SOFTWARE TOOL ARTICLE
shiny-pred: a server for the prediction of protein disordered regions [version 1; referees: awaiting peer review]

Mauricio Oberti, Iosif Vaisman
School of Systems Biology, George Mason University, Manassas, Virginia, 20110, USA

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 \textit{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 \url{https://gmu-binf.shinyapps.io/shiny-pred}

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

This article is included in the \texttt{RPackage} gateway.

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

Author roles: \texttt{Oberti M}: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Resources, Software, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; \texttt{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 Licence, 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; referees: awaiting peer review] \textit{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)
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](https://doi.org/10.1016/S0968-0004(99)00073-8)) 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](https://doi.org/10.1093/nar/gkh959)), at faster speeds.

**Methods**

**Implementation**

shiny-pred is written in the R programming language ([R Core Team, 2017](https://www.r-project.org/)) and the shiny web application framework is implemented using the Shiny R package v1.1.0 ([Chang, 2018](https://cran.r-project.org/web/packages/shiny/)). 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](https://keras.rstudio.com/)).

**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 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 form 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://github.com/mauricioob/shiny-pred](https://github.com/mauricioob/shiny-pred)

Source code available from: [https://github.com/mauricioob/shiny-pred](https://github.com/mauricioob/shiny-pred)

Archived source code as at time of publication: [https://doi.org/10.5281/zenodo.256725](https://doi.org/10.5281/zenodo.256725) (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


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