Structural bioinformatics

MUFold-SSW: a new web server for predicting protein secondary structures, torsion angles and turns

Chao Fang^{1,†}, Zhaoyu Li^{1,†}, Dong Xu ^{1,2,*} and Yi Shang^{1,*}

¹Department of Electrical Engineering and Computer Science and ²Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA

*To whom correspondence should be addressed.

[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: Arne Elofsson

Received on February 11, 2019; revised on September 11, 2019; editorial decision on September 12, 2019; accepted on September 13, 2019

Abstract

Motivation: Protein secondary structure and backbone torsion angle prediction can provide important information for predicting protein 3D structures and protein functions. Our new methods MUFold-SS, MUFold-Angle, MUFold-BetaTurn and MUFold-GammaTurn, developed based on advanced deep neural networks, achieved state-of-the-art performance for predicting secondary structures, backbone torsion angles, beta-turns and gamma-turns, respectively. An easy-to-use web service will provide the community a convenient way to use these methods for research and development.

Results: MUFold-SSW, a new web server, is presented. It provides predictions of protein secondary structures, torsion angles, beta-turns and gamma-turns for a given protein sequence. This server implements MUFold-SS, MUFold-Angle, MUFold-BetaTurn and MUFold-GammaTurn, which performed well for both easy targets (proteins with weak sequence similarity in PDB) and hard targets (proteins without detectable similarity in PDB) in various experimental tests, achieving results better than or comparable with those of existing methods.

Availability and implementation: MUFold-SSW is accessible at <http://mufold.org/mufold-ss-angle>.

Contact: xudong@missouri.edu or shangy@missouri.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Protein tertiary structure prediction is an important and challenging problem, which has been an active research topic for half a century (Dill and MacCallum, 2012). Some closely related problems, such as protein secondary structure prediction and protein backbone torsion angle prediction, have been studied extensively and their results have been very useful for improving the accuracy of protein tertiary structure prediction.

In recent years, significant improvement has been achieved in protein secondary structure prediction through new deep neural network techniques. In Wang *et al.* (2016), a deep convolutional neural network integrated with a conditional random field was proposed and achieved 68.3% Q8 (8-state) accuracy and 82.3% Q3 (3-state) accuracy on the benchmark CB513 dataset. In Li and Yu (2016), a deep neural network with multi-scale convolutional layers followed by three stacked bidirectional recurrent layers achieved 69.7% Q8 accuracy on CB513. In Busia and Jaitly (2017), a deep convolutional neural network with the next-step conditioning technique achieved 71.4% Q8 accuracy. In Heffernan *et al.* (2017), a long short-term memory bidirectional recurrent neural network (LSTM-biRNN) was proposed and a tool SPIDER3 was developed, achieving 84%

Q3 accuracy. In addition, SPIDER3 also obtained a reduction of 5% in 10% in mean absolute error for Psi-Phi angle prediction Heffernan *et al.* (2015), which represented the best performance before our new methods were published in 2018.

We recently developed MUFold-SS (Fang *et al.*, 2018a) for protein secondary structure prediction, MUFold-Angle (Fang *et al.*, 2018b) for backbone torsion angle prediction, MUFold-BetaTurn (Fang *et al.*, 2018c) for beta-turn prediction, and MUFold-GammaTurn (Fang *et al.*, 2018d) for gamma-turn prediction. Our methods achieved state-of-the-art performance by the time they were published better than or comparable with existing tools. Here, we present the new MUFold-SSW (MUFold Secondary Structure Web server) web server that implemented these four methods for the community to use.

2 Method and results

MUFold-SSW (<http://mufold.org/mufold-ss-angle/>) provides an easy-to-use web service for predicting protein secondary structure (SS), backbone torsion angle (TA), beta-turn (BT) and gamma-turn (GT). It accepts protein sequences as inputs and generates outputs

consisting of predictions of SS, TA, BT and GT. It is a web-based implementation and extension of MUFold-SS, MUFold-Angle, MUFold-BetaTurn and MUFold-GammaTurn that employ different deep learning techniques and architectures (Fang et al., 2018a, b, c, d). In particular, MUFold-GammaTurn used capsule networks (Sabour et al., 2017) to predict gamma turns, the first biological sequence analysis utilizing capsule networks at that time (Fang et al., 2018d). The neural network architecture in MUFOLD-SS is composed of a sequence of nested inception modules and maps the input feature matrix to the predicted probabilities of either eight states or three states of secondary structures (Szegedy et al., 2017). The architecture enables effective processing of local and global interactions between amino acids in making accurate prediction. Similar ideas have been used in the other MUFold methods (He et al., 2016).

