Web Server

jmb
Journal of Molecular Biology
Special Issue:
Computation Resources for Molecular Biology

ISPRED-SEQ: Deep Neural Networks and Embeddings for Predicting Interaction Sites in Protein Sequences

Matteo Manfredi[†] **Castrense Savojardo**[†] **Pier Luigi Martelli**[*] **and Rita Casadio**

Biocomputing Group, *Dept. of Pharmacy and Biotechnology, University of Bologna, Italy*

Correspondence to Pier Luigi Martelli: *pierluigi.martelli@unibo.it (P.L. Martelli)*

https://doi.org/10.1016/j.jmb.2023.167963
Edited by Michael Sternberg

Abstract

The knowledge of protein–protein interaction sites (PPIs) is crucial for protein functional annotation. Here we address the problem focusing on the prediction of putative PPIs considering as input protein sequences. The issue is important given the huge volume of protein sequences compared to experimental and/or computed structures. Taking advantage of protein language models, recently developed, and Deep Neural networks, here we describe ISPRED-SEQ, which overpasses state-of-the-art predictors addressing the same problem. ISPRED-SEQ is freely available for testing at https://ispredws.biocomp.unibo.it.

Introduction

Proteins are key players in most biological processes. Proteins are social entities and interact with membranes, within themselves or with other proteins, and/or biomolecules (including nucleic acids) to accomplish their functions within the cell. Among all the different features that protein functional annotation requires, it is also important to determine the likelihood of protein–protein interaction. Therefore, effective computational tools for the prediction of protein–protein interactions are important to characterize protein function and to expand interactomes of different species.[1–3]

The identification of Protein-Protein Interaction (PPI) sites, namely protein residues involved in physical interactions within interacting proteins, can be addressed using two complementary approaches. On one hand, different biochemical and biophysical experimental methods (such as X-ray crystallography, nuclear magnetic resonance , alanine scanning mutagenesis and chemical cross-linking) can be applied to determine protein–protein interfaces at the atomic or residue level.[4] Although very accurate, the applicability of these methods to large-scale characterization of PPI is still hampered by economical and technical issues.

On the other hand, computational methods are cost-effective solutions to complement experimental approaches in identifying and characterizing PPI sites. Docking programs are the major class of computational tools to study PPIs [for review, see ref 2]. Very accurate models can be obtained through docking when the two interacting partners are known in advance.

However, when the interacting partner/s is/are not known, machine-learning approaches can compute PPI sites on unbound protein chains. Historically, these methods have been relying on several physicochemical features extracted from protein sequence and/or structure and they can discriminate between interacting and non-interacting residues.[2]

The most accurate approaches are based on information extracted from protein 3D structures. Very informative features include protein solvent accessibility, protrusion, depth indexes, secondary structures, B-factors, and general geometrical features.[5]

Prediction of PPI sites from protein sequence alone is still challenging and methods developed

0022-2836/© 2023 The Author(s). Published by Elsevier Ltd.

for this specific task are less performing than those based on 3D structures. Methods implemented so far for PPI prediction from protein sequence include in input evolutionary information, conservation scores and physical–chemical properties of amino acids (e.g., hydrophobicity, polarity, charge and/or conformational propensities). Additionally, structural features computed from protein sequence with specific classifiers, such as predicted solvent accessibility and secondary structure, are also included with the aim of filling the scoring gap with structure-based approaches. Several methods have been developed in the past and recent years,[2] based mainly on different types of machine learning, including shallow and deep neural networks.[6–15]

Recently, protein language models trained on large volumes of sequence datasets have been proven to be effective in providing protein/residue representations that are alternative and competitive with canonical hand-crafted features such as evolutionary information and physicochemical properties.[17–20] Representations/embeddings provided by these models have been successfully adopted in many prediction tasks.[21–25]

Here we present ISPRED-SEQ, a novel webserver based on a deep-learning model to predict PPI-sites from protein sequence encoded with an embedding procedure. The method stands on a deep architecture combining convolutional blocks and three cascading fully connected layers. ISPRED-SEQ is trained on a dataset of 6,066 protein chains derived from a dataset available in literature[14]. The main novelty of ISPRED-SEQ is the input generation, obtained using two state-of-the-art protein language models, ESM1-b[17] and ProtT5.[18]

We benchmark ISPRED-SEQ on four different independent test data derived from literature.[9,14–15,26–27] All proteins included in the training dataset have less than 25% sequence similarity with sequences in the testing sets, adopting a stringent homology-reduction procedure. Results show that ISPRED-SEQ performs at the state-of-the-art, reporting MCC scores higher than those obtained by other approaches in all the benchmarks performed.

The ISPRED-SEQ web server is freely accessible at https://ispredws.biocomp.unibo.it.

