Sequence-based prediction of protein-protein interactions: a structure-aware interpretable deep learning model

Samuel Sledzieski[1⋆], Rohit Singh[1⋆], Lenore Cowen[2(✉)], and Bonnie Berger[1,3(✉)]

[1] Computer Science and Artificial Intelligence Lab., Massachusetts Institute of Technology, Cambridge, MA 02139
{samsl,rsingh,bab}@mit.edu
[2] Department of Computer Science, Tufts University, Medford, MA 02155
cowen@cs.tufts.edu
[3] Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139

The systematic mapping of physical protein-protein interactions (PPIs) in the cell has proven extremely valuable in deepening our understanding of protein function and biology. In species like yeast and human where a large network of experimentally determined PPIs exists, this network information has proven valuable for downstream inference tasks in understanding functional genomics and biological pathway analysis [11,9,2,4,3]. However, despite the introduction of high-throughput methods [5,7,8,10,12] to assay PPIs, in many non-model organisms the number of experimentally determined PPIs can be nearly nonextant. This motivates our study of computational methods to predict PPIs in such species from easily-attainable sequence data alone.

Here, we introduce a new deep learning method, D-SCRIPT (Deep Sequence Contact Residue Interaction Prediction Transfer), for determining if two proteins interact physically in the cell, based on their amino acid sequences. D-SCRIPT, like other recent successful deep learning methods PIPR and DPPI [**?**,6], belongs to the class of methods that perform PPI prediction from protein sequence alone. The advantage of a sequence-based approach is that the input sequence data is almost always available, due to the enormous advances in low-cost genome sequencing. Our key conceptual advance is a well-matched combination of input featurization and model architecture design. This fusion allows the model to be trained solely from sequence data, supervised with only a binary interaction label, and yet produce an accessible intermediate representation that captures the structural mechanism of interaction between the protein pair. Our design enables D-SCRIPT to offer a combination of advantages that have hitherto been unavailable simultaneously: broad applicability, interpretability, and high cross-species accuracy.

The D-SCRIPT model consists of two stages (Figure 1): generation of a rich feature representation for each protein sequence separately, then prediction of interaction based on these features, where the model is trained end-to-end across

⋆ These authors contributed equally to the work

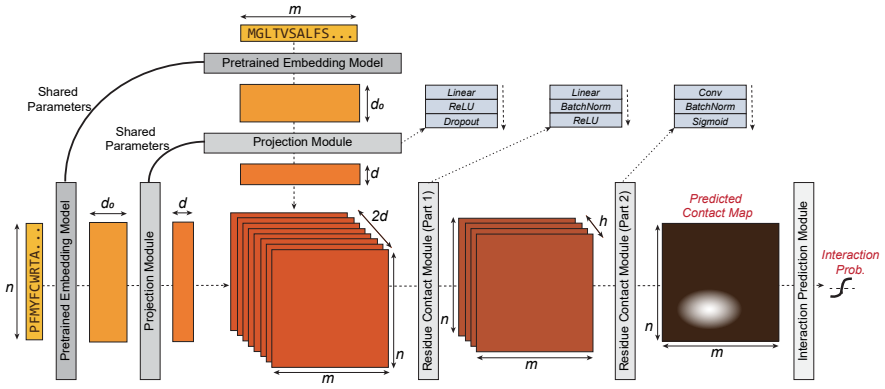

Fig. 1. D-SCRIPT architecture overview. The pretrained language model generates features for each individual protein, which are then reduced to low-dimensional embeddings. These embeddings are combined to compute a sparse *inter-protein* contact map, and a customized max-pooling operation is used to predict probability of interaction.

both stages. The first stage is accomplished by using the pre-trained protein language model from Bepler & Berger [1] followed by a projection module, where the model learns low-dimensional protein embeddings. A key innovation of D-SCRIPT is in our design of a structurally-aware second stage that encodes a *physical* model of protein interaction: we predict two proteins to interact only if there exists a short sequence of residues in the first protein that is highly compatible with a short sequence of residues in the second. In the contact module, the stacked representations of pairs of residues are projected into a lower dimensional space, and the low-dimensional embeddings are used to compute a sparse contact map which predicts the locations of contacts between protein residues. Finally, the interaction module uses a customized max-pooling operation on the contact map to predict the probability of interaction between the proteins.

