SOFTWARE TOOL ARTICLE

shiny-pred: a server for the prediction of protein disordered regions [version 1; peer review: 1 approved with reservations, 1 not approved]

Mauricio Oberti, Iosif Vaisman

School of Systems Biology, George Mason University, Manassas, Virginia, 20110, USA

v1 First published: 28 Feb 2019, 8:230
<https://doi.org/10.12688/f1000research.17669.1>
Latest published: 28 Feb 2019, 8:230
<https://doi.org/10.12688/f1000research.17669.1>

Abstract

Intrinsically disordered proteins or intrinsically disordered regions (IDR) are segments within a protein chain lacking a stable three-dimensional structure under normal physiological conditions. Accurate prediction of IDRs is challenging due to their genome wide occurrence and low ratio of disordered residues, making them a difficult target for traditional classification techniques. Existing computational methods mostly rely on sequence profiles to improve accuracy, which is time consuming and computationally expensive. The shiny-pred application is an *ab initio* sequence-only disorder predictor implemented in R/Shiny language. In order to make predictions, it uses convolutional neural network models, trained using PDB sequence data. It can be installed on any operating system on which R can be installed and run locally. A public version of the web application can be accessed at <https://gmu-binf.shinyapps.io/shiny-pred>

Keywords

Disordered proteins, machine learning, convolutional neural networks, R, Shiny

Open Peer Review

Approval Status ✗ ?

	1	2
version 1 28 Feb 2019	✗ view	? view

1. **Jinbo Xu** , Toyota Technological Institute at Chicago, Chicago, USA
2. **Appadurai Rajeswari**, Indian Institute of Science, Bangalore, India
- Anand Srivastava** , Indian Institute of Science, Bangalore, India

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the **R**Package gateway.

This article is included in the **Artificial Intelligence and Machine Learning** gateway.

Corresponding author: Mauricio Oberti (mauricio.oberti@gmail.com)

Author roles: **Oberti M:** Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Resources, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Vaisman I:** Conceptualization, Supervision, Validation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

Copyright: © 2019 Oberti M and Vaisman I. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Oberti M and Vaisman I. **shiny-pred: a server for the prediction of protein disordered regions [version 1; peer review: 1 approved with reservations, 1 not approved]** F1000Research 2019, 8:230 <https://doi.org/10.12688/f1000research.17669.1>

First published: 28 Feb 2019, 8:230 <https://doi.org/10.12688/f1000research.17669.1>

Introduction

Experimental structure resolution of intrinsically disordered proteins/intrinsically disordered regions (IDP/IDRs) is complex, lengthy and expensive, leading to a variety of computational approaches being developed (He *et al.*, 2009). Over 60 computational protein disorder prediction servers are currently available, although not all publicly. Methods can be classified in one of the following categories (Atkins *et al.*, 2015): (i) *Ab initio* or sequence-based, (ii) clustering, (iii) template based, and (iv) meta or consensus.

shiny-pred is an *ab initio* predictor, which means it relies exclusively on amino acid sequence information to make disordered predictions. It uses prediction models based on convolutional neural networks and reduced protein alphabets. Currently there are three available models, each one built using the same training protein data from PDB (Berman *et al.*, 2000) but differing on the convolutional neural network architecture. Since it doesn't rely in sequence profiles to make predictions, it is fast to be used in proteome-wide disorder scenarios. It performs at the same level or outperforms other state of the art sequence-only methods, achieving accuracy levels of 0.76 and AUC of 0.85 on the publicly available CASP10 dataset (Monastyrskyy *et al.*, 2014), at faster speeds.

Methods

Implementation

shiny-pred is written in the R programming language (R Core Team, 2017) and the shiny web application framework is implemented using the Shiny R package v1.1.0 (Chang, 2018).