Materials and Methods

Datasets

Training dataset. For training the ISPRED-SEQ network we used a set of protein chains derived from a dataset available in literature[28] and already adopted, after some filtering steps, to train the DELPHI method.[14] The DELPHI dataset comprises 9,982 protein chain sequences extracted from the PDB and sharing no more than 25% pairwise sequence identity. Moreover, the sequences in the training set are also non-redundant (25% identity) with respect to all the sequences included in the independent test datasets (see next section). Starting from this set, we further restricted the number of protein sequences by filtering out all the chains (as in the correspondent UniProt file) having a coverage with the associated PDB structure/s less than 80%, in order to validate PPI annotation on structural experimental evidence. After this filtering step, we ended up with 6,066 protein sequences comprising 1,757,296 residues.

Annotation of PPI sites was then retrieved from the original data available from[28] and manually curated. Starting from the PDB structure of the complex, a residue of a given chain is defined in interaction if the distance between an atom of the residue and an atom of another residue in a different chain is below a given distance threshold, which routinely is set equal to the total sum of the van der Waals' radii of the two atoms plus 0.5 Å[28]. PPI annotations are available for the complete UniProt protein sequences after combining all interaction sites obtained from multiple protein complexes in which each protein is represented, adopting SIFTS[29] for the relative mapping of PDB and UniProt.[28] Overall, our dataset comprises 285,751 interaction sites, corresponding to about 16% of the whole set of residues.

We split the training dataset into 10 different subsets for performing the 10-fold cross validation procedure. Before splitting, we further clustered the sequences at 25% sequence identity and 40% alignment coverage using MMseqs2.[30] The cross-validation split was then performed by randomly distributing complete clusters (instead of individual sequences) among the different subsets. This step is required to capture residual local redundancies between pair of sequences that could have survived the first redundancy reduction performed during dataset construction.

Independent test datasets. To evaluate generalization performance of ISPRED-SEQ and to compare it with other state-of-the-art approaches we used four different independent test sets widely used in literature for comparative evaluation of tools.[9,14–15,26–27] Supplementary Table 1 provide an overview of all datasets used in this study.

The first dataset comprises 448 protein chains used in a review comparing different tools for protein interaction site prediction from sequence.[27] The aim of the authors was to collect data including not only protein–protein interaction sites, but also annotations for DNA, RNA and small-ligand binding sites. For this reason, the dataset was obtained starting from the BioLip database,[31] collecting nucleic-acid and ligand binding site annotations. For the set of proteins retrieved from BioLip, authors also extracted protein–protein interaction sites by analyzing corresponding protein complexes available in the PDB. Protein interaction sites are identified using the same definition adopted for the training set (see above). Internal redundancy of the dataset was set to 25% pairwise sequence identity using the Blastclust tool.[32] We refer to this dataset to as the Dset448.

The second dataset used here is referred to as the Dset335 and it is a subset of the Dset448 introduced in[14] for sake of comparing the methods DELPHI and DLPred.[33] The 335 sequences included in the dataset are indeed selected such that they are non-redundant at 25% sequence identity with the DLPred training set, hence enabling a fair comparison with this method. We used Dset335 to also include DLPred in our benchmark.

The third and fourth datasets, referred to as HomoTE and HeteroTE, respectively, were introduced by Hou and coauthors[9,26]. Recently, these sets were also used for evaluating the performance of the PIPENN prediction tool.[15] HomoTE and HeteroTE include 479 and 48 protein chains from homomeric and heteromeric complexes, respectively. Interface residues are defined in HomoTE and HeteroTE using a slightly different definition based on the computation of Accessible Surface Area (ASA) before and after complex formation: interacting residues are those whose ASA value undergoes a change upon complex formation[26]. Nevertheless, as highlighted in literature,[34] this definition provides very similar or equal interaction interfaces as those based on inter-chain distances.

ISPRED-SEQ implementation

The ISPRED-SEQ general architecture is depicted in Figure 1. Starting from a protein sequence, ISPRED-SEQ input is constructed using two alternative protein language models: i) ESM1-b[17], an encoder-only transformer model trained on about 27 million sequences from UniRef50[35], and ii) ProtT5[18], a sequence-to-sequence model derived from the T5 architecture[36], trained on the large Big Fantastic Database (BFD)[37] comprising 2.1 billion sequences and fine-tuned on the UniRef50 database.

For each residue in the input sequence, ESM1-b and ProtT5 provide embeddings of dimension 1280 and 1024, respectively. These are then concatenated to form a single vector comprising 2304 components for each residue.

Since ESM1-b can only accept input sequences of length lower than 1022, all longer sequences are split into non-overlapping chunks of equal length. After this step, the sequence embedding is reconstructed by concatenating all the chunks.

The joint embedding (ESM1-b + ProtT5) is then processed using a four-layer network. The first layer is a 1-dimensional convolutional neural network with 2304 filters (the number of filters is set as to be equal to the input dimension) and a filter width of 31, corresponding to a window comprising 31 flanking residues and centered at each residue position. The positional output of the convolutional layers is processed by two dense, fully connected layers with 128 and 32 hidden units, respectively. The final output consists of a single unit with a sigmoid activation function. Each residue is classified as interaction site if the output value is greater or equal to 0.5, as not in interaction otherwise.

