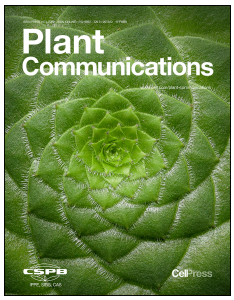
Journal Pre-proof



iCREPCP: a deep learning-based web server for identifying base-resolution *cis*-regulatory elements within plant core promoters

Kaixuan Deng, Qizhe Zhang, Yuxin Hong, Jianbing Yan, Xuehai Hu

PII: S2590-3462(22)00292-9

DOI: https://doi.org/10.1016/j.xplc.2022.100455

Reference: XPLC 100455

To appear in: *PLANT COMMUNICATIONS*

Received Date: 31 July 2022

Revised Date: 16 September 2022

Accepted Date: 23 September 2022

Please cite this article as: Deng, K., Zhang, Q., Hong, Y., Yan, J., Hu, X., iCREPCP: a deep learning-based web server for identifying base-resolution *cis*-regulatory elements within plant core promoters, *PLANT COMMUNICATIONS* (2022), doi: https://doi.org/10.1016/j.xplc.2022.100455.

iCREPCP: a deep learning-based web server for identifying base-resolution *cis*-regulatory elements within plant core promoters

Kaixuan Deng[1, #], Qizhe Zhang[1, #], Yuxin Hong[1], Jianbing Yan[2,*] and Xuehai Hu[1,*]

[1]College of Informatics, Hubei Engineering Technology Research Center of Agricultural Big Data, Huazhong Agricultural University, Wuhan, Hubei, P.R. of China

[2]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan 430070, China

*Correspondence: Xuehai Hu (huxuehai@mail.hzau.edu.cn) and Jianbing Yan (yjianbing@mail.hzau.edu.cn)

[#] These authors contributed equally to this article.

Tel.: +86-18171282783; Fax: +86-27-87288509.

Dear editor,

A central question of plant biology is to specify the temporal and spatial patterns as well as the quantitative level of gene expressions, which are significantly associated with important agronomic traits. There is a growing consensus in the past decade that two key factors determining gene expression level are *cis*-regulatory modules (CRMs) and *trans*-acting factors (TAFs) (Schmitz et al., 2022). Common CRMs include gene-proximal promoters and distal enhancers, which are all considered as the complex assemblies of *cis*-regulatory elements (CREs). It is the binding or interaction between CREs and TAFs (often are transcription factors, TFs) in a ubiquitous or cell-specific manner that determines in which cell, at what time and at what level a gene is expressed. Therefore, the identification of plant CRMs or critical CREs will not only help us understand transcriptional regulatory mechanisms in plants, but also is an essential prerequisite for plant breeding 4.0─breeding by genome editing (Gao, 2021).

However, comparing with rich data resources on CREs in mammalian genomes (Fornes et al., 2020), related works in plants have lagged far behind (Schmitz *et al.*, 2022). The bottleneck mainly lies in two aspects: (1) the lack of a big project like ENCODE in plants makes epigenomic features absent or fragmented, leading only a handful of putative plant CREs from genome-wide identification; (2) too few transient transfection systems (only two of protoplasts and tobacco leaves (Jores et al., 2021)), together with difficult validation assays like self-transcribing active regulatory region-sequencing (STARR-seq) in plants, make fewer experimental-validated CREs.

Plant **C**ore **P**romoter (**PCP**), with the minimal sequence region of 50-100bp around transcription start site (TSS), is a large group of CRMs that are rich in CREs, and can drive basal level of target gene transcriptions (Schmitz *et al.*, 2022). The promoter strength of PCP is defined as the ability to drive expression of a barcoded green fluorescent protein (GFP) reporter gene via transient transfection systems. To our best knowledge, there is no existing computational tool for identifying CREs within PCPs. Here, we developed a deep learning-based web server (http://www.hzau-hulab.com/icrepcp/) to **i**dentify which **CRE**s a given **P**lant **C**ore **P**romoter (**iCREPCP**) contains, with a focus on base-resolution position of each CRE and its contribution to the promoter strength.

We first downloaded a large-scale PCP dataset of 18,329 *Arabidopsis*, 34,415 maize and 27,094 sorghum core promoters, whose strengths were measured by STARR-seq assays in six transient transfection systems (tobacco leaves with enhancer in dark, tobacco leaves without enhancer in dark, tobacco leaves with enhancer in light, tobacco leaves without enhancer in light, maize protoplasts with enhancer in dark and maize protoplasts without enhancer in dark) (Jores *et al.*, 2021). We will take

'sequence' as input and 'enrichment' as output of a total of about 76,000 samples from all three species for training and testing deep learning models.

