuracy of SPIDER3_single is shown by a dashed line for comparison.

Table 2
Observed accuracy of the profile-based and sequence-based predictors evaluated on the PDB409 dataset as described in Section 3.1. Shown are the accuracies of two predictors, both tested on a sequence- and profiles-based test dataset.

	Tested on sequences
Trained from sequences	72.3%
Trained from profiles	68.6%
	Tested on profiles
Trained from sequences	74.5%
Trained from profiles	81.5%

current protocols where variable levels of redundancy are observed as a result of profiles being calculated from alignments of homologous sequences that contain identical subsets. The results reported in Table 2 show that profile-based predictors are unable to sustain their performances without this redundancy. This result is also visible when performing the opposite experiment, i.e. adding some redundancy between the training and test sequences of a sequence-based predictor, by replacing test sequences with profiles, which ends up boosting the predictor's accuracy. Taken together, these results show that the ∼ 10% accuracy difference between profile-based and sequence-based predictors is in part due to large quantities of highly similar or identical profiles in the training and test datasets, a regime where machine learning methods are naturally expected to be more accurate.

4.2. Consequences for current evaluation protocols

The results obtained in this study are also evidence that the current evaluation protocols used by the community are not adequate to: (1) reliably assess the accuracy of a profile-based predictor; and (2) compare profile-based predictors. Indeed, results reported in Figs. 2, 3, and 6 show that the mean accuracy of a profile-based predictor will strongly depend on the abundance of profile fragments highly similar between training and test sets. From the same results, one can also see that this abundance is not constant at all when using a 25% sequence identity threshold to reduce the redundancy between training and test datasets, leading to important variations of the estimated accuracy of a predictor from a test set to another (up to 6.1% between the results reported in Fig. 2 and 3 for instance). Even comparing the predictors on identical datasets would not solve this issue as the abundance of high similarity profile fragments is also dependent on other factors such as: the method used to generate the profiles; the protein database used to find homologous proteins; and even the type of profiles used in input of the predictors. An evaluation protocol assessing the accuracy of a profile-based predictor as a function of the similarity level between training and test profiles, such as the one implemented in this study, can actually solve this issue since predictors can then be compared at each profile similarity level. Such a protocol not only allows for a fair comparison of the predictors on equally hard or easy examples, but also reduces the impact of choosing a given similarity cutoff when selecting test proteins and is simple enough to not require a lot of computation time. If