For sake of assessing the contribution of the input encoding, we also trained alternative models based on different types of inputs, including: the sequence one-hot encoding, providing 20 values per residue, the position-specific scoring matrix (PSSM), computed using two runs of HHblits[38] against the UniClust30 database[39] and providing 20 values per residues, ESM1-b embedding only (1280 values per residue) and ProtT5 embedding only (1024 values per residue). For all the models trained, we adopted the same architecture shown in Figure 1, and changing the number of convolutional filters to be equal to the input dimension (20 for one-hot and PSSMs, 1280 for ESM1-b and 1024 for ProtT5).

Training is performed using minibatches of 64 residues adopting an early stopping procedure that halts the training after 10 epochs without a decrease in the validation loss. The loss that we implemented is a binary cross-entropy and we adopted an Adam optimizer.[40]

To fix all the hyperparameters of the model we performed a grid search using a strict 10-fold cross validation. After that, we retrained the final model on the whole training dataset, and we evaluated it on the different benchmark sets.

Scoring measures

The following measures were used to score performance of the different methods:

- Accuracy (Q_2):

$$Q_2 = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- Precision:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

- Recall:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

- F1-score, the harmonic mean of precision and recall:

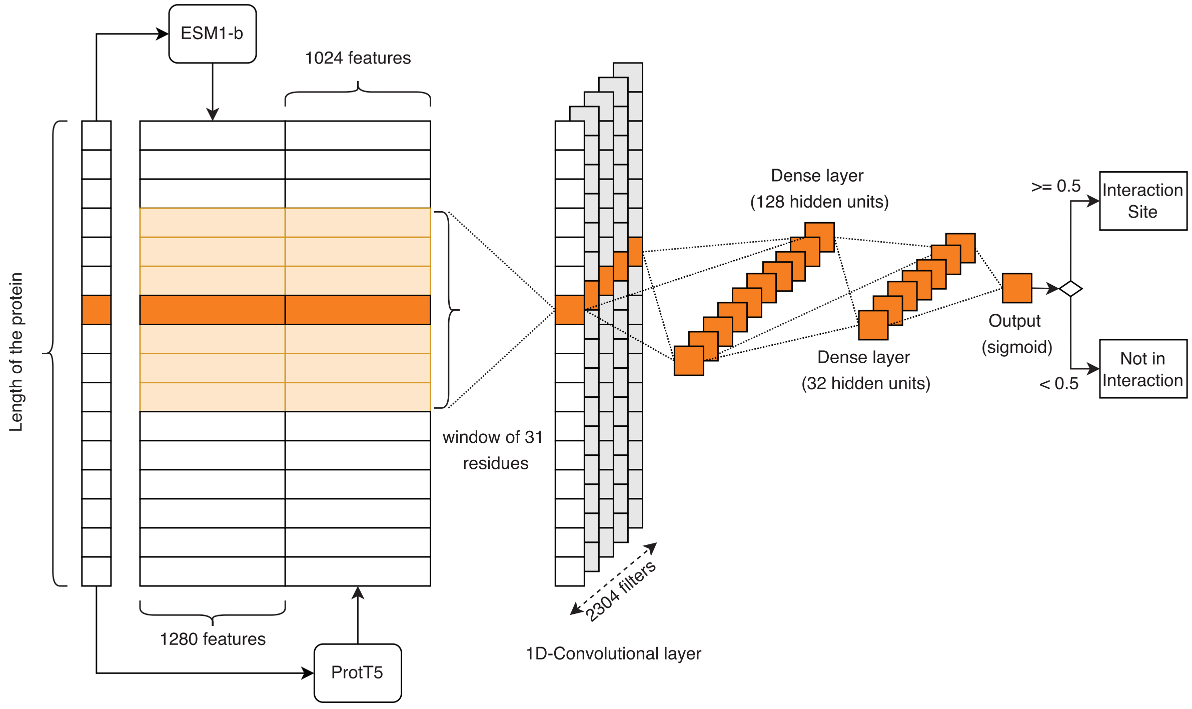


Figure 1. The ISPRED-SEQ deep network architecture. The input sequence is encoded using the two language models (ESM1-b[17] and ProtT5[18]), producing a joint embedding of 2304 features. These are processed using a 1D-Convolutional layer with 2304 filters of size 31. The convolutional output is then processed by two fully connected Dense layers with 128 and 32 hidden units, respectively. The final output is a single unit with sigmoid activation function: each residue is classified as Interaction Site when the output value is greater or equal to 0.5, non-interaction site otherwise (see Materials and Methods for details).

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

- Area Under the Receiver Operating Characteristic Curve (ROC-AUC).
- Matthews Correlation Coefficient (MCC):

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{5}$$

Routinely, the probability value discriminating between positive and negative predictions is set to 0.5. For benchmarking on blind test sets ISPRED-SEQ towards other approaches[14–15,27], we adopted a methodological strategy previously described.[27] According to this procedure, for each of the different methods, a method-specific threshold is introduced to set the number of positive predictions equal to the number of real positive examples.[14–15,27] This procedure allows comparing different methods on the same number of predictions.[27] AUC values are however independent of this procedure.