We next trained a deep learning architecture of 'DenseNet' (Huang, 2017) to fit promoter strengths with their DNA sequences. DenseNet has won the best paper award of CVPR-2017, and it can alleviate the vanishing-gradient problem (Figure 1A, Supplementary Information). As expected, **iCREPCP** can accurately fit the experimental results in all six transfection systems: the mean training R^2 ranges from 0.490 to 0.782, and all models have low variances, implying their feasibilities (Figure 1B). We next investigate its generalizability by an independent testing dataset (Supplementary Information). Remarkably, **iCREPCP** achieves good testing R^2 ranging from 0.420 to 0.752 and obviously improves the previous work who employed a simple convolutional neural network (Jores *et al.*, 2021) (Figure 1B), also implying its strong generalizability. Moreover, the small differences between training R^2 and testing R^2 (ranging from 0.03 to 0.07) demonstrate that **iCREPCP** have little problems of overfitting, further suggesting that they have potential transfer abilities for other plant species.

To investigate the biological interpretability and practicability of **iCREPCP**, we here are more concerned on the contribution of each base during the promoter strength prediction of PCP rather than the prediction accuracy. Because several successive bases having high contributions are potential critical CREs, which are ideal targets of genome editing engineering (Gao, 2021). To this end, we employed a powerful interpretability tool of DeepLIFT (Shrikumar, 2017) to assign a DeepLIFT contribution score to each base of a given PCP. We employed two known PCP examples of maize YIGE1 gene and rice IPA1 gene for demonstrating the detecting power of **iCREPCP** together with DeepLIFT (DeepLIFT contribution scores are visualized as high characters with colors that help readers easily find critical bases). YIGE1 is a newly-reported maize gene contributing to ear length and grain yield, and a single-nucleotide polymorphism (SNP) located in its regulatory region had a large effect on its promoter strength (Luo et al., 2022). Using the trained model of tobacco leaves without enhancer in light, **iCREPCP** successfully located a large-contributed regulatory region flanking the important SNP (also repeatedly detected by two additional interpretability tools of in-silico tilling deletion and in-silico mutagenesis, Figure 1C), suggesting its detecting power. For trans-species circumstance, IPA1 is a rice star gene that is a master regulator of rice plant architecture. Its' function was known to increase grains per panicle but reduce tillers, but a recent breakthrough reported that a 54-base pair *cis*-regulatory deletion can both increase grains per panicle and tiller number (Song et al., 2022). Surprisingly, **iCREPCP** successfully detected a 12bp region (-128~-117) with large contributions that exactly covers the An-1 binding site within the deletion (Figure

1D, Supplementary Figure 1), implying that **iCREPCP** has great potentials for trans-species identifications of base-resolution critical CREs.

For a rough estimation of precision and recall of iCREPCP, we constructed a benchmark of *Arabidopsis* CREs, which was employed for an evaluation: precision and recall are 0.447 and 0.344 respectively (Supplementary Figure 3 and Supplementary materials).

To investigate biological implications of several successive bases with high DeepLIFT contribution scores, we next naturally ask whether they are TF motifs and then employed a new motif discovery algorithm of TF-MoDISco (Shrikumar, 2018), that was specifically developed for deep learning, to identify high-quality, non-redundant TF motifs within PCPs (Supplementary Information). For the trained model of tobacco leaves without enhancer in light, TF-MoDISco totally identified 21 clustered seqlets, 14 out of which have perfect matching in JASPAR database (Figure 1E, Supplementary Figure 2 and Table 1). To further quantify the population-level effect size of the 14 enriched TF motifs, we performed a global importance analysis (Koo et al., 2021) and found that 8 (including TATATA motif, TCP8 and AP1) out of 14 have positive global importance, whereas 6 (including ERF3 and ABI3) out of 14 have negative effects (Figure 1F). Finally, we scanned all 75,375 PCPs using the 14 PWMs of the enriched TF motifs and gave a comprehensive statistic about their occurrence numbers in each PCP sample (Figure 1G, Supplementary Table 2). Notably, the TATATA motif has the most occurrence numbers within PCPs having large promoter strengths in all three species, whereas the ERF3 motif has more occurrences within PCPs having small promoter strengths in both sorghum and maize, which is consistent with their global importance analysis results.

In summary, iCREPCP (Figure 1H) provides a user-friendly platform to identify critical CREs that importantly contribute to the promoter strength of any given PCPs with base resolution. These resources, including the six trained prediction models and a powerful visualization tool, will greatly help plant scientists in at least two respects: (i) easily obtain an accurate prediction value of the promoter strength with the only need of the 170bp DNA sequence around its TSS; (ii) precisely detect the base-resolution position of each CRE and its contribution to the promoter strength. The later function will provide important candidate targets of genome editing and will be of general interests in the plant community. The main limitation of iCREPCP is that it was trained with promoter strength measured in vitro via tobacco leaves or maize protoplasts, implying that iCREPCP might work not well on some genes needing distinct expression pattern in vivo. Another limitation is that the prediction accuracy is sensitive to the boundary (Supplementary Figure 4), imply that our models only can be used on (-165, +5) of TSS. Further

improvements of iCREPCP will focus on the accurate identification of distal CREs: (i) take longer genomic sequences as the inputs, covering more distal CREs (such as enhancers) influencing gene expressions; (ii) develop more sophisticated models for capturing long-range dependency information.