The predictions of SS, TA, BT and GT are no longer assessed in CASP (<http://predictioncenter.org/>); instead, the performances of our four new methods were compared with existing tools in our published papers using various public benchmark datasets. In addition, for secondary structure prediction, MUFOLD-SS was also compared to existing tools using our own datasets created through stringent homology reduction processes.

In MUFold-SSW, for Psi-phi angle, beta-turn and gamma-turn predictions, the predictive models are all from our previously published papers (Fang et al., 2018b, c, d).

As part of the MUFold-SSW server development, we retrained and tested the MUFOLD-SS model using a new dataset with stringent homology reduction as follows. We downloaded the CullPDB dataset on June 15, 2017 with 30% sequence identity cutoff, resolution cutoff 3.0 Å and R-factor cutoff 0.25. Then CD-Hit (Li and Godzik, 2006) was applied to it using a 25% sequence identity cutoff to get a new non-redundant dataset that contains 6,000 protein sequences. The new non-redundant CullPDB dataset was divided randomly into training, validation and test subsets with a ratio of 8:1:1. The MUFOLD-SS model was trained using the training and validation subsets, and its performance was reported based on the test subset. Multiple runs of this process were carried out. Finally, the average and standard deviation of the prediction accuracy of the MUFOLD-SS models on the test subsets were Q3 0.8507 (± 0.0013) and Q8 0.7415 (± 0.0014), consistent with the results reported in the original MUFOLD-SS paper (Fang et al., 2018a). The model with the best test performance is used in server MUFold-SSW.

Since the publication of our MUFOLD-SS method, some new methods for secondary structure prediction have been published with very good results. To compare with these newly published methods, we created a new test dataset and conducted an objective performance evaluation as follows. First, we collected recently released proteins from PDBs between February 1, 2019 and May 15, 2019. Then, for each protein, PSI-blast (Altschul et al., 1997) was applied against the whole non-redundant CullPDB dataset we used in retraining of the MUFOLD-SS model for the server implementation as mentioned in the last paragraph. In the PSI_blast results, if a protein had a hit with E-value smaller than 0.005 (i.e. possible homologies), then the protein was removed. This process removed proteins similar to the ones used in training our MUFold-SS models. Finally, 150 proteins were randomly selected from the remaining proteins to form the new test set, which is called PDB150 in this paper (Supplementary Table S1). This test set was used in performance comparison between MUFold-SSW and other secondary structure prediction tools and the results are shown in Supplementary Table S2. It is worth mentioning that for the other tools, their training sets may contain proteins similar to the ones in our new test set PDB150, which would give them some advantage over MUFold-SSW.

Based on experimental results published in (Fang et al., 2018a, b, c, d) and Supplementary Tables S2–S6, MUFold-SSW outperformed SPIDER3 in Psi angle prediction accuracy by more than 5 degrees and Phi angle prediction accuracy by more than 2 degrees. Compared to the new tools, SPOT-1D (Hanson, 2019), NetSurfP-2.0 (Klausen et al., 2019) and Porter5 (Torrisi et al., 2018), MUFold-SSW was slightly worse than SPOT-1D in prediction accuracy, but better than or similar to the other six predictors. Yet,

MUFold-SSW was about 6 times faster than SPOT-1D. In terms of beta-turn prediction, MUFold-SSW outperformed the previous best performing tool, BetaTPred3, significantly in both the overall prediction accuracy and the nine-type beta-turn classification. For gamma-turn prediction, MUFold-SSW achieved an MCC of 0.45 on the GT320 benchmark dataset, which is significantly better than the previous best method with an MCC of 0.38.

The first version of MUFold-SSW was released in June 2017. Since then, there have been about 100–150 submissions per month from dozens of sites around the world, although we have not formally published the web server. Our MUFold-SSW server has been currently used by many researchers (Drori et al., 2018; Guo et al., 2018, 2019; Hanson et al., 2019; Khalatbari et al., 2019; Lu et al., 2018; Ma et al., 2018; Moreno et al., 2019; Oldfield et al., 2019; Zhang et al., 2018).

3 Web server usage

Users may submit protein sequences using our web interface. Submissions will be put into a queue on our web server and processed in a first-come-first-serve manner.

Input. The input to MUFold-SSW includes up to 10 protein sequences in the FASTA format, type of predictions (SS, TA, BT and/or GT), and, optionally, user email address and job name. Currently, the maximum protein sequence length is 700. We will support longer sequences later.