Results

ISPRED-SEQ performance

For fine tuning ISPRED-SEQ, we tested the network architecture using a 10-fold cross-validation procedure to compare different input encodings. Specifically, we evaluated five different models trained on different inputs, including: i) the sequence one-hot encoding, ii) the sequence profile, iii) the ESM1-b embedding only, iv) the ProtT5 embedding only and v) the joint embedding obtained combining ESM1-b and ProtT5. Supplementary Table 2 lists the results.

Models incorporating canonical features (one-hot and sequence profiles) are both outperformed by embedding-based approaches. MCCs obtained with embedding-based approaches score with values above 0.30 and higher that the 0.14 value obtained with only the sequence profile as input (Supplementary Table 2). Data are shown in Supplementary Table 2, obtained adopting a cross validation procedure. This highlights the effectiveness of language model representations in the task of predicting PPI sites. The two different language models (ESM1-b and ProtT5) provide similar contributions individually achieving comparable MCC scores (0.30 and 0.31, respectively). When combined, the value of MCC is 0.34 (adopting a threshold for positive predictions equal to 0.5), suggesting that the ESM1-b and ProtT5 are complementary, and their combination is advantageous for the problem at hand. This conclusion is further supported by data shown in Supplementary Table 3, where we can observe that predictions made using the two models disagree on roughly 25% of the data (on 14.9% ProtT5 is correct, on 9.3% ESM1-b is correct).

We compared ISPRED-SEQ with state-of-the-art tools, including DELPHI[14], PIPENN,[15] PITHIA[16], SCRIBER[11], SSRWF[8], CRFPPI[41] and LORIS.[42] Table 1 shows the results, and Supplementary Table 4 shows more details regarding the tools adopted for the comparison.

Performance of all methods, with the exclusion of ISPRED-SEQ, are extracted from literature[14–15]. Specifically, performance on Dset448 and Dset335 for DELPHI, SCRIBER, SSRWF, CRFPPI and LORIS are derived from[14], results of PIPENN in all datasets are taken from the original reference paper,[15] and results fro PITHIA are taken from.[16]

All the benchmarked methods provide numerical prediction scores representing the propensity of each input residue to be a PPI site. A threshold must be set to obtain a binary prediction. To compare ISPRED-SEQ performance with other state-of-the-art tools, we adopted the same strategy described in[14–15] and defined in[27] by which binary predictions are obtained using a different threshold for each method so that the number of positive predictions (FP + TP) is equal to the number of real positive examples (TP + FN), or equivalently FP = FN. For our ISPRED-SEQ, performance measures obtained using this strategy are labelled as “th⇒FP = FN” in Table. 1. A direct comparison with the state-of-the-art methods is therefore possible. For sake of completeness, we also show ISPRED-SEQ score obtained using the threshold of 0.5 on the output prediction score. This threshold assumes a probability meaning for the output of ISPRED-SEQ and it is the one adopted in the web server.

Regardless of the method adopted for choosing the threshold, Table 1 indicates that ISPRED-SEQ outperforms all the methods in all the considered datasets. In the Dset448 (the most recent and complete dataset released in literature so far[27]), ISPRED-SEQ achieves a MCC value of 0.39, seven percentage points higher than the one obtained by the second top-performing method, PITHIA.

In the Homo-TE dataset containing homomeric interfaces, ISPRED-SEQ reaches a MCC value of 0.46, again significantly higher than the one registered by PIPENN. Performance on the small Hetero-TE, containing only 48 chains, are lower. However, also in this case, ISPRED-SEQ

Table 1 Comparative benchmark on different independent test sets.

Method	Dataset	MCC	F1	Precision	Recall	Q2	AUC
ISPRED-SEQ (th = 0.5)[°]	Dset448	0.34	0.42	0.29	0.78	0.71	0.82
ISPRED-SEQ (th⇒FP = FN)[°]	Dset448	0.39	0.47	0.47	0.47	0.86	0.82
PITHIA[16] [*]	Dset448	0.32	0.41	0.41	0.41	0.84	0.78
DELPHI[14] [†]	Dset448	0.27	0.37	0.37	0.37	0.83	0.74
PIPENN[15] [‡]	Dset448	0.25	0.39	0.39	0.39	0.79	0.73
SCRIBER[11] [†]	Dset448	0.23	0.33	0.33	0.33	0.82	0.72
SSWRF[8] [†]	Dset448	0.18	0.29	0.29	0.29	0.81	0.69
CRFPPI[41] [†]	Dset448	0.15	0.27	0.26	0.27	0.81	0.68
LORIS[42] [†]	Dset448	0.15	0.27	0.26	0.26	0.81	0.66
ISPRED-SEQ (th = 0.5)[°]	Dset335	0.33	0.40	0.27	0.77	0.72	0.82
ISPRED-SEQ (th⇒FP = FN)[°]	Dset335	0.39	0.46	0.46	0.46	0.87	0.82
PITHIA[16] [*]	Dset335	0.30	0.38	0.38	0.38	0.85	0.76
DELPHI[14] [†]	Dset335	0.28	0.36	0.36	0.36	0.85	0.75
SCRIBER[11] [†]	Dset335	0.23	0.32	0.32	0.32	0.84	0.72
DLPred[33] [†]	Dset335	0.21	0.31	0.31	0.31	0.84	0.72
ISPRED-SEQ (th = 0.5)[°]	Homo_TE	0.42	0.56	0.42	0.83	0.71	0.84
ISPRED-SEQ (th⇒FP = FN)[°]	Homo_TE	0.46	0.58	0.58	0.58	0.81	0.84
PIPENN[15] [‡]	Homo_TE	0.34	0.49	0.49	0.49	0.77	0.77
ISPRED-SEQ (th = 0.5)[°]	Hetero_TE	0.20	0.27	0.17	0.68	0.65	0.72
ISPRED-SEQ (th⇒FP = FN)[°]	Hetero_TE	0.16	0.24	0.24	0.24	0.86	0.72
PIPENN[15] [‡]	Hetero_TE	0.11	0.20	0.20	0.20	0.85	0.66

