DELPHI: accurate deep ensemble model for protein interaction sites prediction

Yiwei Li[1], **G. Brian Golding**[2], **and Lucian Ilie**[1,*]

[1] University of Western Ontario, CANADA

[2] McMaster University, CANADA

*Corresponding author: ilie@uwo.ca

Abstract

- Protein-protein interaction (PPI) binding sites prediction – an vital problem for biology
  - Experimental methods are time and labor intensive
  - Many computational approaches are proposed: sequence-based ones are very promising
  - The prediction performance of current programs are far from satisfaction
- **DELPHI** - a new sequence-based Deep Learning model for PPI binding sites prediction
  - A novel ensemble model architecture
  - Three novel features
  - More accurate than the leading sequence-based programs

Results

Dataset	Proteins	Residues			binding
		total	binding	non-binding	% of total
Dset_448[9]	448	116,500	15,810	100,690	13.57
Dset_355	355	95,940	11,467	84,473	11.95
Dset_186[2]	186	36,219	5,517	30,702	15.23
Dset_72[2]	72	18,140	1,923	16,217	10.60
Dset_164[1]	164	33,681	6,096	27,585	18.10
Train+Validate	9,982	4,254,198	427,687	3,826,511	10.05

Table 1: The datasets used for training, validation, and testing. The columns give, in order, the dataset names, the number of proteins in each dataset, the total number of residues, the number of binding, and the number of non-binding residues in each dataset, and the percentage of the binding residues out of total.

Table 2: Performance comparison on Dset_448 and Dset_355. Programs are sorted in ascending order by AUPRC. Darker colours represent better results. The evaluation of the programs marked with * is by Zhang.[9]

Predictor	Sens.	Spec.	Prec.	Acc.	F1	MCC	AUROC	AUPRC
Dset_448								
SPPIDER[3]*	0.202	0.870	0.194	0.781	0.198	0.071	0.517	0.159
SPRINT[5]*	0.183	0.873	0.183	0.781	0.183	0.057	0.570	0.167
PSIVER[2]*	0.191	0.874	0.191	0.783	0.191	0.066	0.581	0.170
SPRINGS[4]*	0.229	0.882	0.228	0.796	0.229	0.111	0.625	0.201
LORIS[1]*	0.264	0.887	0.263	0.805	0.263	0.151	0.656	0.228
CRFPPI[7]*	0.268	0.887	0.264	0.805	0.266	0.154	0.681	0.238
SSWRF[6]*	0.288	0.891	0.286	0.811	0.287	0.178	0.687	0.256
SCRIBER[9]	0.334	0.896	0.332	0.821	0.333	0.230	0.715	0.287
DELPHI	0.371	0.901	0.371	0.829	0.371	0.272	0.737	0.337
Dset_355								
SPPIDER	0.180	0.889	0.180	0.804	0.180	0.068	0.515	0.138
SPRINT	0.168	0.886	0.167	0.801	0.168	0.054	0.571	0.150
PSIVER	0.178	0.888	0.177	0.803	0.177	0.065	0.583	0.155
SPRINGS	0.211	0.892	0.210	0.811	0.211	0.103	0.608	0.178
LORIS	0.242	0.896	0.240	0.818	0.241	0.137	0.637	0.203
CRFPPI	0.247	0.897	0.245	0.819	0.246	0.143	0.662	0.214
SSWRF	0.268	0.901	0.268	0.825	0.268	0.168	0.667	0.228
DLPred[8]	0.308	0.906	0.308	0.835	0.308	0.214	0.724	0.272
SCRIBER	0.322	0.908	0.322	0.838	0.322	0.230	0.719	0.275
DELPHI	0.364	0.914	0.364	0.848	0.364	0.278	0.746	0.326

Table 3: Performance comparison on Dset_186, Dset_164, and Dset_72 using the same metrics. Darker colours represent better results.

Predictor	Sens.	Spec.	Prec.	Acc.	F1	MCC	AUROC	AUPRC
Dset_186								
SPPIDER	0.194	0.848	0.186	0.748	0.190	0.041	0.499	0.165
SCRIBER	0.279	0.870	0.279	0.780	0.279	0.150	0.647	0.246
DLPred	0.320	0.878	0.320	0.793	0.320	0.198	0.694	0.290
DELPHI	0.351	0.884	0.351	0.803	0.351	0.235	0.710	0.319
Dset_164								
SPPIDER	0.264	0.828	0.253	0.726	0.258	0.090	0.528	0.220
PSIVER	0.217	0.826	0.216	0.716	0.216	0.043	0.554	0.205
SSWRF	0.266	0.838	0.266	0.734	0.266	0.103	0.606	0.243
CRFPPI	0.280	0.841	0.280	0.739	0.280	0.121	0.608	0.267
SCRIBER	0.327	0.851	0.327	0.756	0.327	0.179	0.657	0.301
DLPred	0.338	0.854	0.338	0.760	0.338	0.192	0.672	0.330
DELPHI	0.352	0.857	0.352	0.765	0.352	0.209	0.685	0.332
Dset_72								
SPPIDER	0.188	0.898	0.179	0.823	0.183	0.084	0.522	0.134
PSIVER	0.152	0.899	0.152	0.820	0.152	0.052	0.604	0.141
CRFPPI	0.248	0.911	0.248	0.840	0.248	0.158	0.669	0.200
SSWRF	0.246	0.911	0.246	0.840	0.246	0.157	0.678	0.198
SCRIBER	0.232	0.909	0.232	0.837	0.232	0.141	0.680	0.198
DLPred	0.246	0.901	0.246	0.826	0.246	0.148	0.688	0.215
DELPHI	0.274	0.914	0.274	0.847	0.274	0.189	0.711	0.237