Data availability

The datasets and codes used to build the DenseNet model, to compute the DeepLIFT contribution scores and to perform TF-MoDISco analysis are available at https://github.com/kaixuanDeng95/iCREPCP.

Acknowledgments

We acknowledge Prof. Weibo Xie for helpful discussions and we thank Dr. Yun Luo for providing the example of maize YIGE1 gene. We also thank four anonymous reviewers for their helpful suggestions that have greatly improved the original manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (32070689 to Xuehai Hu) and the Joint Funds of the National Natural Science Foundation of China (U1901201 to Jianbing Yan).

Conflicts of interest

The authors declare no conflicts of interest.

Authors contributions

X.-H.H. and J.-B.Y. designed the research and wrote the manuscript. K.-X.D., Q.-Z.Z. and Y.-X.H. collected the data and built the model. K.-X.D. and Q.-Z.Z. performed the DeepLIFT analysis. K.-X.D. performed the motif analysis and developed the web server of iCREPCP. All authors read and approved the final manuscript.

References

Fornes, O., Castro-Mondragon, J.A., Khan, A., van der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M., Baranašić, D., et al. (2020). JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic acids research **48**:D87-d92. 10.1093/nar/gkz1001.

Gao, C. (2021). Genome engineering for crop improvement and future agriculture. Cell **184**:1621-1635. 10.1016/j.cell.2021.01.005.

Huang, G.L., Z; Van Der Maaten, L; Weinberger, KQ. (2017). Densely Connected Convolutional Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR):17355312. DOI: 10.1109/CVPR.2017.243.

Jores, T., Tonnies, J., Wrightsman, T., Buckler, E.S., Cuperus, J.T., Fields, S., and Queitsch, C. (2021). Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. Nature plants **7**:842-855. 10.1038/s41477-021-00932-y.

Koo, P.K., Majdandzic, A., Ploenzke, M., Anand, P., and Paul, S.B. (2021). Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. PLoS computational biology **17**:e1008925. 10.1371/journal.pcbi.1008925.

Luo, Y., Zhang, M., Liu, Y., Liu, J., Li, W., Chen, G., Peng, Y., Jin, M., Wei, W., Jian, L., et al. (2022). Genetic variation in YIGE1 contributes to ear length and grain yield in maize. The New phytologist **234**:513-526. 10.1111/nph.17882.

Schmitz, R.J., Grotewold, E., and Stam, M. (2022). Cis-regulatory sequences in plants: Their importance, discovery, and future challenges. The Plant cell **34**:718-741. 10.1093/plcell/koab281.

Shrikumar, A., Greenside, P., Kundaje, A. (2017). Learning Important Features Through Propagating Activation Differences. Proceedings of the 34th International Conference on Machine Learning **70**:3145-3153.
Shrikumar, A., Tian, K., Avsec, Ž., Shcherbina, A., Banerjee, A., Sharmin, M., Nair, S., Kundaje, A. (2018). TF-MoDISco v0.4.2.2-alpha: technical note. Preprint at arXiv https://arxiv.org/abs/1811.00416.
Song, X., Meng, X., Guo, H., Cheng, Q., Jing, Y., Chen, M., Liu, G., Wang, B., Wang, Y., Li, J., et al. (2022). Targeting a gene regulatory element enhances rice grain yield by decoupling panicle number and size. Nature biotechnology 10.1038/s41587-022-01281-7.

Figure 1. The workflow of **iCREPCP**.

(**A**) The deep learning architecture of DenseNet.

(**B**) The prediction performances via R^2 on training sets and on independent testing sets for six transient systems.

(**C**) The example of maize YIGE1 gene for demonstrating the detecting power of iCREPCP. Top panel, a snapshot of core promoter region: chr1_51127917-51128086; The second panel is the FIMO scanning results; The third panel is the DeepLIFT contribution score; The fourth and fifth panels are used to demonstrate the results of in-silico tilling deletion, which measure the difference of predicted promoter strength with a sliding window of 5-bp deletion across the whole sequence; The bottom panel is the heatmap for demonstrating in-silico mutagenesis results.

(**D**) A trans-species example of rice IPA1 gene with the same layout of (C).

(**E**) A total of 14 seqlets identified by TF-MoDISco of the model of tobacco leaves without enhancer in light and their similar TF motifs in JASPAR.

(**F**) Motif occurrence frequencies and global importances of 14 enriched TF motifs of the model of tobacco leaves without enhancer in light.

(**G**) The heatmap for demonstrating occurrence numbers of 14 enriched TF motifs within all 75,375 PCPs. Each row represents a PCP and each column represents a specific TF motif. The row order (from top to bottom) is based on promoter strength (from high to low) within each species and the column order (from left to right) is based on the total occurrence number of TF motifs across three species (from more to less).

(**H**) The homepage of iCREPCP.