[*] Data taken from.[16]

[†] Data taken from.[14]

[‡] Data taken from.[15]

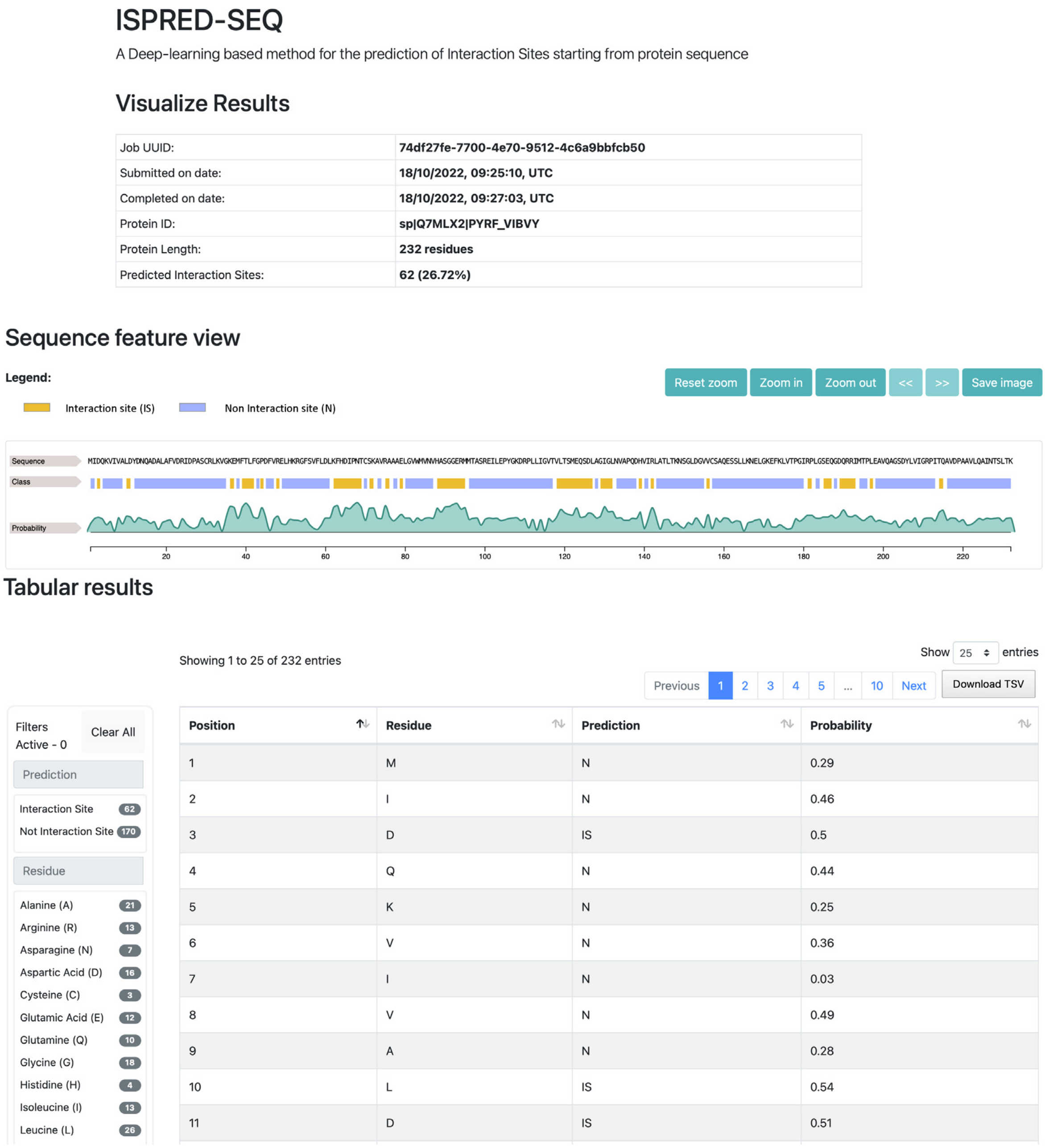
[°] th, threshold value (see Materials and Methods). Performance of all methods different from ISPRED-SEQ are reported considering a prediction threshold that makes equal the numbers of false positive and false negative predictions.[27] Results of ISPRED-SEQ adopting the same strategy are reported (th⇒ FP = FN) as well as those obtained adopting a probability threshold equal to 0.5 (th = 0.5).

outperforms the other tested method (PIPENN) by 5 percentage points, considering the MCC value.

