Plant Communications
Correspondence

CellPress
Partner Journal

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

Dear Editor,

A central question of plant biology is how to specify the temporal and spatial patterns as well as the quantitative levels of gene expression, which are significantly associated with important agronomic traits. There has been a growing consensus in the past decade that the 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 both considered to be complex assemblies of *cis*-regulatory elements (CREs). It is the binding or interaction between CREs and TAFs (often 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 serve as an essential prerequisite for plant breeding 4.0—breeding by genome editing (Gao, 2021).

However, by comparison with rich data resources on CREs in mammalian genomes (Fornes et al., 2020), related work in plants has lagged far behind (Schmitz et al., 2022). This bottleneck has two main aspects: (1) the lack of a large project like ENCODE in plants makes epigenomic features absent or fragmented, leading to only a handful of putative plant CREs from genome-wide identification; (2) too few transient transfection systems (only two, in protoplasts and tobacco leaves; Jores et al., 2021), together with difficult validation assays, such as self-transcribing active regulatory region sequencing in plants, has resulted in fewer experimentally validated CREs.

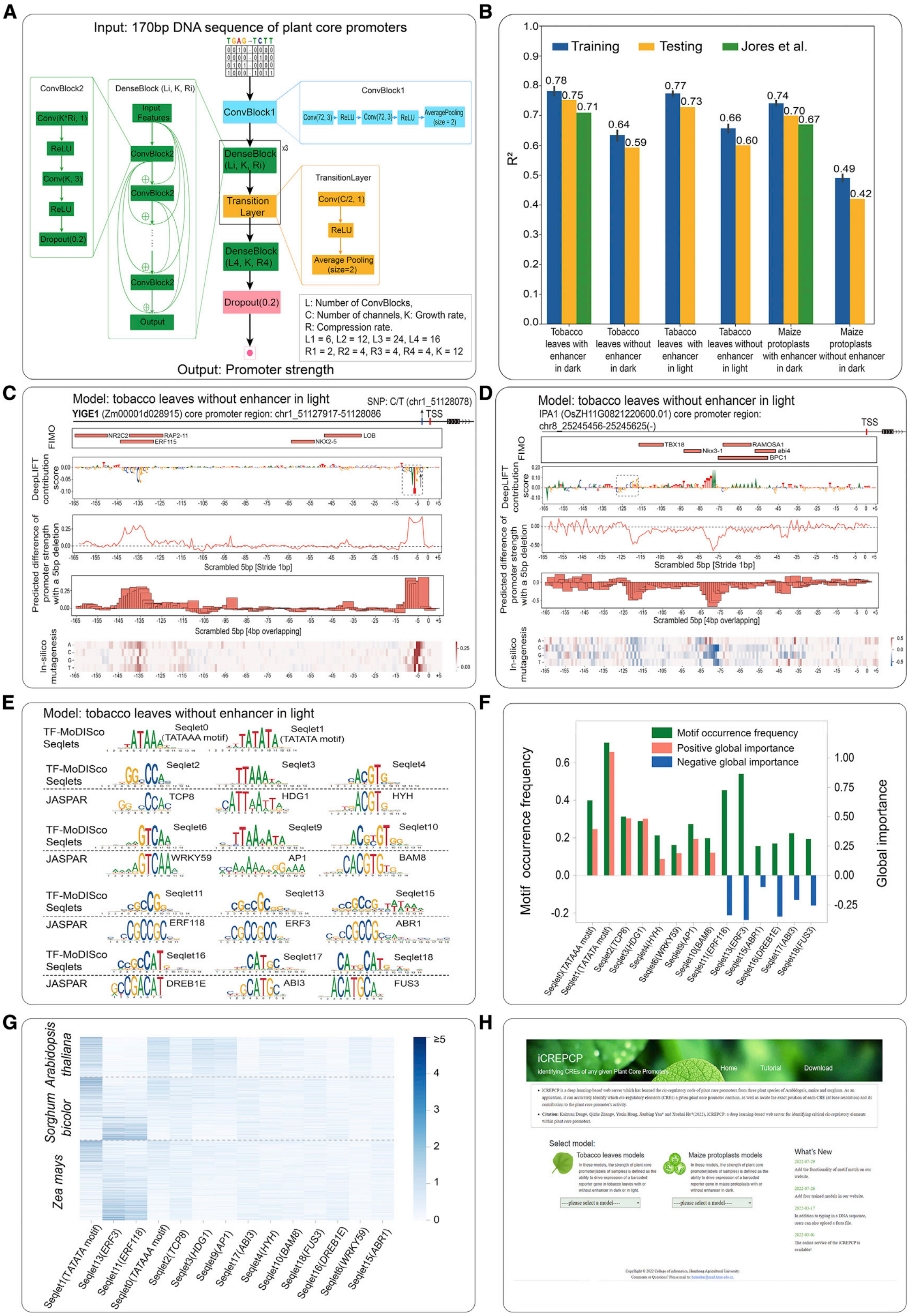
The plant core promoter (PCP), with a minimal sequence region of 50–100 bp around the transcription start site (TSS), is a large group of CRMs that are rich in CREs and can drive the basal level of target gene transcription (Schmitz et al., 2022). The promoter strength of the PCP is defined as its ability to drive the expression of a barcoded green fluorescent protein reporter gene via transient transfection systems. To the best of our 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 identify the CREs contained in a given PCP (iCREPCP), with a focus on the base-resolution position of each CRE and its contribution to 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 self-transcribing active regulatory region sequencing assays in six transient transfection systems (tobacco leaves with enhancer in the dark, tobacco leaves without enhancer in the dark, tobacco leaves with enhancer in the light, tobacco leaves without enhancer in the light, maize protoplasts with enhancer in the dark, and maize protoplasts without enhancer in the dark) (Jores et al., 2021). We took "sequence" as input and "enrichment" as output from a total of ~76 000 samples of all three species for training and testing deep learning models.