Currently, three convolutional neural network models are made available by our application:

(i) cnn-64-ker-local, is a one layer convolutional network (step size 1 and window size of 32) with 64 kernels and local max pooling model; (ii) cnn-128-ker-local, implements one convolutional layer (step size 1 and window size of 32) with 128 kernels and local max pooling model; and (iii) cnn-2-conv-local implements two convolutional layers (64 and 32 kernels) with local max pooling.

The models were created, trained and accessed using the keras R package v2.1.6 (Allaire & Chollet, 2018).

Operation

Our tool has two operation modes; predicting disordered residues in protein sequences (prediction) and benchmarking the predictor performance against sequences with known disorder information (benchmark). The mode is selected automatically based on the format of the input sequences. Users can either upload a sequence file, type/paste a sequence into the text area or select pre-loaded examples from a list.

When in prediction mode, the amino acid sequences are expected to be in FASTA format (Figure 1). In benchmark mode, input sequences in FASTA format are expected to have an additional line

containing the disorder information (D=disorder, O=ordered). Multiple sequences can be submitted at once; several examples for different types of submissions (prediction and benchmark modes) are made available as examples. In both modes, the application will show a result panel, where for each input sequence a graph with the probability of disorder per residue is plotted (Figure 2).

(1) Prediction mode

The workflow for protein disorder prediction is:

- (i) Input the target sequences (in FASTA format) in the text area;
- (ii) Select the model to use for the prediction (default is cnn-128-ker-local) and submit the sequence for prediction;
- (iii) Visualize and download results.

(2) Benchmark mode

In benchmark mode, input sequences are expected to have an extra line with the actual disorder information to be used as benchmark. Result tables will populate two extra columns (actual class and match) with the actual disorder information and if the prediction was accurate for the current residue. An extra panel (Benchmark) shows the ROC curve along with other common binary metrics (sensitivity, specificity, balance accuracy and Matthews correlation coefficient).

Use cases

We use shiny-pred to predict disordered regions within the publicly available CASP10 benchmark dataset. The dataset contains 94 target sequences, each one annotated with the disorder/order information at the residue level. The annotated dataset is provided as an example ('CASP_all') and it can be selected from the example selection list on the 'Sequence Input' tab. Figure 3 shows the input panel after the dataset is selected and loaded. Predictions per sequence can be viewed and downloaded from the 'Results' tab while the 'Benchmark' tab provides a summary of the performance using binary and statistical metrics. Figure 4 shows the server performance for the input dataset, achieving an AUC value of 0.85 and balance accuracy of 0.75.

Summary

This article presents shiny-pred, a sequence-only *ab initio* web application for predicting protein disorder. It's based on reduced amino acid alphabets and convolutional neural networks, being fast and accurate, it is suitable for large proteome-wide experiments.

Software availability

Software available from: <https://gmu-binf.shinyapps.io/shiny-pred>

Source code available from: <https://github.com/mauricioob/shiny-pred>

Archived source code as at time of publication: <https://doi.org/10.5281/zenodo.2567259> (Mauricio, 2019).

License: GNU public license (GPL-3)

Figure 1. Input sequence format (prediction mode).

Figure 2. Prediction results.

Figure 3. Input sequence format (benchmark mode).

Figure 4. Predictor benchmarking.

Grant information

The author(s) declared that no grants were involved in supporting this work.

Acknowledgments

The authors are grateful for the computational facilities provided by Novartis Institutes of Biomedical Research.

References

Atkins JD, Boateng SY, Sorensen T, *et al.*: **Disorder Prediction Methods, Their Applicability to Different Protein Targets and Their Usefulness for Guiding Experimental Studies.** *Int J Mol Sci.* 2015; **16**(8): 19040–19054.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Berman HM, Westbrook J, Feng Z, *et al.*: **The Protein Data Bank.** *Nucleic Acids Res.* 2000; **28**(1): 235–242.

[PubMed Abstract](#) | [Free Full Text](#)

He B, Wang K, Liu Y, *et al.*: **Predicting intrinsic disorder in proteins: an overview.** *Cell Res.* 2009; **19**(8): 929–949.