Independently of the procedure adopted for evaluating the scoring indexes, ISPRED-SEQ overpasses the performance of all other methods. This is also evident when considering the AUC values reported in Table 1, totally independent of the strategy adopted for the other scoring indexes.

The ISPRED-SEQ web server

ISPRED-SEQ webserver is available at https://ispredws.biocomp.unibo.it/. The server input interface accepts a single protein sequence in FASTA format with length ranging between 50 and 5000 residues. Upon submission, the user is redirected to the page where results will be available after job completion. The page automatically refreshes every 60s and shows to the user the current status of the job (queued or running). The server also provides the user with a universal job identifier, which can be thereafter used to retrieve job results. The result page (Figure 2) provides information about the job, including i) the identifier, ii) submission and completion time, iii) protein ID, iv) protein length and v) counts of positive and negative predictions. After that, the output of the predictor is shown



ISPRED-SEQ

A Deep-learning based method for the prediction of Interaction Sites starting from protein sequence

Visualize Results

Job UUID:	74df27fe-7700-4e70-9512-4c6a9bbfcb50
Submitted on date:	18/10/2022, 09:25:10, UTC
Completed on date:	18/10/2022, 09:27:03, UTC
Protein ID:	sp|Q7MLX2|PYRF_VIBVY
Protein Length:	232 residues
Predicted Interaction Sites:	62 (26.72%)

Sequence feature view

Legend:

🟨 Interaction site (IS) 🟦 Non Interaction site (N)

Sequence MIDQKVIVALDYDNQADALAFVDRIDPASCRLKVGKEMFTLFGPDFVRELHKRGFSVFLDLKFHDIPNTCSKAVRAAAELGVWMVNVHASGGERMMTASREILEPYGKDRPLLIGVTVLTSMEQSDLAGIGLNVAPQDHVIRLATLTKNSGLDGVVCSAQESSLLKNELGKEFKLVTPGIRPLGSEQGDQRRIMTPLEAVQAGSDYLVIGRPITQAVDPAAVLQAINTSLTK

Tabular results

Showing 1 to 25 of 232 entries

Interaction Site	62
Not Interaction Site	170

Alanine (A)	21
Arginine (R)	13
Asparagine (N)	7
Aspartic Acid (D)	16
Cysteine (C)	3
Glutamic Acid (E)	12
Glutamine (Q)	10
Glycine (G)	18
Histidine (H)	4
Isoleucine (I)	13
Leucine (L)	26

Position	Residue	Prediction	Probability
1	M	N	0.29
2	I	N	0.46
3	D	IS	0.5
4	Q	N	0.44
5	K	N	0.25
6	V	N	0.36
7	I	N	0.03
8	V	N	0.49
9	A	N	0.28
10	L	IS	0.54
11	D	IS	0.51

Figure 2. The ISPRED-SEQ web server output page.

using an interactive viewer. This allows to visualize the whole PPI-site probability profile computed for each residue during the procedure. The page highlights in yellow the predicted PPIs along the sequence. Results are as well summarized in tabular format (Figure 2).

Conclusions

In this paper we present ISPRED-SEQ, a novel method for the prediction of PPI sites from sequence. ISPRED-SEQ novelty is the adoption of input encodings based on embeddings generated by two state-of-the-art protein language models, ESM1-b and ProtT5. In our tests, residue representations based on embeddings outperform canonical feature descriptors such as one-hot encoding and sequence profiles. The scoring index values, although good, still need improvement. However, the major bias is due to the fact that still we do not have a complete picture of all the possible PPIs in a cell, as discussed before.[1–2]

We evaluated ISPRED-SEQ using several independent datasets released in literature and compared its performances against recently state-of-the-art approaches, also based on deep-learning algorithms. In all the tests performed, ISPRED-SEQ significantly outperforms top-scoring methods, reaching MCC scores of 0.39 on recent benchmark datasets containing more than 300 proteins.

We propose ISPRED-SEQ as a valuable tool for the characterization of protein interface residues starting from the protein primary sequence.

We released ISPRED-SEQ as a publicly accessible web server available at https://ispredws.biocomp.unibo.it.

CRediT authorship contribution statement

Matteo Manfredi: Data curation, Formal analysis, Software, Validation, Writing – original draft, Writing – review & editing. **Castrense Savojardo:** Conceptualization, Supervision, Formal analysis, Software, Validation, Writing – original draft, Writing – review & editing. **Pier Luigi Martelli:** Conceptualization, Supervision, Formal analysis, Validation, Writing – original draft, Writing – review & editing. **Rita Casadio:** Conceptualization, Supervision, Formal analysis, Validation, Writing – original draft, Writing – review & editing.

DATA AVAILABILITY

All data are avaiable at: https://ispredws.biocomp.unibo.it/sequence/about/download/

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.