Output. After a job is submitted, it will be assigned a unique job ID and a web URL for tracking job status. If a user submits a valid email address, they will receive the web URL in an email. They can also track their job status on the server web page using their email address or job IDs. On the job running status webpage, the real-time job status and job running log are displayed. When a job is finished, the user will receive another email notification with the link to the job result page. The result webpage includes a job summary, detailed prediction result, distributions and histograms of various predictions and a link to download the final prediction result in a zip file.

Supplementary Figures S1–S3 show examples of job submission page, job running status page and job result page, respectively.

Acknowledgements

We would like to thank the editor and anonymous reviewers for their constructive suggestions to improve the quality of our manuscript and server.

Funding

This work was partially supported by National Institutes of Health grant R01-GM100701.

Conflict of Interest: none declared.

References

- Altschul, S.F. et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Busia, A. and Jaitly, N. (2017) Next-step conditioned deep convolutional neural networks improve protein secondary structure prediction. *arXiv Preprint arXiv: 1702.03865*.
- Dill, K.A. and MacCallum, J.L. (2012) The protein-folding problem, 50 years on. *Science*, **338**, 1042–1046.
- Drori, I. et al. (2018) High quality prediction of protein q8 secondary structure by diverse neural network architectures. *arXiv preprint arXiv: 1811.07143*.
- Fang, C. et al. (2018a) MUFOLD-SS: new deep inception-inside-inception networks for protein secondary structure prediction. *Proteins Struct. Funct. Bioinf.*, **86**, 592–598.
- Fang, C. et al. (2018b) Prediction of protein backbone torsion angles using deep residual inception neural networks. *IEEE/ACM Trans. Comput. Biol. Bioinf.*, **16**, 1020–1028.
- Fang, C. et al. (2018c) MUFold-BetaTurn: a deep dense inception network for protein beta-turn prediction. *Proteins Struct. Funct. Bioinf.*, **86**, 592.

- Fang, C. *et al.* (2018d) Improving protein gamma-turn prediction using inception capsule networks. *Sci. Rep.*, **8**, 15741.
- Guo, Y. *et al.* (2018) Protein secondary structure prediction improved by recurrent neural networks integrated with two-dimensional convolutional neural networks. *J. Bioinf. Comput. Biol.*, **16**, 1850021–1850021.
- Guo, Y. *et al.* (2019) DeepACLSTM: deep asymmetric convolutional long short-term memory neural models for protein secondary structure prediction. *BMC Bioinformatics*, **20**, 341.
- Hanson, J. *et al.* (2019) 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*, **35**, 2403–2410.
- He, K. *et al.* (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.
- Heffernan, R. *et al.* (2017) 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*, **33**, 2842–2849.
- Heffernan, R. *et al.* (2015) Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci. Rep.*, **5**, 11476.
- Khalatbari, L. *et al.* (2019) MCP: a multi-component learning machine to predict protein secondary structure. *Comput. Biol. Med.*, **110**, 144.
- Klausen, M.S. *et al.* (2019) NetSurfP-2.0: improved prediction of protein structural features by integrated deep learning. *Proteins Struct. Funct. Bioinf.*, **87**, 520–527.
- Li, Z. and Yu, Y. (2016) Protein secondary structure prediction using cascaded convolutional and recurrent neural networks. *arXiv preprint arXiv:1604.07176*.
- Li, W. and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
- Lu, Y. *et al.* (2018) A multifunctional oxidosqualene cyclase from *Tripterygium regelii* that produces both α - and β -amyrin. *RSC Adv.*, **8**, 23516–23521.
- Ma, Y. *et al.* (2018) Protein secondary structure prediction based on data partition and semi-random subspace method. *Sci. Rep.*, **8**, 9856.
- Moreno, G.R. *et al.* (2019) Saliency map on CNNs for protein secondary structure prediction. In: *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1249–1253. IEEE.
- Oldfield, C.J. *et al.* (2019) Computational prediction of secondary and supersecondary structures from protein sequences. *Methods Mol. Biol.*, **1958**, 73–100.
- Sabour, S. *et al.* (2017) Dynamic routing between capsules. *Adv. Neural Inf. Process. Syst.*
- Szegedy, C. *et al.* (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. *AAAI*, **4**, 12.
- Torrisi, M. *et al.* (2018) Porter 5: fast, state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes. *bioRxiv*, 289033.
- Wang, S. *et al.* (2016) Protein secondary structure prediction using deep convolutional neural fields. *Sci. Rep.*, **6**, 18962.
- Zhang, B. *et al.* (2018) Prediction of 8-state protein secondary structures by a novel deep learning architecture. *BMC Bioinformatics*, **19**, 293.