[PubMed Abstract](#) | [Publisher Full Text](#)

Allaire JJ, Chollet F: **keras: R Interface to “Keras”.** [Accessed: 15 January 2019]. 2018.

[Reference Source](#)

Mauricio: **mauricioob/shiny-pred: Initial release (Version v1.0).** *Zenodo.* 2019.

<http://www.doi.org/10.5281/zenodo.2567259>

Monastyrskyy B, Kryshchuk A, Moulton J, *et al.*: **Assessment of protein disorder region predictions in CASP10.** *Proteins.* 2014; **82** Suppl 2: 127–137.

[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

R Core Team: **R: A Language and Environment for Statistical Computing.**

[Accessed: 13 January 2019]. 2017.

[Reference Source](#)

Chang W: **shiny: Web Application Framework for R.** [Accessed: 13 January 2019]. 2018.

[Reference Source](#)

Open Peer Review

Current Peer Review Status:

Version 1

Reviewer Report 08 July 2019

<https://doi.org/10.5256/f1000research.19321.r50204>

© 2019 Srivastava A et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Appadurai Rajeswari

Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India

Anand Srivastava

Molecular Biophysics Unit, Biological Sciences Division, Indian Institute of Science, Bangalore, India

The authors presented yet another neural network-based disorder prediction tool written in R, trained on PDB data and benchmarked on CASP10 dataset and they claim that the tool outperforms other existing tools in terms of both calculation speed and performance.

We tried using the tool for predicting the known disordered sequences and found that the predictions are accurate and similar to other tools such as IUPRED, DISOPRED3 servers for the well-known disordered sequences such as p53 and Histatin5.

In terms of concerns, I have following comments to make:

1. In general, I find the paper does not describe the motivation, methods and the results in a self-sufficient manner and these could be elaborated further.
2. As the authors state in the paper, there are over 60 tools already existing for disorder prediction. The justification for requiring another tool is not clearly stated.
3. The authors mention they have used PDB data for training the neural network. Do they take all the currently available PDB datasets for training? Does any overlap exist between the datasets trained and benchmarked? The reason why I am asking this is the CASP10 dataset that the authors used for benchmarking has been released in 2012, which would be a subset of the training PDB dataset if they have taken all the PDB data published till date.
4. The authors claim that their method is faster than the existing methods. It would be nice to provide evidence towards that and provide some benchmarking data.
5. AUC and balance accuracy are the two metrics used for evaluating the performance of the

tool. However, a clear definition of these terms are not described in the method section.

6. The tool should be tested and bench marked against a larger data set such as Disorder-723, which contains 723 disorder sequences.

Is the rationale for developing the new software tool clearly explained?

No

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

No

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

No

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

No

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Membrane Biophysics, Protein Structures and Folding, Mechanotransduction, Statistical mechanics of Biological Systems, Integrative Modeling, Multiscale Biomolecular Simulations

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Reviewer Report 11 April 2019

<https://doi.org/10.5256/f1000research.19321.r46176>

© 2019 Xu J. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Jinbo Xu

Toyota Technological Institute at Chicago, Chicago, IL, 60637, USA

This manuscript describes a new protein disorder prediction web server that makes use of (shallow) convolutional neural networks.

There are already many disorder predictors, some of which are based upon deep convolutional neural network and can do prediction directly on amino acid sequence instead of sequence profile. This manuscript does not have sufficient results to justify why one more web server for disorder prediction is needed. Here are some concerns:

1. Please compare with existing, similar methods.
2. It is better to test the method on more recent CASP datasets and make sure that there is no redundancy between training and test data. Ideally, a much larger test set shall be used to evaluate the method.
3. AUC may not be a good metric for disorder prediction since the ratio of disordered residues is quite small.
Precision and Recall may be better.
4. Existing work shall be cited.

Is the rationale for developing the new software tool clearly explained?

Partly

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational biology, machine learning, optimization.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research