Acknowledgments

The work was supported by PRIN2017 grant (project 2017483NH8_002), delivered to CS by the Italian Ministry of University and Research. We acknowledge ELIXIR-IIB, the Italian node of the ELIXIR infrastructure.

Appendix A. Supplementary Data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jmb.2023.167963.

Received 21 November 2022;
Accepted 10 January 2023;
Available online 13 January 2023

Keywords:
end-to-end models;
embedding;
deep neural networks;
PPI prediction;
protein sequence

† The authors equally contributed to the work.

References

1. Li, S., Wu, S., Wang, L., Li, F., Jiang, H., Bai, F., (2022). Recent advances in predicting protein–protein interactions with the aid of artificial intelligence algorithms. *Curr. Opin. Struct. Biol.* **73**, 102344. https://doi.org/10.1016/j.sbi.2022.102344.
2. Casadio, R., Martelli, P.L., Savojardo, C., (2022). Machine learning solutions for predicting protein–protein interactions. *WIREs Comput. Mol. Sci.*. https://doi.org/10.1002/wcms.1618.
3. Lyon, A.S., Peeples, W.B., Rosen, M.K., (2021). A framework for understanding the functions of biomolecular condensates across scales. *Nat. Rev. Mol. Cell Biol.* **22**, 215–235. https://doi.org/10.1038/s41580-020-00303-z.
4. Rodrigues, J.P.G.L.M., Karaca, E., Bonvin, A.M.J.J., (2015). Information-driven structural modelling of protein-protein interactions. *Methods Mol. Biol.* **1215**, 399–424. https://doi.org/10.1007/978-1-4939-1465-4_18.
5. Savojardo, C., Fariselli, P., Martelli, P.L., Casadio, R., (2017). ISPRED4: interaction sites PREDiction in protein structures with a refining grammar model. *Bioinformatics* **33**, 1656–1663. https://doi.org/10.1093/bioinformatics/btx044.

6. Ofran, Y., Rost, B., (2003). Predicted protein-protein interaction sites from local sequence information. *FEBS Lett.* **544**, 236–239.
7. Murakami, Y., Mizuguchi, K., (2010). Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein-protein interaction sites. *Bioinformatics* **26**, 1841–1848. https://doi.org/10.1093/bioinformatics/btq302.
8. Wei, Z.-S., Han, K., Yang, J.-Y., Shen, H.-B., Yu, D.-J., (2016). Protein-protein Interaction Sites Prediction by Ensembling SVM and Sample-weighted Random Forests. *Neurocomput.* **193**, 201–212. https://doi.org/10.1016/j.neucom.2016.02.022.
9. Hou, Q., De Geest, P.F.G., Vranken, W.F., Heringa, J., Feenstra, K.A., (2017). Seeing the trees through the forest: sequence-based homo- and heteromeric protein-protein interaction sites prediction using random forest. *Bioinformatics* **33**, 1479–1487. https://doi.org/10.1093/bioinformatics/btx005.
10. Hou, Q., De Geest, P.F.G., Griffioen, C.J., Abeln, S., Heringa, J., Feenstra, K.A., (2019). SeRenDIP: SEquential REmasteriNg to DerIve profiles for fast and accurate predictions of PPI interface positions. *Bioinformatics* **35**, 4794–4796. https://doi.org/10.1093/bioinformatics/btz428.
11. Zhang, J., Kurgan, L., (2019). SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics* **35**, i343–i353. https://doi.org/10.1093/bioinformatics/btz324.
12. Qiu, J., Bernhofer, M., Heinzinger, M., Kemper, S., Norambuena, T., Melo, F., Rost, B., (2020). ProNA2020 predicts protein–DNA, protein–RNA, and protein–protein binding proteins and residues from sequence. *J. Mol. Biol.* **432**, 2428–2443. https://doi.org/10.1016/j.jmb.2020.02.026.
13. Zeng, M., Zhang, F., Wu, F.-X., Li, Y., Wang, J., Li, M., (2020). Protein–protein interaction site prediction through combining local and global features with deep neural networks. *Bioinformatics* **36**, 1114–1120. https://doi.org/10.1093/bioinformatics/btz699.
14. Li, Y., Golding, G.B., Ilie, L., (2021). DELPHI: accurate deep ensemble model for protein interaction sites prediction. *Bioinformatics* **37**, 896–904. https://doi.org/10.1093/bioinformatics/btaa750.
15. Stringer, B., de Ferrante, H., Abeln, S., Heringa, J., Feenstra, K.A., Haydarlou, R., (2022). PIPENN: protein interface prediction from sequence with an ensemble of neural nets. *Bioinformatics* **38**, 2111–2118. https://doi.org/10.1093/bioinformatics/btac071.
16. Hosseini, S., Ilie, L., (2022). PITHIA: Protein Interaction Site Prediction Using Multiple Sequence Alignments and Attention. *Int. J. Mol. Sci.* **23**, 12814. https://doi.org/10.3390/ijms232112814.
17. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., et al., (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U S A.* **118** https://doi.org/10.1073/pnas.2016239118. e2016239118.
18. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., et al., (2021). ProtTrans: towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Trans. Pattern. Anal. Mach. Intell.* **PP** https://doi.org/10.1109/TPAMI.2021.3095381.
19. Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., Rost, B., (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinf.* **20**, 723. https://doi.org/10.1186/s12859-019-3220-8.
20. Bepler, T., Berger, B., (2021). Learning the protein language: Evolution, structure, and function. *Cell Syst.* **12**, 654–669.e3. https://doi.org/10.1016/j.cels.2021.05.017.
21. Stärk, H., Dallago, C., Heinzinger, M., Rost, B., (2021). Light attention predicts protein location from the language of life. *Bioinformat. Adv.* **1**, vbab035. https://doi.org/10.1093/bioadv/vbab035.
22. Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T., Rost, B., (2021). Embeddings from deep learning transfer GO annotations beyond homology. *Sci. Rep.* **11**, 1160. https://doi.org/10.1038/s41598-020-80786-0.
23. Teufel, F., Almagro Armenteros, J.J., Johansen, A.R., Gíslason, M.H., Pihl, S.I., Tsirigos, K.D., Winther, O., Brunak, S., et al., (2022). SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nat. Biotechnol.* **40**, 1023–1025. https://doi.org/10.1038/s41587-021-01156-3.
24. Mahbub, S., Bayzid, M.S., (2022). EGRET: edge aggregated graph attention networks and transfer learning improve protein–protein interaction site prediction. *Brief. Bioinform.* **23**, bbab578. https://doi.org/10.1093/bib/bbab578.
25. Singh, J., Litfin, T., Singh, J., Paliwal, K., Zhou, Y., (2022). SPOT-Contact-LM: improving single-sequence-based prediction of protein contact map using a transformer language model. *Bioinformatics* **38**, 1888–1894. https://doi.org/10.1093/bioinformatics/btac053.
26. Hou, Q., Dutilh, B.E., Huynen, M.A., Heringa, J., Feenstra, K.A., (2015). Sequence specificity between interacting and non-interacting homologs identifies interface residues – a homodimer and monomer use case. *BMC Bioinf.* **16**, 325. https://doi.org/10.1186/s12859-015-0758-y.
27. Zhang, J., Kurgan, L., (2018). Review and comparative assessment of sequence-based predictors of protein-binding residues. *Brief. Bioinform.* **19**, 821–837. https://doi.org/10.1093/bib/bbx022.
28. Zhang, J., Ma, Z., Kurgan, L., (2019). Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Brief. Bioinform.* **20**, 1250–1268. https://doi.org/10.1093/bib/bbx168.
29. Dana, J.M., Gutmanas, A., Tyagi, N., Qi, G., O'Donovan, C., Martin, M., Velankar, S., (2019). SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res.* **47**, D482–D489. https://doi.org/10.1093/nar/gky1114.
30. Steinegger, M., Söding, J., (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028. https://doi.org/10.1038/nbt.3988.
31. Yang, J., Roy, A., Zhang, Y., (2013). BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Res.* **41**, D1096–D1103. https://doi.org/10.1093/nar/gks966.
32. Altschul, S., (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.