We compared D-SCRIPT with PIPR [**?**], a best-performing state-of-the-art method for sequence-based human PPI prediction. We trained each model on 38,345 human PPIs and evaluated it using PPI data sets from *H. sapiens* and five other model organisms (*M. musculus*, *D. melanogaster*, *C. elegans*, *S. cerevisiae*, and *E. coli*). While D-SCRIPT under-performs PIPR in same-species (human) PPI prediction (0.516 vs. 0.844 AUPR), it significantly outperforms PIPR cross-species and maintains a high performance across all species, even those which are highly evolutionarily distant from human (0.547 vs. 0.352 average AUPR across five model species). In fact, its AUPR in these species remains comparable to that seen in human cross-validation, while PIPR's AUPR drops off significantly in other species. An investigation of the intermediate stages of the model shows that D-SCRIPT substantially models the physical process of protein interaction.

Full manuscript and software available at: http://dscript.csail.mit.edu

Acknowledgements We thank Tristan Bepler for helpful discussions and technical assistance. SS, RS and BB were supported by the NIH grant R01 GM081871. LC was supported by NSF HDR grant 1939263.

References

1. Bepler, T., Berger, B.: Learning protein sequence embeddings using information from structure. In: 7th International Conference on Learning Representations, ICLR 2019 (2019)
2. Cho, H., Berger, B., Peng, J.: Compact integration of multi-network topology for functional analysis of genes. Cell Systems **3**(6), 540–548 (2016)
3. Choobdar, S., Ahsen, M.E., Crawford, J., Tomasoni, M., Fang, T., Lamparter, D., Lin, J., Hescott, B., Hu, X., Mercer, J., et al.: Assessment of network module identification across complex diseases. Nature Methods **16**(9), 843–852 (2019)
4. Cowen, L., Ideker, T., Raphael, B.J., Sharan, R.: Network propagation: a universal amplifier of genetic associations. Nature Reviews Genetics **18**(9), 551 (2017)
5. Fields, S., Song, O.k.: A novel genetic system to detect protein–protein interactions. Nature **340**(6230), 245–246 (1989)
6. Hashemifar, S., Neyshabur, B., Khan, A.A., Xu, J.: Predicting protein-protein interactions through sequence-based deep learning. In: Bioinformatics. vol. 34, pp. i802–i810. Oxford University Press (sep 2018). https://doi.org/10.1093/bioinformatics/bty573, https://academic.oup.com/bioinformatics/article/34/17/i802/5093239
7. Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., et al.: Global landscape of protein complexes in the yeast saccharomyces cerevisiae. Nature **440**(7084), 637–643 (2006)
8. Kumar, A., Snyder, M.: Protein complexes take the bait. Nature **415**(6868), 123–124 (2002)
9. Navlakha, S., Kingsford, C.: The power of protein interaction networks for associating genes with diseases. Bioinformatics **26**(8), 1057–1063 (2010)
10. Sahni, N., Yi, S., Taipale, M., Bass, J.I.F., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G.I., Wang, Y., et al.: Widespread macromolecular interaction perturbations in human genetic disorders. Cell **161**(3), 647–660 (2015)
11. Sharan, R., Ulitsky, I., Shamir, R.: Network-based prediction of protein function. Molecular Systems biology **3**(1), 88 (2007)
12. Taipale, M., Tucker, G., Peng, J., Krykbaeva, I., Lin, Z.Y., Larsen, B., Choi, H., Berger, B., Gingras, A.C., Lindquist, S.: A quantitative chaperone interaction network reveals the architecture of cellular protein homeostasis pathways. Cell **158**(2), 434–448 (2014)