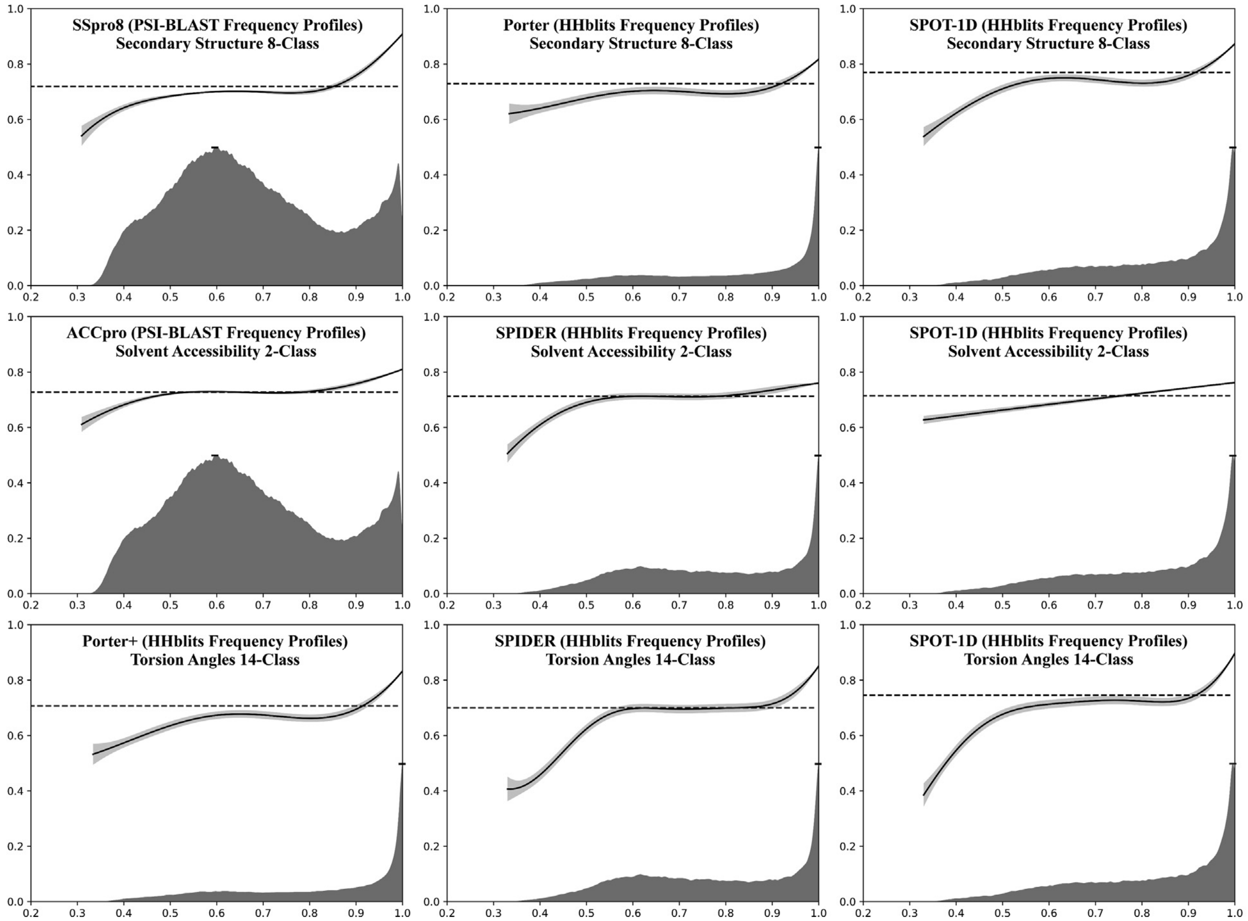

Fig. 6. Evaluation results for profile-based predictors of structural features other than the secondary structure 3-class on their own test dataset using the same representation as in Fig. 2. GPR interpolations of the predictors' accuracies as a function of profile fragment similarity are shown as continuous lines surrounded by 95% confidence intervals. Relative profile fragment frequencies are indicated as a grey area, scaled to reach 50% at their peak for improved visibility.

a protocol, such as the one described here, is used to assess and compare predictors, the need to separate training and test sets using a strict, but arbitrary, sequence identity threshold becomes redundant, leading to the possibility of adopting larger training sets and designing predictors that have a higher accuracy over a larger portion of the protein space.

4.3. Software availability

An implementation of the evaluation protocol proposed in this study, named EVALpro, is available for download, with a full documentation, from the SCRATCH suite [7] at http://scratch.proteomics.ics.uci.edu, or more directly at: www.download.igb.uci.edu/#evalpro. It can be used to either reproduce the analysis presented here or, more generally, to evaluate other profile-based predictors and training/test sets.

5. Conclusion

By probing the mechanisms behind the increase in average accuracy of profile-based predictors versus sequence-based predictors, we have identified an important bias in the way profile-based protein structure predictors have been designed and evaluated in the past 30 years. This increase largely relies on the presence and degree of redundancy between profiles in the training and test sets. As a result, this redundancy produces evaluation biases when current evaluation protocols are used. Despite these results, and somewhat paradoxically, the usefulness of including evolutionary profiles in the predictors' inputs remains unchanged. This is for the same reasons, and with the same limitations, as using template-based prediction methods when templates are available in structural databases. In both cases, a significant improvement in accuracy can be expected; but this improvement occurs only when certain conditions of overlap are met. The continued growth of protein databases benefits profile-based predictors by increasing the number of situations where these favorable overlap conditions occur. Nevertheless, at the opposite end, our work clearly shows that the evaluation protocols used in the field need to be revised to account for the biases associated with these overlaps. In particular, we have shown that measuring average accuracy alone on a protein data set is not particularly meaningful or reliable. Instead, one should measure accuracy as a function of profile similarity. Such a protocol provides a means for evaluating profile-based predictors, and compare them with each other and with sequence-based predictors, in a fairer way.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

Work in part supported by NSF NRT grant 1633631, and NIH grant GM123558 to PB. The work of MT was also supported by the Irish Research Council [GOIPG/2015/3717].

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, athttps://doi.org/10.1016/j.csbj.2020.08.015.

References

[1] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J. Mol. Biol. 1990;215:403–10. https://doi.org/10.1016/S0022-2836(05)80360-2.

[2] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402. https://doi.org/10.1093/nar/25.17.3389. 9254694[pmid].

[3] Benner SA, Gerloff D. Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. Adv. Enzyme Regul. 1991;31:121–81.

[4] Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc. Natl. Acad. Sci. U.S.A. 1998;95:6073–8. https://doi.org/10.1073/pnas.95.11.6073. 9600919[pmid].

[5] Buchan DWA, Jones DT. The PSIPRED Protein Analysis Workbench: 20 years on. Nucleic Acids Res. 2019. https://doi.org/10.1093/nar/gkz297.

[6] Chandonia JM, Fox NK, Brenner SE. SCOPe: Structural Classification of Proteins-extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res. 2013;42:D304–9. https://doi.org/10.1093/nar/gkt1240.