Ablation study

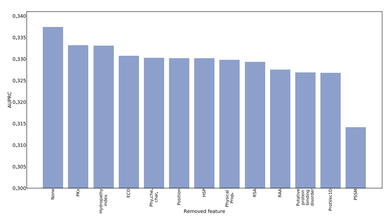

Figure 1: The areas under PR curves with the removal of one out of the twelve features on Dset_448. One feature is removed each time, and the DELPHI model is trained, validated, and tested using the remaining eleven features. The x-axis shows the removed features where 'None' indicates using all twelve features, and the y-axis is the AUPRC achieved. The features are sorted decreasingly by the AUPRC values.

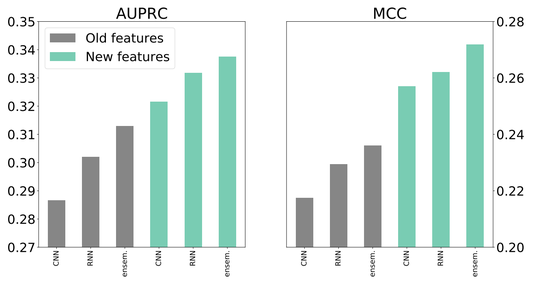

Figure 2: The evaluation of the DELPHI model architecture and the three novel features. The area under PR curves (left) and MCC (right) are plotted separately. Each plot contains the performance of using CNN, RNN, and the ensemble model on Dset_448. Two different colors indicate with and without the three new features.

Evolutionary conservation

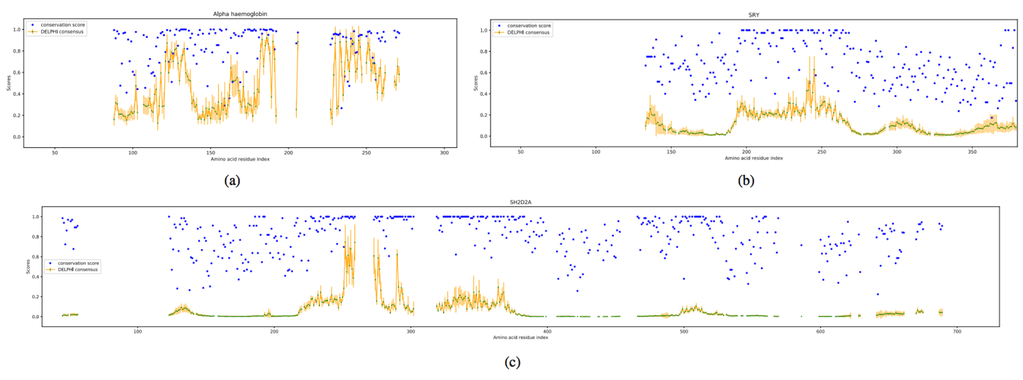

Figure 3: Three proteins were evaluated to compare the PPI predicted by DELPHI (green and orange) with the degree of site-by-site conservation (blue). Only sites represented in ten or more taxa are included resulting in some apparent gaps. The proteins are (a) alpha haemoglobin, (b) SRY and (c) SH2D2A.

Methods

Model Architecture

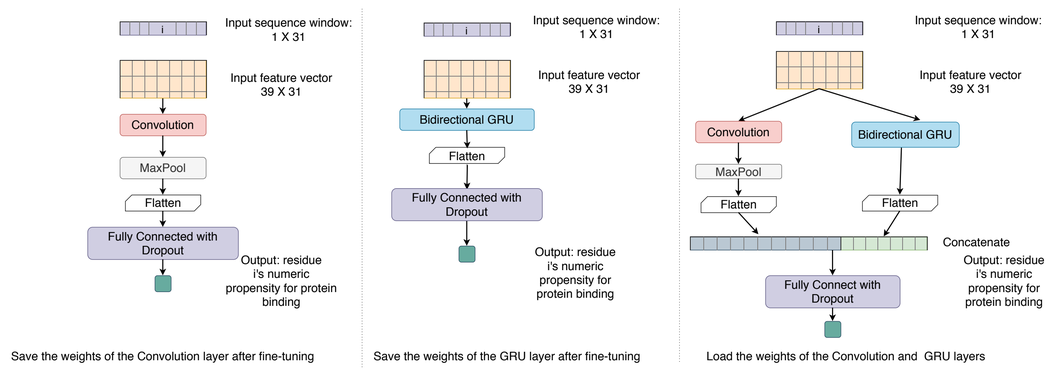

Figure 4: The architecture of DELPHI. Left: the CNN component of the model. Middle: the RNN component of the model. Right: The ensemble model.