We next trained a deep learning architecture of "DenseNet" (Huang et al., 2017) to fit promoter strengths with their DNA sequences. DenseNet won the best paper award at CVPR-2017 and can alleviate the vanishing-gradient problem (Figure 1A supplemental information). As expected, iCREPCP accurately fit the experimental results from all six transfection systems: the mean training R^2 ranged from 0.490 to 0.782, and all models had low variances, implying their feasibility (Figure 1B). We next investigated its generalizability using an independent testing dataset (supplemental information). iCREPCP achieved good testing R^2 values from 0.420 to 0.752 and clearly improved on previous work that used a simple convolutional neural network (Jores et al., 2021) (Figure 1B), implying its strong generalizability. Moreover, the small differences between training R^2 and testing R^2 (ranging from 0.03 to 0.07) demonstrated that iCREPCP has few problems with overfitting, further suggesting that it has potential transfer abilities for other plant species.

To investigate the biological interpretability and practicality of iCREPCP, we are more concerned here with the contribution of each base during promoter strength prediction of the PCP rather than on prediction accuracy. Several successive bases that make high contributions are potential critical CREs and are therefore ideal targets for genome-editing engineering (Gao, 2021). To identify such bases, we employed a powerful interpretability tool, DeepLIFT (Shrikumar et al., 2017), to assign a DeepLIFT contribution score to each base of a given PCP. We employed two known PCP examples, the maize *YIGE1* gene and the rice *IPA1* gene, to demonstrate the detection power of iCREPCP together with DeepLIFT (DeepLIFT contribution scores are visualized as tall characters with colors to help readers easily identify the critical bases). *YIGE1* is a newly reported maize gene that contributes to ear length and grain yield; a single-nucleotide polymorphism located in its regulatory region has a large effect on its promoter strength (Luo et al., 2022). Using the trained model from tobacco leaves without enhancer in the light, iCREPCP successfully located a regulatory region with a

Published by the Plant Communications Shanghai Editorial Office in association with Cell Press, an imprint of Elsevier Inc., on behalf of CSPB and CEMPS, CAS.