[7] Cheng J, Randall AZ, Sweredoski MJ, Baldi P. Scratch: a protein structure and structural feature prediction server. Nucleic Acids Res. 2005;33:W72–6.

[8] Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J. 1986;5:823–6. 3709526[pmid], PMC1166865 [pmcid].

[9] Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins: Struct. Function Bioinf. 2000;40:502–11.

[10] Doolittle R. Similar amino acid sequences: chance or common ancestry?. Science 1981;214:149–59. https://doi.org/10.1126/science.7280687.

[11] Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. Nucleic Acids Res. 2015;43:W389–94. https://doi.org/10.1093/nar/gkv332.

[12] Gilliland G, Berman HM, Weissig H, Shindyalov IN, Westbrook J, Bourne PE, Bhat TN, Feng Z. The Protein Data Bank. Nucleic Acids Res. 2000;28:235–42. https://doi.org/10.1093/nar/28.1.235.

[13] Godzik A, Li W. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22:1658–9. https://doi.org/10.1093/bioinformatics/btl158.

[14] Hanson J, Paliwal K, Litfin T, Yang Y, Zhou Y. Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks. Bioinformatics 2018. https://doi.org/10.1093/bioinformatics/bty1006.

[15] Heffernan R, Paliwal K, Lyons J, Dehzangi A, Sharma A, Wang J, Sattar A, Yang Y, Zhou Y. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Sci. Rep. 2015;5:11476EP.

[16] Heffernan R, Paliwal K, Lyons J, Singh J, Yang Y, Zhou Y. Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. J. Comput. Chem. 2018;39:2210–6. https://doi.org/10.1002/jcc.25534.

[17] Heffernan R, Yang Y, Paliwal K, Zhou Y. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. Bioinformatics 2017;33:2842–9. https://doi.org/10.1093/bioinformatics/btx218.

[18] Jiang Q, Jin X, Lee SJ, Yao S. Protein secondary structure prediction: a survey of the state of the art. J. Mol. Graph. Model. 2017;76:379–402. https://doi.org/10.1016/j.jmgm.2017.07.015.

[19] Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 1999;292:195–202. https://doi.org/10.1006/jmbi.1999.3091.

[20] Lipman D, Pearson W. Rapid and sensitive protein similarity searches. Science 1985;227:1435–41. https://doi.org/10.1126/science.2983426.

[21] Magnan CN, Baldi P. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. Bioinformatics 2014;30:2592–7. https://doi.org/10.1093/bioinformatics/btu352.

[22] Orengo CA, Lee D, Lees JG, Ashford P, Das S, Lewis TE, Sillitoe I, Dawson NL. CATH: an expanded resource to predict protein function through structure and sequence. Nucleic Acids Res. 2016;45:D289–95. https://doi.org/10.1093/nar/gkw1098.

[23] Pollastri G, McLysaght A. Porter: a new, accurate server for protein secondary structure prediction. Bioinformatics 2004;21:1719–20. https://doi.org/10.1093/bioinformatics/bti203.

[24] Pollastri G, Przybylski D, Rost B, Baldi P. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. Proteins: Struct Function Bioinf 2002;47:228–35. https://doi.org/10.1002/prot.10082.

[25] Qian N, Sejnowski TJ. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 1988;202:865. https://doi.org/10.1016/0022-2836(88)90564-5.

[26] Remmert M, Biegert A, Hauser A, Söding J. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. Nat. Methods 2011;9:173EP.

[27] Rost B. Twilight zone of protein sequence alignments. Protein Eng Design Selection 1999;12:85–94. https://doi.org/10.1093/protein/12.2.85.

[28] Rost B, Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci USA 1993;90:7558–62. https://doi.org/10.1073/pnas.90.16.7558. prefixhttps://www.ncbi.nlm.nih.gov/pubmed/8356056.

[29] Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins: Struct Funct Bioinf 1991;9:56–68. https://doi.org/10.1002/prot.340090107.

[30] Sankoff D, Kruskal J. Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. Addison-Wesley Publishing Company; 1983.

[31] Sauder JM, Arthur JW, Dunbrack Jr RL. Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins: Struct Function Bioinf 2000;40:6–22.

[32] Torrisi M, Kaleel M, Pollastri G. Deeper profiles and cascaded recurrent and convolutional neural networks for state-of-the-art protein secondary structure prediction. Sci Rep 2019;9:1–12.

[33] Wang G, Dunbrack Roland L. J. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589–91. https://doi.org/10.1093/bioinformatics/btg224.

[34] Yang Y, Gao J, Wang J, Heffernan R, Hanson J, Paliwal K, Zhou Y. Sixty-five years of the long march in protein secondary structure prediction: the final stretch?. Briefings Bioinf 2016;19:482–94. https://doi.org/10.1093/bib/bbw129.