Nucleic Acids Res. **25**, 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

33. Zhang, B., Li, J., Quan, L., Chen, Y., Lü, Q., (2019). Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network. *Neurocomputing* **357**, 86–100. https://doi.org/10.1016/j.neucom.2019.05.013.
34. Ezkurdia, I., Bartoli, L., Fariselli, P., Casadio, R., Valencia, A., Tress, M.L., (2009). Progress and challenges in predicting protein-protein interaction sites. *Brief. Bioinformatics.* **10**, 233–246. https://doi.org/10.1093/bib/bbp021.
35. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C. H., (2015). UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932. https://doi.org/10.1093/bioinformatics/btu739.
36. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21** 140:5485–140:5551.
37. Steinegger, M., Mirdita, M., Söding, J., (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nat. Methods.* **16**, 603–606. https://doi.org/10.1038/s41592-019-0437-4.
38. Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S.J., Söding, J., (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinf.* **20**, 473. https://doi.org/10.1186/s12859-019-3019-7.
39. Mirdita, M., von den Driesch, L., Galiez, C., Martin, M.J., Söding, J., Steinegger, M., (2017). Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **45**, D170–D176. https://doi.org/10.1093/nar/gkw1081.
40. Kingma D. P. & Ba, J. (2017). Adam: A Method for Stochastic Optimization, ArXiv:1412.6980 [Cs]. http://arxiv.org/abs/1412.6980 (accessed October 19, 2020).
41. Wei, Z.-S., Yang, J.-Y., Shen, H.-B., Yu, D.-J., (2015). A cascade random forests algorithm for predicting protein-protein interaction sites. *IEEE Trans. Nanobiosci.* **14**, 746–760. https://doi.org/10.1109/TNB.2015.2475359.
42. Dhole, K., Singh, G., Pai, P.P., Mondal, S., (2014). Sequence-based prediction of protein-protein interaction sites with L1-logreg classifier. *J. Theor. Biol.* **348**, 47–54. https://doi.org/10.1016/j.jtbi.2014.01.028.