- iCREPCP is a deep learning-based web server which has learned the *cis*-regulatory code of plant core promoters from three plant species of Arabidopsis, maize and sorghum. As an application, it can accurately identify which *cis*-regulatory elements (CREs) a given plant core promoter contains, as well as locate the exact position of each CRE (at base resolution) and its contribution to the plant core promoter's activity.
- **Citation:** Kaixuan Deng#, Qizhe Zhang#, Yuxin Hong, Jianbing Yan* and Xuehai Hu*(2022), iCREPCP: a deep learning-based web server for identifying critical cis-regulatory elements within plant core promoters.

In these models, the strength of plant core promoter(labels of samples) is defined as the ability to drive expression of a barcoded reporter gene in tobacco leaves with or without enhancer in dark or in light.

In these models, the strength of plant core promoter(labels of samples) is defined as the ability to drive expression of a barcoded reporter gene in maize protoplasts with or without enhancer in dark.

2022-07-29
Add the functionality of motif match on our website.

2022-07-26
Add four trained models in our website.

2022-03-17
In addition to typing in a DNA sequence, users can also upload a fasta file.

2022-03-01
The online service of the iCREPCP is available!

CopyRight © 2022 College of informatics, Huazhong Agricultural University.
Comments or Questions? Please mail to: huxuehai@mail.hzau.edu.cn.

Figure 1. The workflow of iCREPCP.
(A) The deep learning architecture of DenseNet.
(B) The prediction performances via R^2 on training sets and independent testing sets from six transient expression systems (The error bar represents the R^2 fluctuation among 10 times training with random division of the training set and the validation set).

(legend continued on next page)

large contribution flanking the important single-nucleotide polymorphism (also repeatedly detected by two additional interpretability tools, *in silico* tilling deletion and *in silico* mutagenesis, Figure 1C), demonstrating its detection power. For the *trans*-species circumstance, *IPA1* is a key rice gene that is a master regulator of rice plant architecture. Its function is known to increase grains per panicle but reduce tillers; however, a recent breakthrough showed that a 54-base pair *cis*-regulatory deletion can increase both grains per panicle and tiller number (Song et al., 2022). Surprisingly, iCREPCP successfully detected a 12-bp region (–128 to approximately –117) with large contributions that exactly covered the An-1 binding site within the deletion (Figure 1D and Supplemental Figure 1), implying that iCREPCP has great potential for *trans*-species identification of critical CREs with base-level resolution.

To obtain a rough estimate of the precision and recall of iCREPCP, we constructed a benchmark of *Arabidopsis* CREs that was used for evaluation: precision and recall were 0.447 and 0.344, respectively (Supplemental Figure 3 and supplemental information).

To investigate the biological implications of several successive bases with high DeepLIFT contribution scores, we next asked whether they were TF motifs and then used a new motif discovery algorithm, TF-MoDISco (Shrikumar et al., 2018), which was specifically developed for deep learning, to identify high-quality, non-redundant TF motifs within PCPs (supplemental information). For the model trained on tobacco leaves without enhancer in the light, TF-MoDISco identified 21 clustered seqlets, 14 of which have perfect matches in the JASPAR database (Figure 1E and Supplemental Figure 2; Supplemental 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 the TATATA motif, TCP8, and AP1) had positive global importance, whereas 6 (including ERF3 and ABI3) had negative effects (Figure 1F). Finally, we scanned all 75 375 PCPs using the 14 PWMs of the enriched TF motifs and obtained comprehensive statistics for their occurrence numbers in each PCP sample (Figure 1G; Supplemental Table 2). Notably, the TATATA motif had the highest occurrence numbers in PCPs with large promoter strengths in all three species, whereas the ERF3 motif had more occurrences in PCPs with low promoter strengths in sorghum and maize, consistent with their results in the global importance analysis.