Model Input and Output

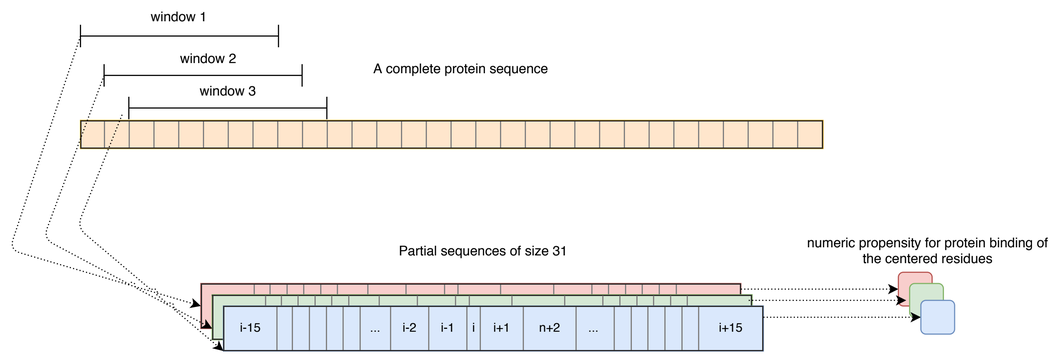

Figure 5: The many-to-one prediction. Sliding windows of size 31, stride 1 are put on top of an input protein sequence. Each time, a sub-sequence of length 31 is extracted. The model predicts the protein-binding propensity of the middle amino acid for each sub-sequence.

Input Features

Feature	Program	Dimension
High-scoring segment pair (HSP)	Compute	1
3-mer amino acid embedding (ProtVec1D)	Load/compute	1
Position information	Compute	1
Position-specific scoring matrix (PSSM)	Psi-Blast	20
Evolutionary conservation (ECO)	Hhblits	1
Putative relative solvent accessibility (RSA)	ASAquick	1
Relative amino acid propensity (RAA)	Load	1
Putative protein-binding disorder	ANCHOR	1
Hydropathy index	Load	1
Physicochemical characteristics	Load	3
Physical properties	Load	7
PKx	Load	1

Table 4: The feature names, computation programs, and dimensions of each feature used by DELPHI. The first three features are novel.

DELPHI Web Server

Figure 6: The interface of the DELPHI web server. It takes protein sequences in FASTA format as input, and the result will be emailed to the user.

Conclusions

- DELPHI is the most accurate sequence-based PPI sites predictor.
- The three novel features and the ensemble architecture can be potentially used in other protein sequence classifiers.

Availability

The source code of DELPHI is freely available from github.com/lucian-ilie/DELPHI/.
All datasets and results as well the DELPHI web server is available from www.csd.uwo.ca/~yli922/index.php.

References

[1] Kaustubh Dhole, Gurdeep Singh, Priyadarshini P Pai, and Sukanta Mondal. Sequence-based prediction of protein–protein interaction sites with l1-logreg classifier. *Journal of theoretical biology*, 348:47–54, 2014.

[2] Yoichi Murakami and Kenji Mizuguchi. Applying the naïve bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics*, 26(15):1841–1848, 2010.

[3] Aleksey Porollo and Jarosław Meller. Prediction-based fingerprints of protein–protein interactions. *Proteins: Structure, Function, and Bioinformatics*, 66(3):630–645, 2007.

[4] Gurdeep Singh, Kaustubh Dhole, Priyadarshini P Pai, and Sukanta Mondal. Springs: prediction of protein-protein interaction sites using artificial neural networks. Technical report, PeerJ PrePrints, 2014.

[5] Ghazaleh Taherzadeh, Yuedong Yang, Tuo Zhang, Alan Wee-Chung Liew, and Yaoqi Zhou. Sequence-based prediction of protein–peptide binding sites using support vector machine. *Journal of computational chemistry*, 37(13):1223–1229, 2016.

[6] Zhi-Sen Wei, Ke Han, Jing-Yu Yang, Hong-Bin Shen, and Dong-Jun Yu. Protein–protein interaction sites prediction by ensembling svm and sample-weighted random forests. *Neurocomputing*, 193:201–212, 2016.

[7] Zhi-Sen Wei, Jing-Yu Yang, Hong-Bin Shen, and Dong-Jun Yu. A cascade random forests algorithm for predicting protein-protein interaction sites. *IEEE transactions on nanobioscience*, 14(7):746–760, 2015.

[8] Buzhong Zhang, Jinyan Li, Lijun Quan, Yu Chen, and Qiang Lü. Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network. *Neurocomputing*, 357:86–100, 2019.

[9] Jian Zhang and Lukasz Kurgan. Scriber: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics*, 35(14):i343–i353, 2019.

Acknowledgements

The research of L.I. is funded by an NSERC Discovery Grant (R3143A01) and a Research Tools and Instruments Grant (R3143A07). The research of G.B.G. is funded by an NSERC Discovery Grant RGPIN-2020-05733.