In summary, iCREPCP (Figure 1H) provides a user-friendly platform for the identification of critical CREs that make an important contribution to the promoter strength of any given PCP with base-level resolution. These resources, including the six trained prediction models and a powerful visualization tool, will help plant scientists to: (1) easily obtain an accurate promoter strength prediction based on only the 170-bp DNA sequence around the TSS; and (2) precisely detect the position of each CRE with base-level resolution and its contribution to promoter strength. The latter function will provide important candidate targets for genome editing and will be of general interest to 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 may not work well on some genes that exhibit distinct expression patterns *in vivo*. Another limitation is that prediction accuracy is sensitive to the boundary (Supplemental Figure 4), implying that our models can only be used on the region (–165, +5) of the TSS. Further improvements in iCREPCP will focus on the accurate identification of distal CREs: (1) taking longer genomic sequences as inputs in order to cover more distal CREs (such as enhancers) that influence gene expression; and (2) developing more sophisticated models for capturing long-range dependency information.

DATA AVAILABILITY

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

SUPPLEMENTAL INFORMATION

Supplemental information is available at *Plant Communications Online*.

FUNDING

This work was supported by the National Natural Science Foundation of China (32070689 to X.H.H.) and the Joint Funds of the National Natural Science Foundation of China (U1901201 to J.-B.Y.).

AUTHOR 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 for iCREPCP. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

We acknowledge Prof. Weibo Xie for helpful discussions, and we thank Dr. Yun Luo for providing the example of the maize *YIGE1* gene. We also thank four anonymous reviewers for their helpful suggestions that have greatly improved the original manuscript. No conflict of interest declared.

Received: July 31, 2022
Revised: September 16, 2022
Accepted: September 23, 2022
Published: September 28, 2022

(C) The maize *YIGE1* gene is used as an example to demonstrate the detection power of iCREPCP. The top panel shows a snapshot of the core promoter region: chr1_51127917-51128086; the second panel shows the FIMO scanning results; the third panel shows the DeepLIFT contribution scores; the fourth and fifth panels show the results of *in silico* tilling deletion, in which the difference in predicted promoter strength is measured with a sliding window of 5-bp deletion across the whole sequence; the bottom panel is a heatmap demonstrating the *in silico* mutagenesis results.
(D) A *trans*-species example of the rice *IPA1* gene, using the same layout as **(C)**.
(E) Fourteen seqlets identified by TF-MoDISco using the model of tobacco leaves without enhancer in the light and their similar TF motifs in JASPAR.
(F) Motif occurrence frequencies and global importance values of 14 enriched TF motifs in the model of tobacco leaves without enhancer in the light.
(G) Heatmap demonstrating the 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 numbers of TF motifs across the three species (from more to fewer).
(H) The iCREPCP homepage.

Kaixuan Deng[1,3], *Qizhe Zhang*[1,3], *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. China

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

[3]These authors contributed equally to this article.

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

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

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 Res. **48**:D87–D92. https://doi.org/10.1093/nar/gkz1001.

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

Huang, G.L.Z., Van Der Maaten, L., and Weinberger, K.Q. (2017). Densely connected convolutional networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), p. 17355312. https://doi.org/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. Nat. Plants **7**:842–855. https://doi.org/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 Comput. Biol. **17**:e1008925. https://doi.org/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. New Phytol. **234**:513–526. https://doi.org/10.1111/nph.17882.

Schmitz, R.J., Grotewold, E., and Stam, M. (2022). Cis-regulatory sequences in plants: their importance, discovery, and future challenges. Plant Cell **34**:718–741. https://doi.org/10.1093/plcell/koab281.

Shrikumar, A., Greenside, P., and 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., and 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. Nat. Biotechnol. **40**:1403–1411. https://doi.org/10.1038/s41587-022-01281-7.